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

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

CAT-ASVAB Equating: Sample Size Targets

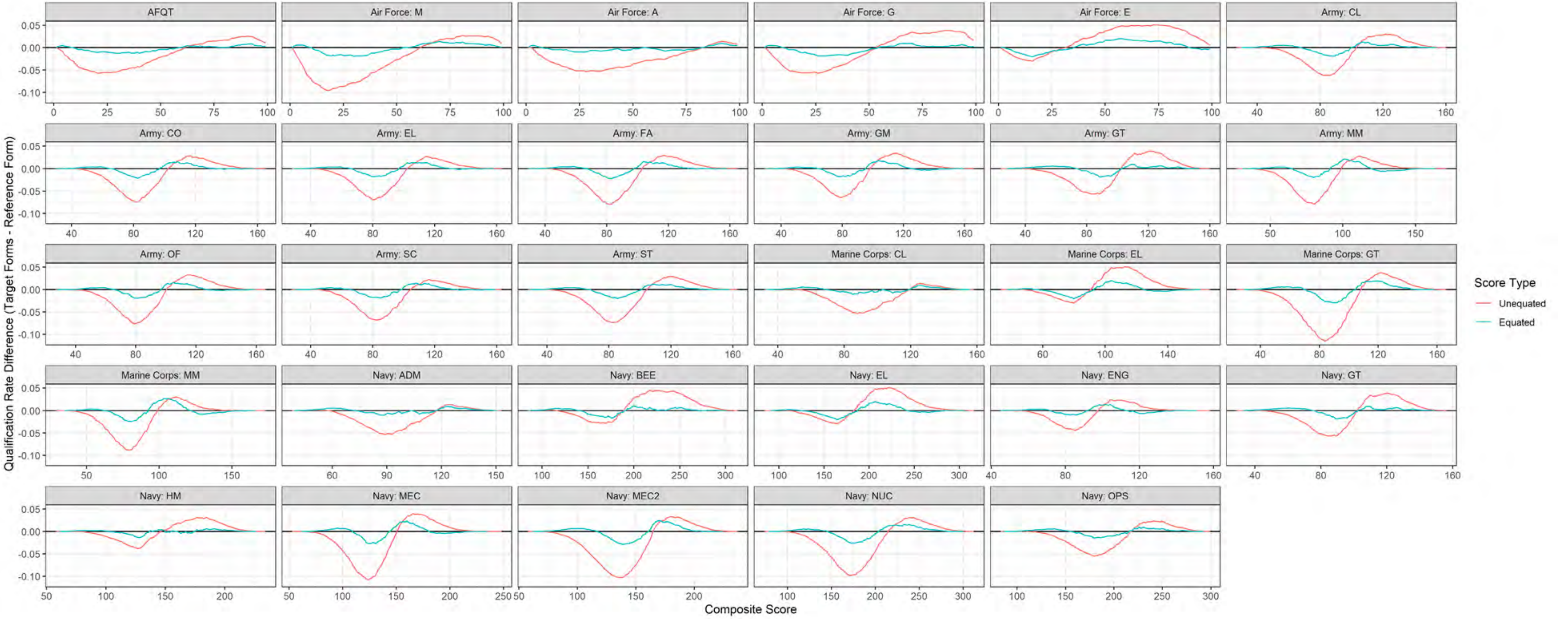
Form	Assignment Probability	Phase 1 Target	Phase 2 Target	Phase 3 Target
Reference	1/7	500	1,500	10,000
Operational	1/7	500	1,500	10,000
New Form A	1/7	500	1,500	10,000
New Form B	1/7	500	1,500	10,000
New Form C	1/7	500	1,500	10,000
New Form D	1/7	500	1,500	10,000
New Form E	1/7	500	1,500	10,000
Total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

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)



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?

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)

Word Knowledge (WK)

Auto Information (AI)

Shop Information (SI)

Math Knowledge (MK)

Electronics Information (EI)

Paragraph Comprehension (PC)

Mechanical Comprehension (MC)

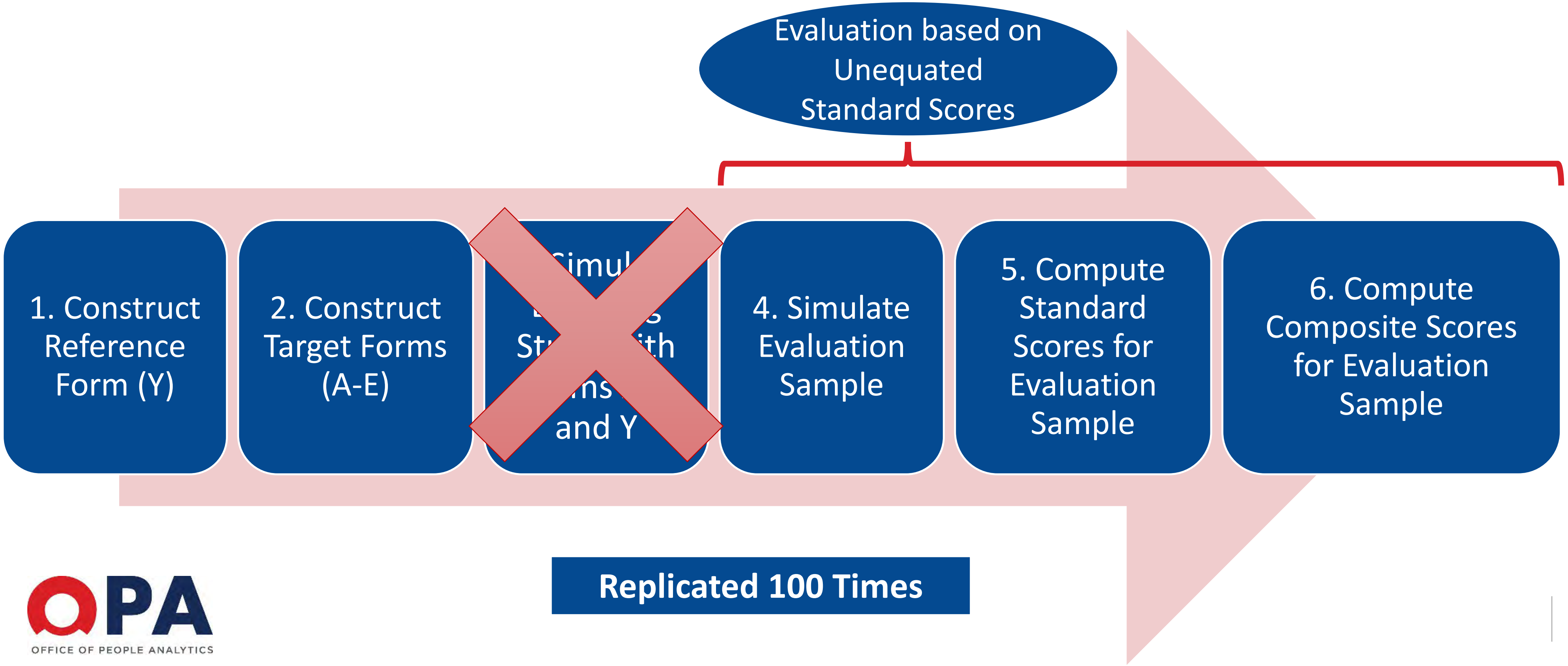
Arithmetic Reasoning (AR)

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

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



Evaluation based on Unequated Standard Scores

1. Construct Reference Form (Y)

2. Construct Target Forms (A-E)

Simul St ith ns and Y

4. Simulate Evaluation Sample

5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

Replicated 100 Times

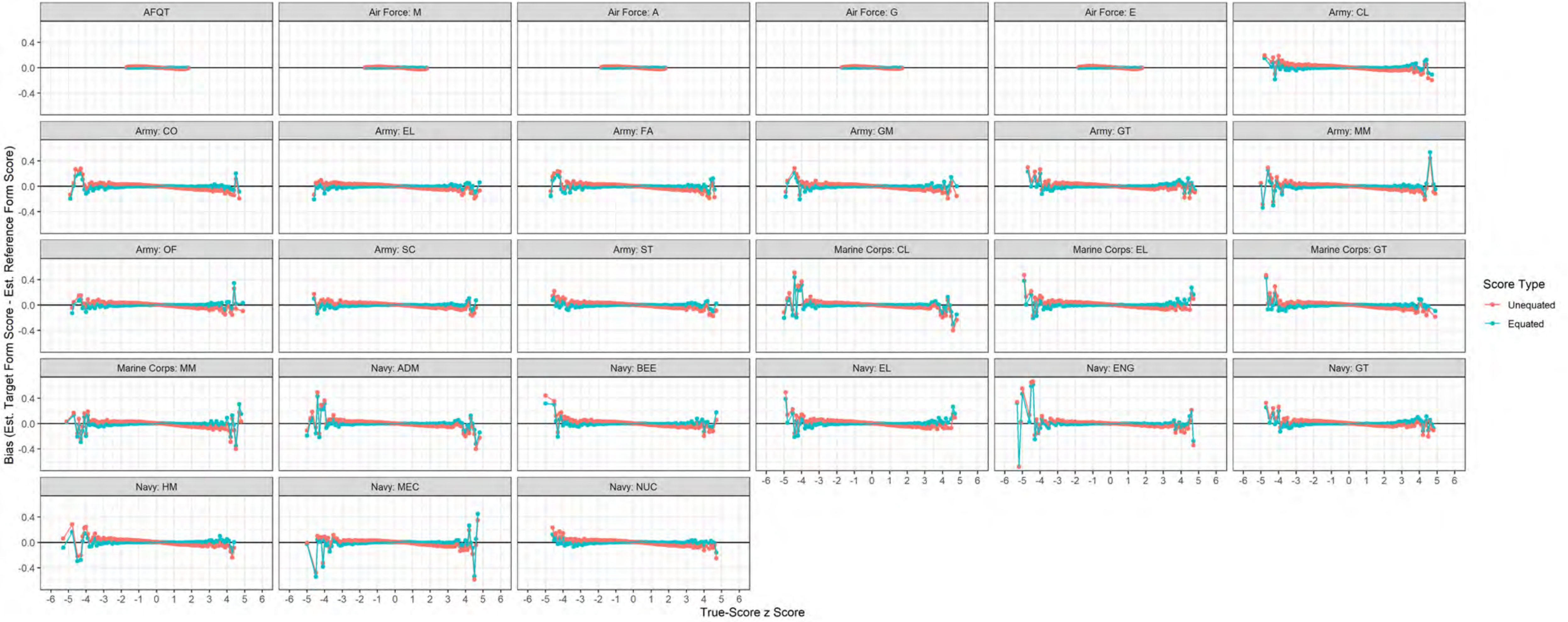
Evaluation of Conditional Score Bias

- 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

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

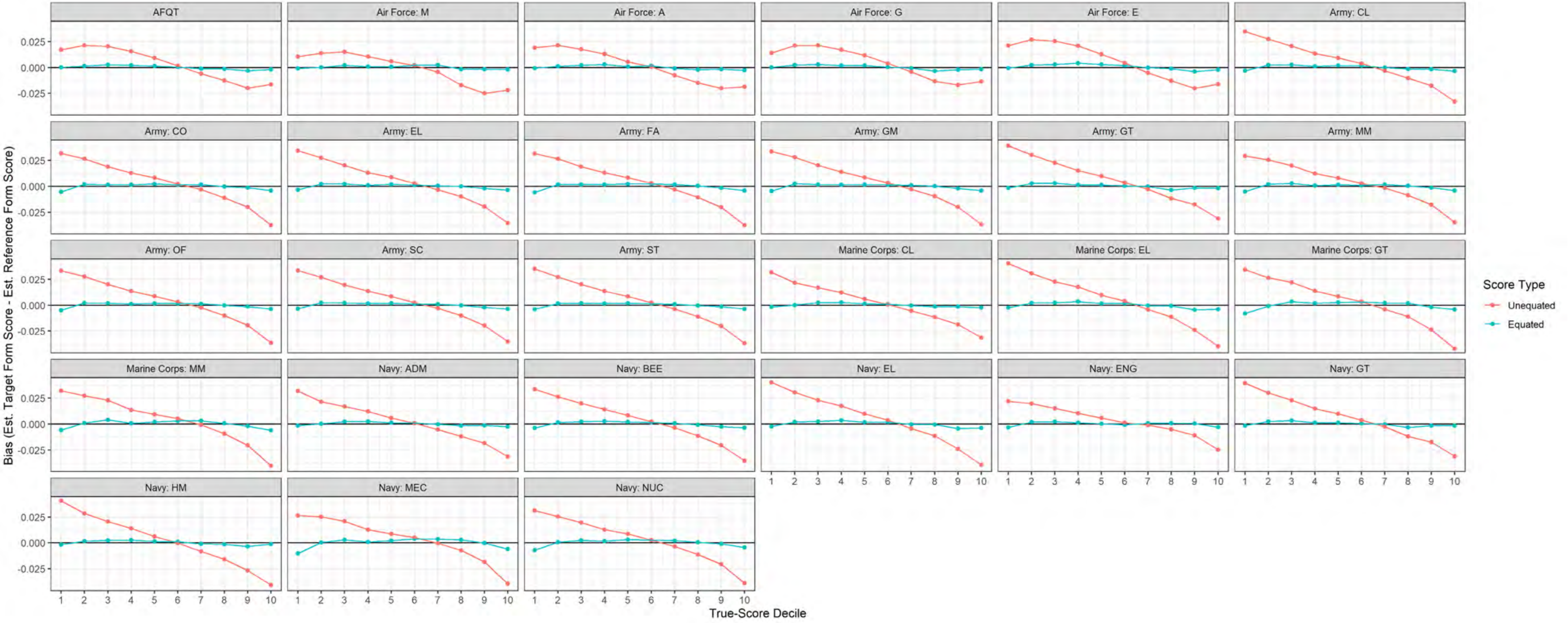
  - Scores evaluated in bias analyses were centered and scaled using the mean and SD of true scores (generating thetas converted to composite scores using generating TCs)
- The following plots depict mean bias effects across forms and replications

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



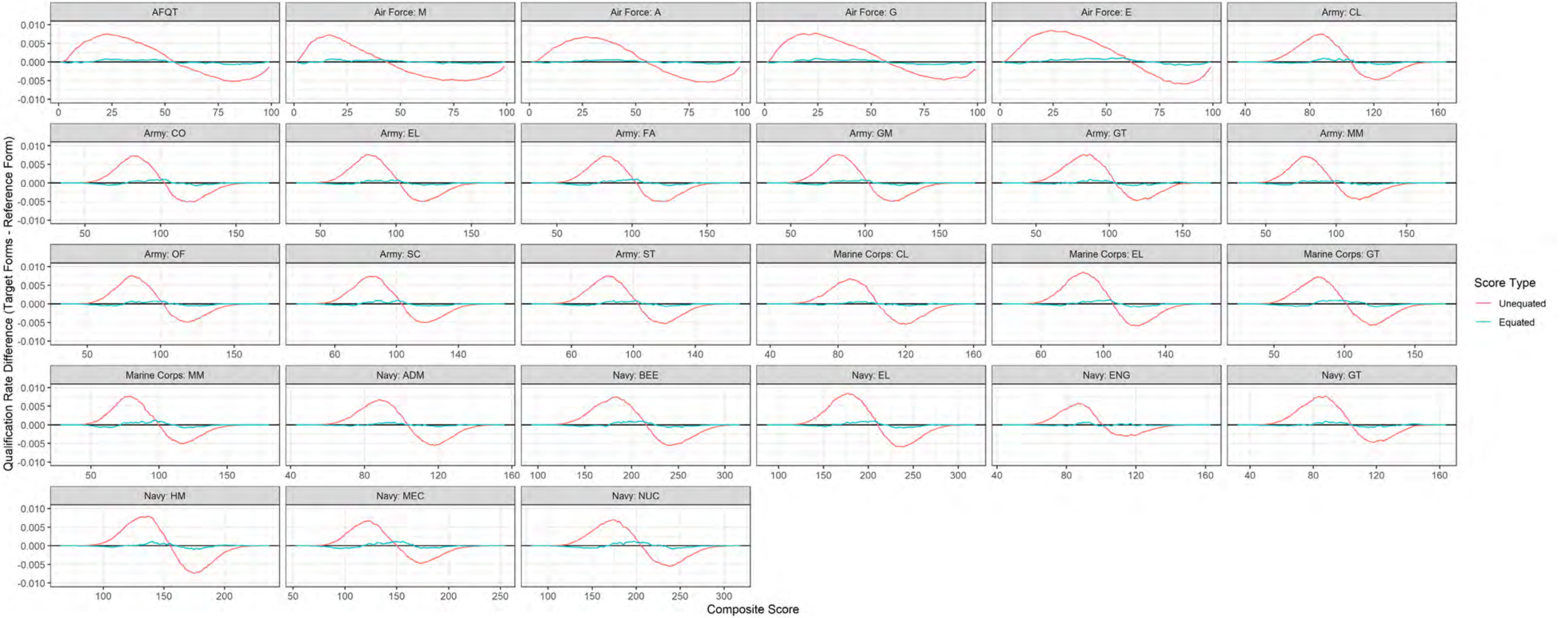
OFFICE OF PEOPLE ANALYTICS

Evaluation of Composite Score Bias by True-Score Decile



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

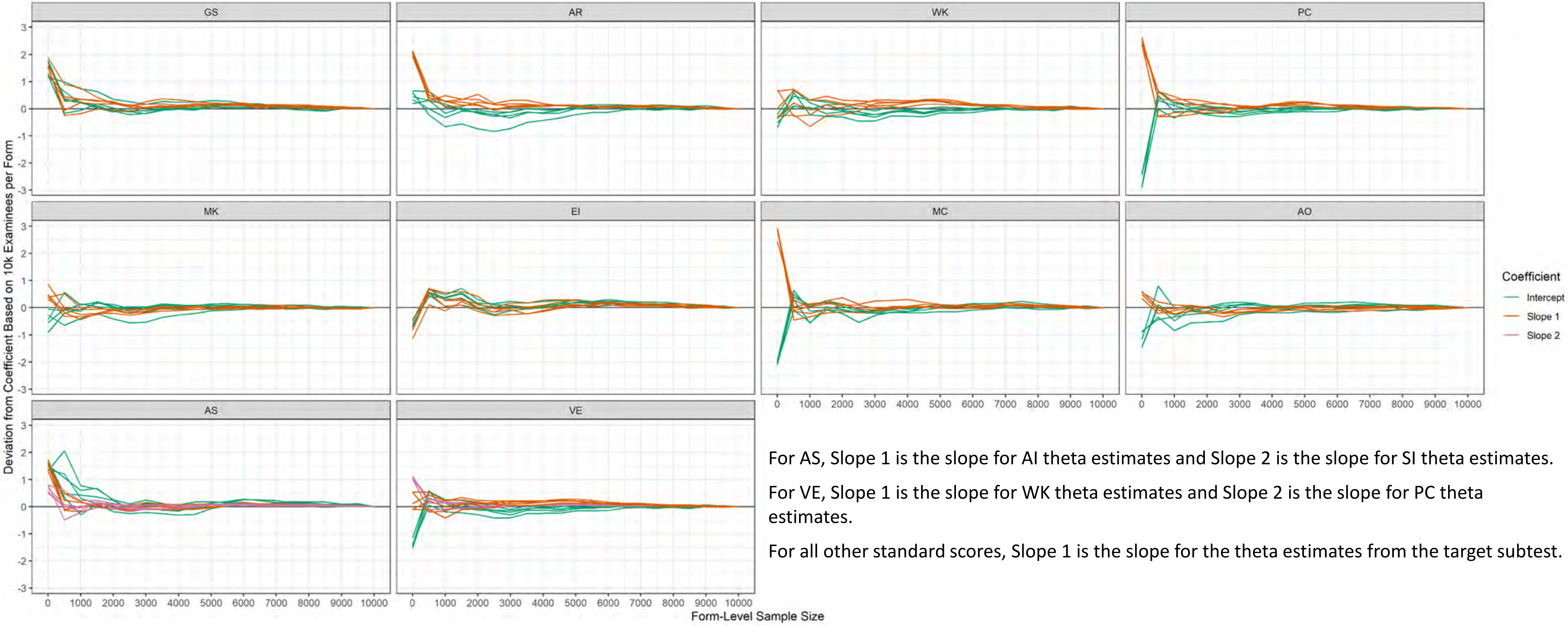
OFFICE OF PEOPLE ANALYTICS

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

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



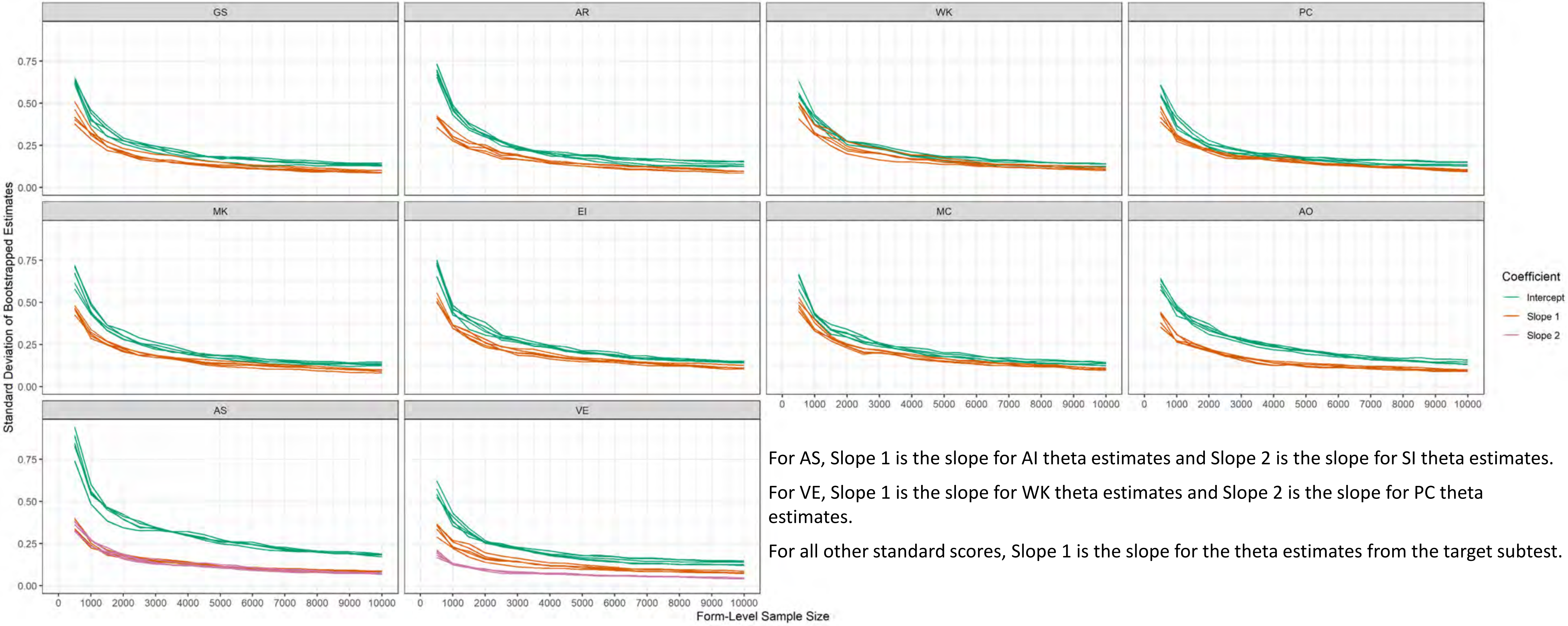
For AS, Slope 1 is the slope for AI theta estimates and Slope 2 is the slope for SI theta estimates.

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.

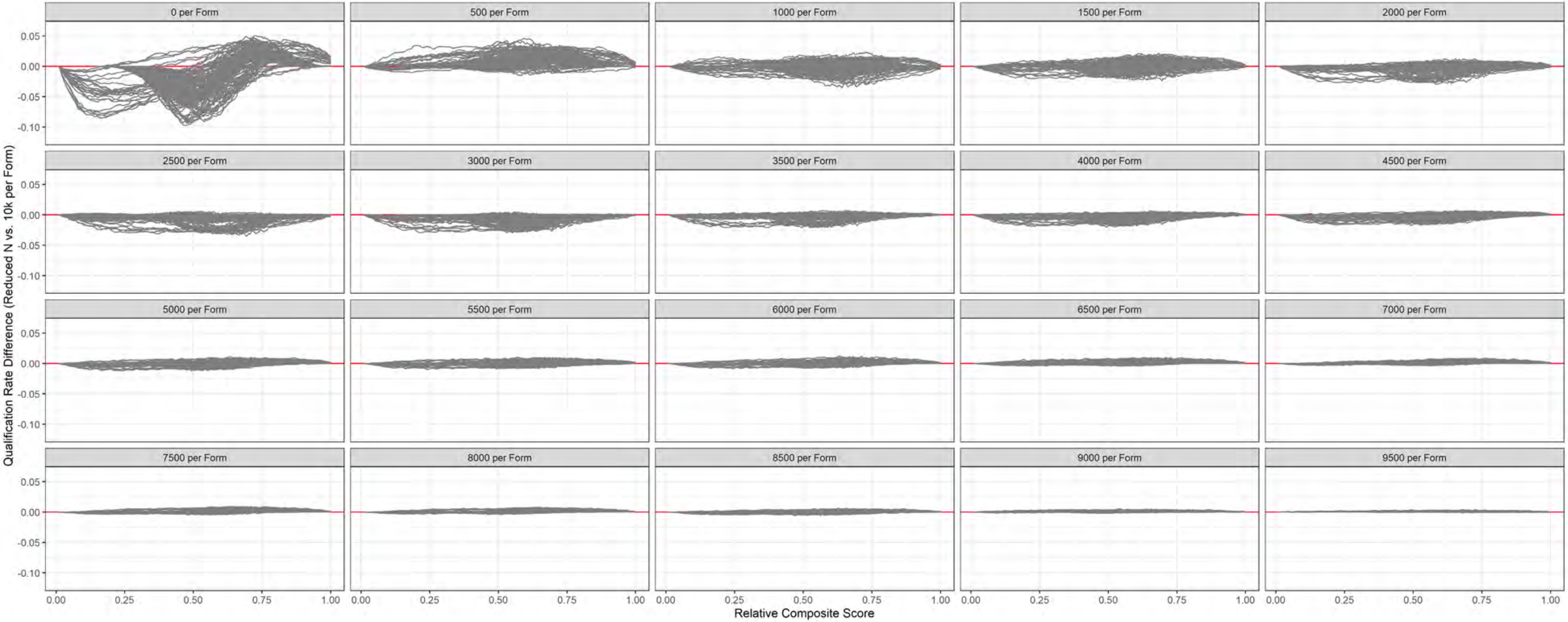
For all other standard scores, Slope 1 is the slope for the theta estimates from the target subtest.

OPA
OFFICE OF PEOPLE ANALYTICS

Qualification Rate Differences: Within-Form Convergence

OFFICE OF PEOPLE ANALYTICS

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

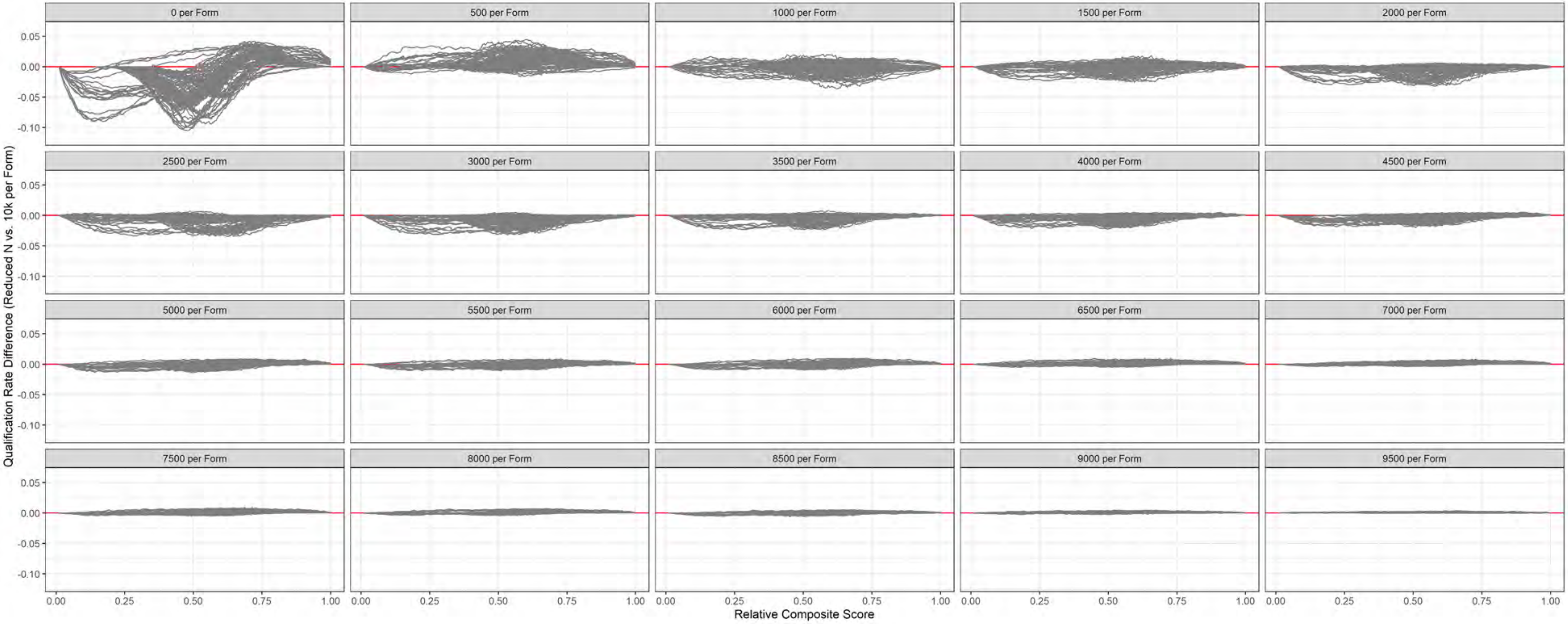


OFFICE OF PEOPLE ANALYTICS

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)

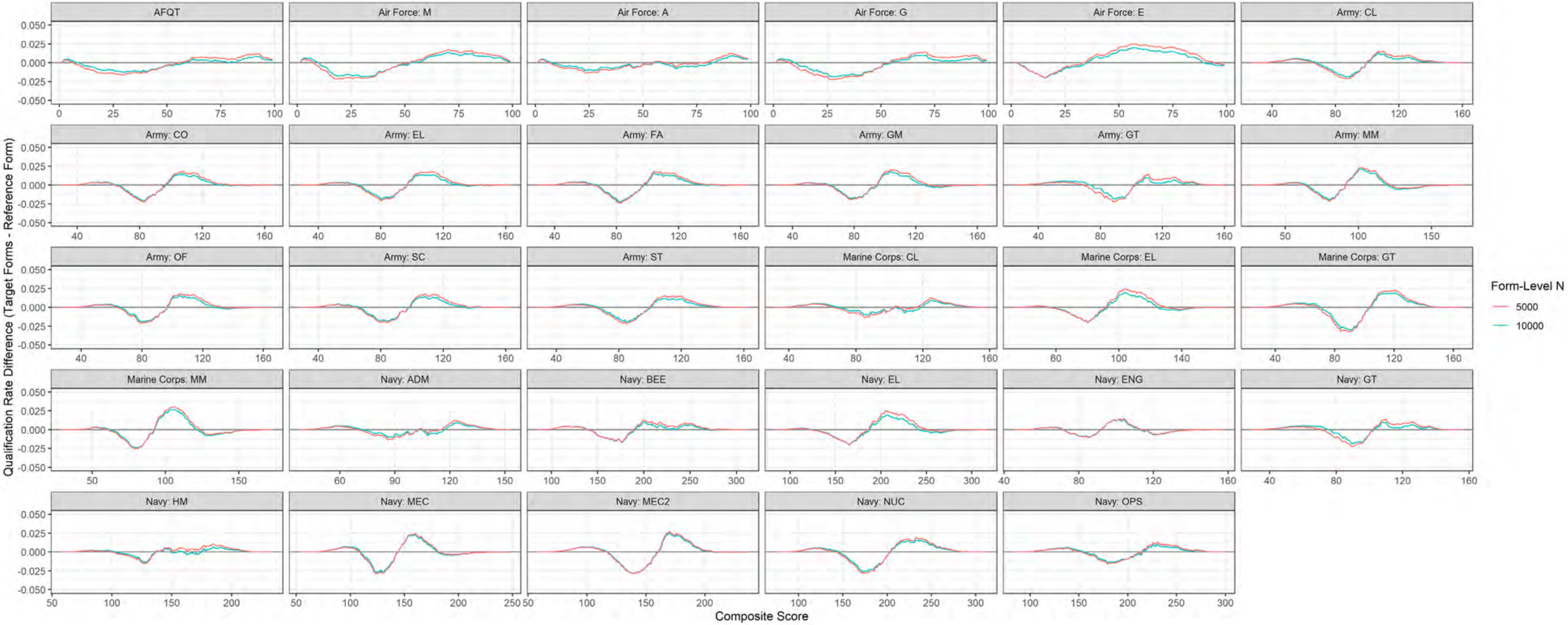


OFFICE OF PEOPLE ANALYTICS

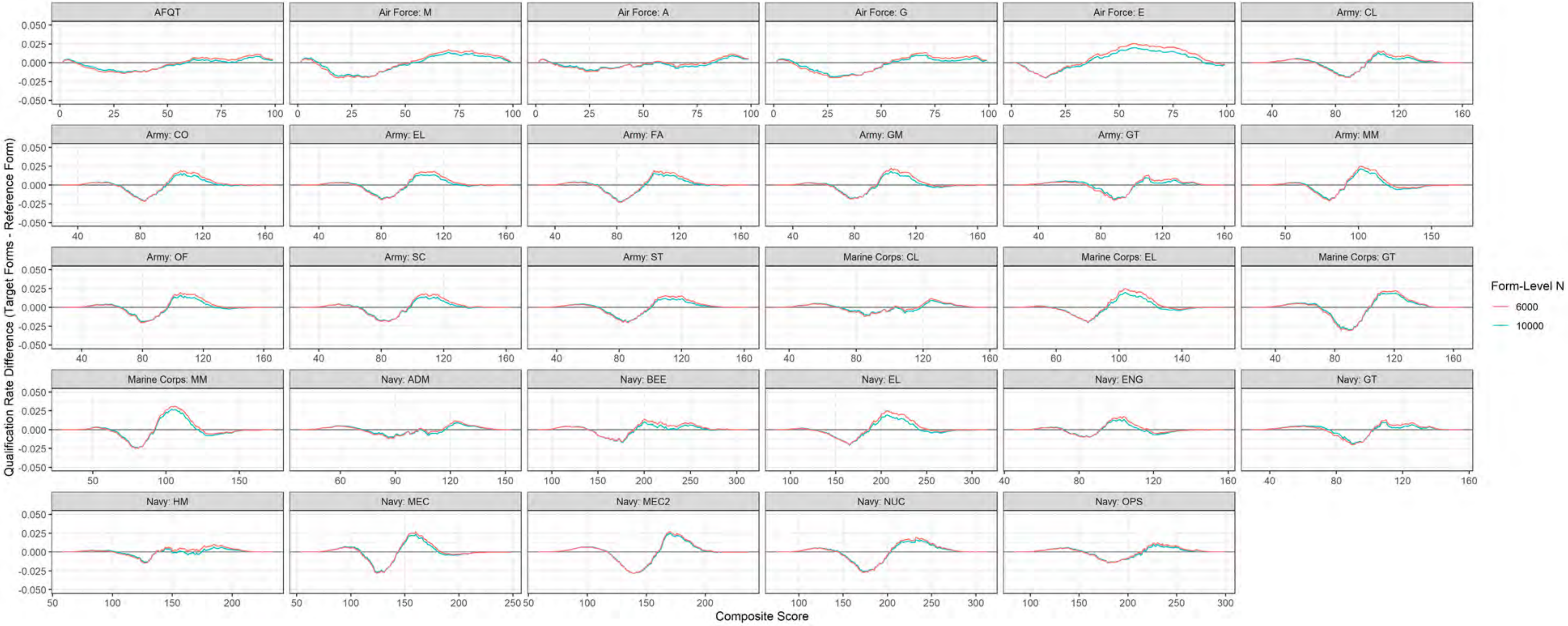
Qualification Rate Differences: Comparison with Reference Form

OPA
OFFICE OF PEOPLE ANALYTICS

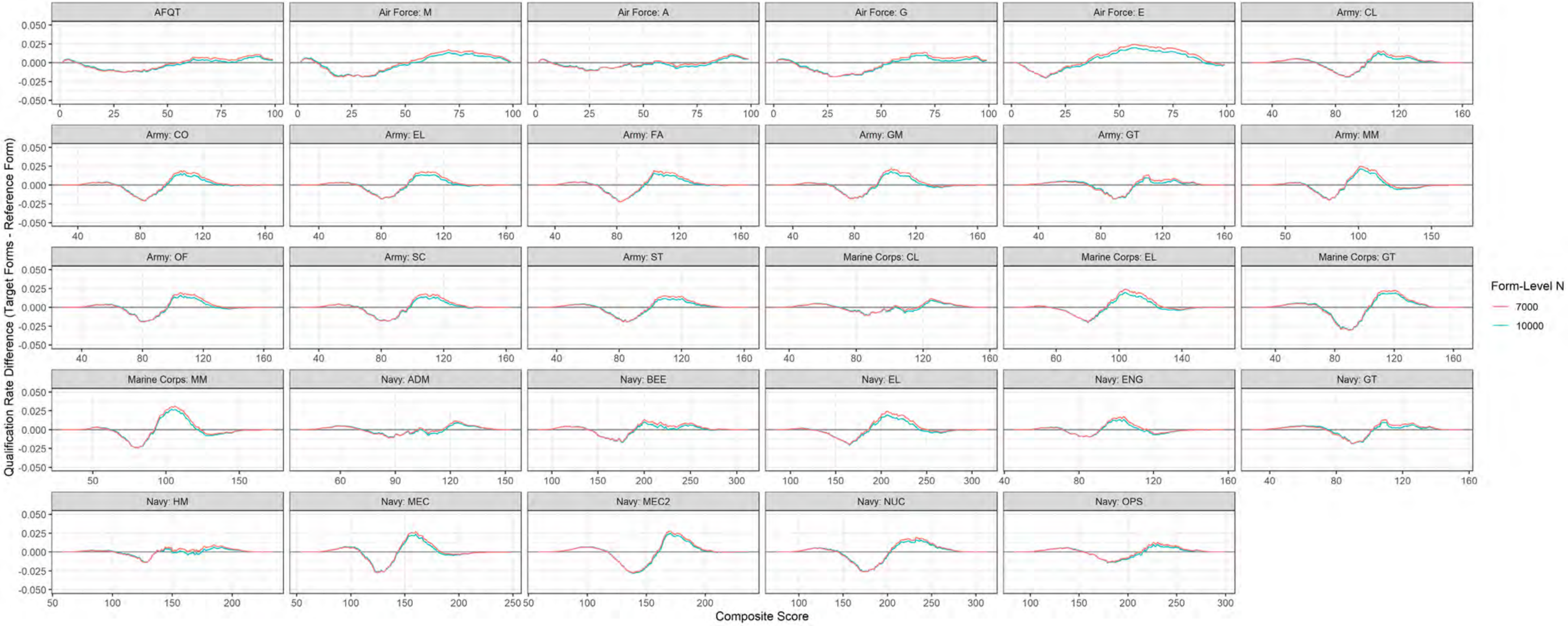
Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 5k vs. *N* = 10k per Form (Equating Sample)



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

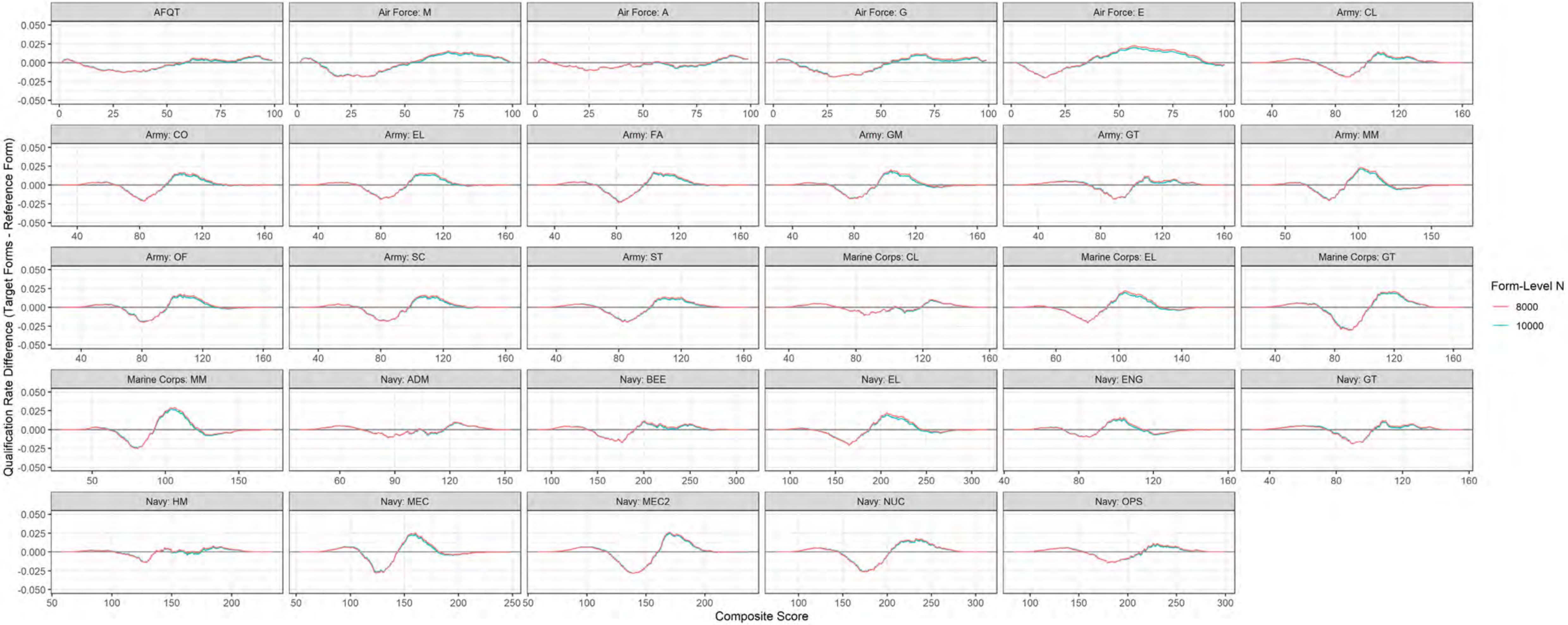


Qualification Rate Differences Relative to the Reference Form for Equating Conditions with *N* = 7k vs. *N* = 10k per Form (Equating Sample)



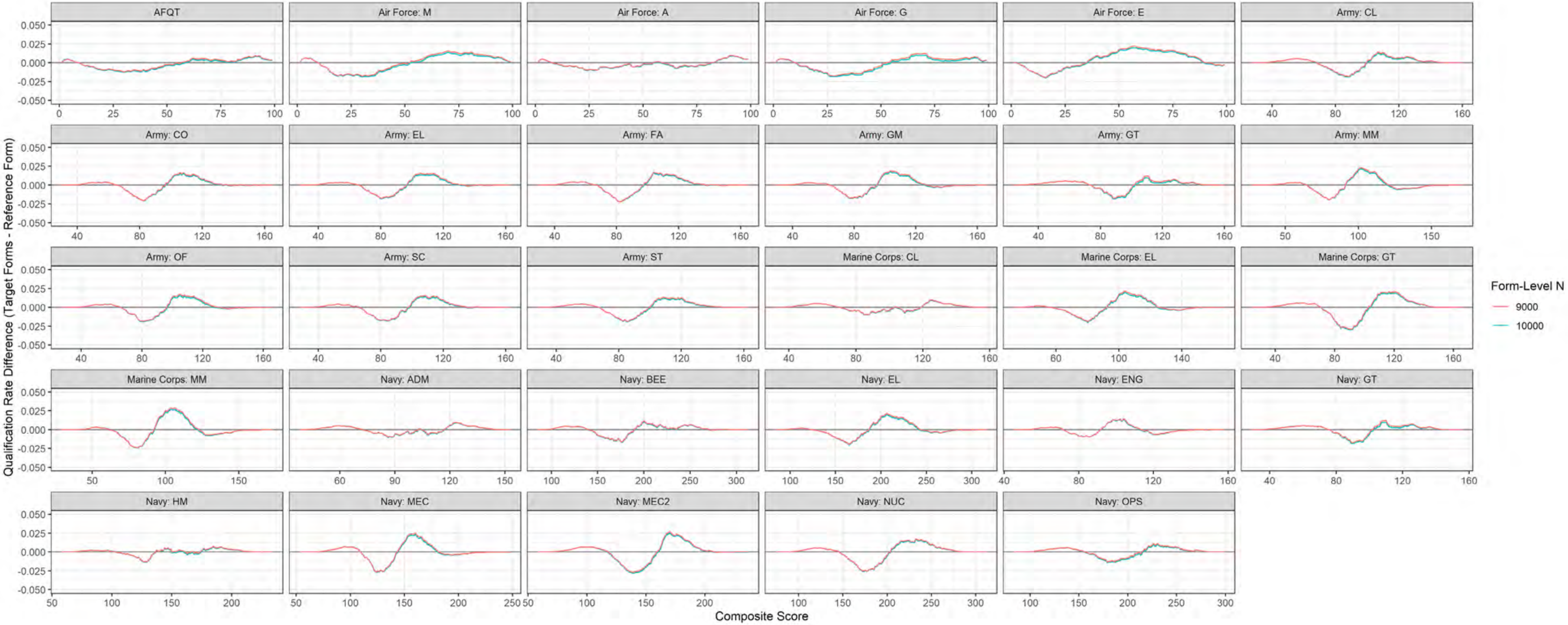
OFFICE OF PEOPLE ANALYTICS

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)



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

Evaluation Strategy

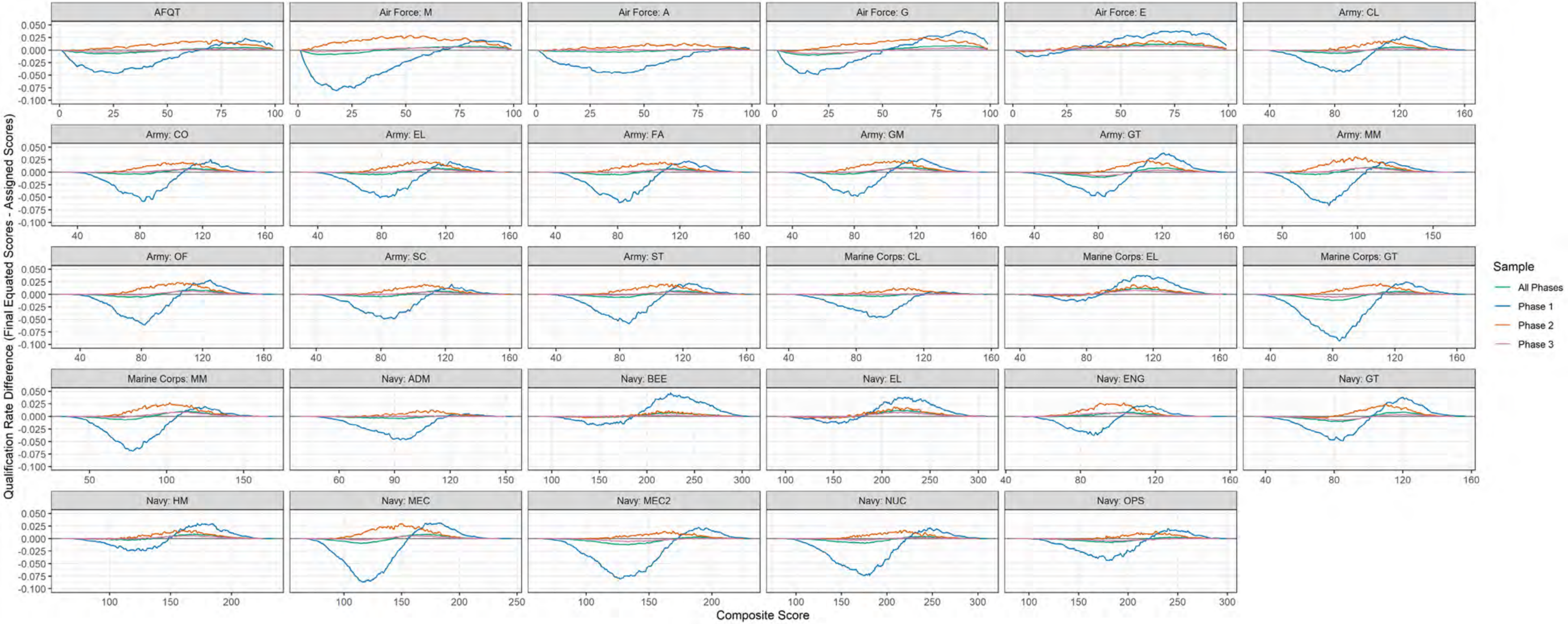
- Each sample was constructed by selecting examinees from the equating data set from CAT-ASVAB Forms 11–15 in the order their results were recorded
- We used a series of four sequential evaluations to identify a recommended configuration for future equating studies:
  1. Using a final form-level sample 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

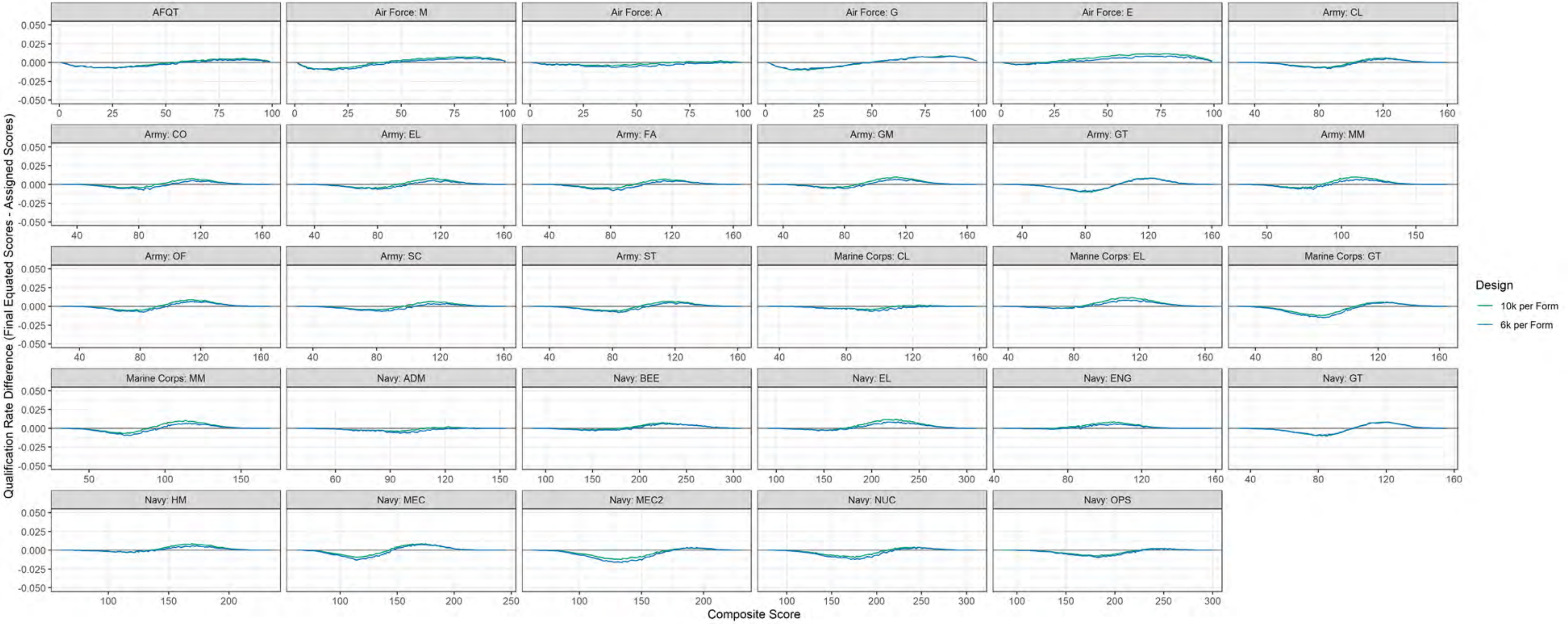
- This accounts for the additional testing that occurs while temporary equating solutions are being computed, replicated, implemented, and released
  - E.g., although the current Phase 1 *N* is 500 per form, the processing lag in our analyses means 500 additional people take each form before the provisional 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

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



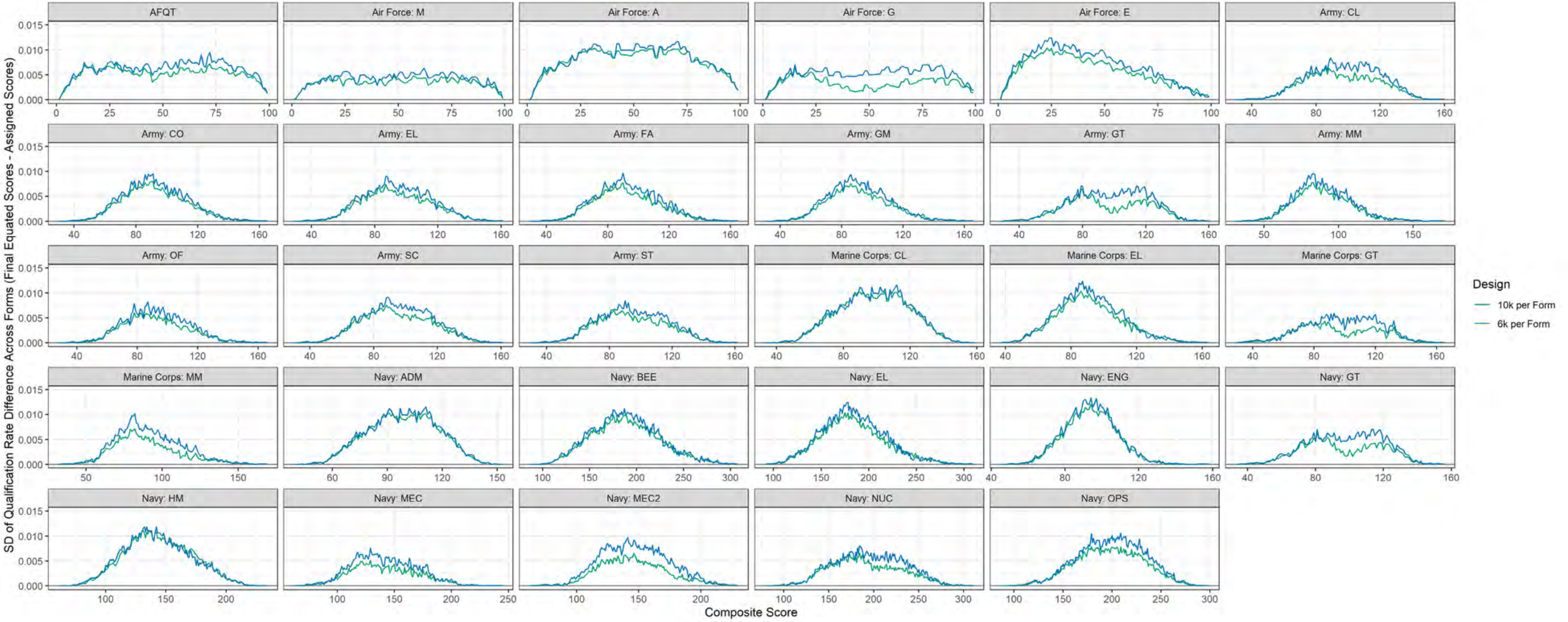
Note: Phase-specific samples are non-cumulative in this figure.

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



OFFICE OF PEOPLE ANALYTICS

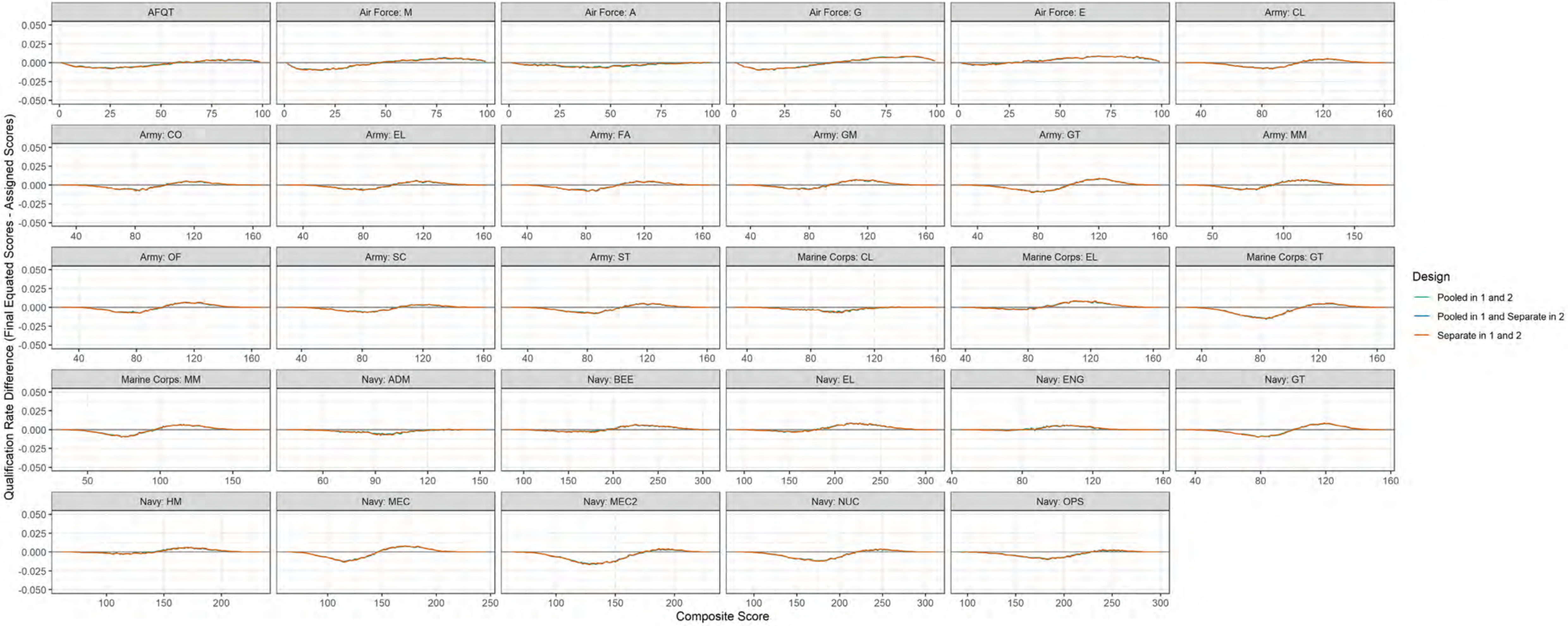
Evaluation 1: Using Final Form-Level *N* of 10k vs. 6k (Standard Deviations of Qualification Rate Differences Across Forms)



Evaluation 1 Winner: 6k Examinees per Form

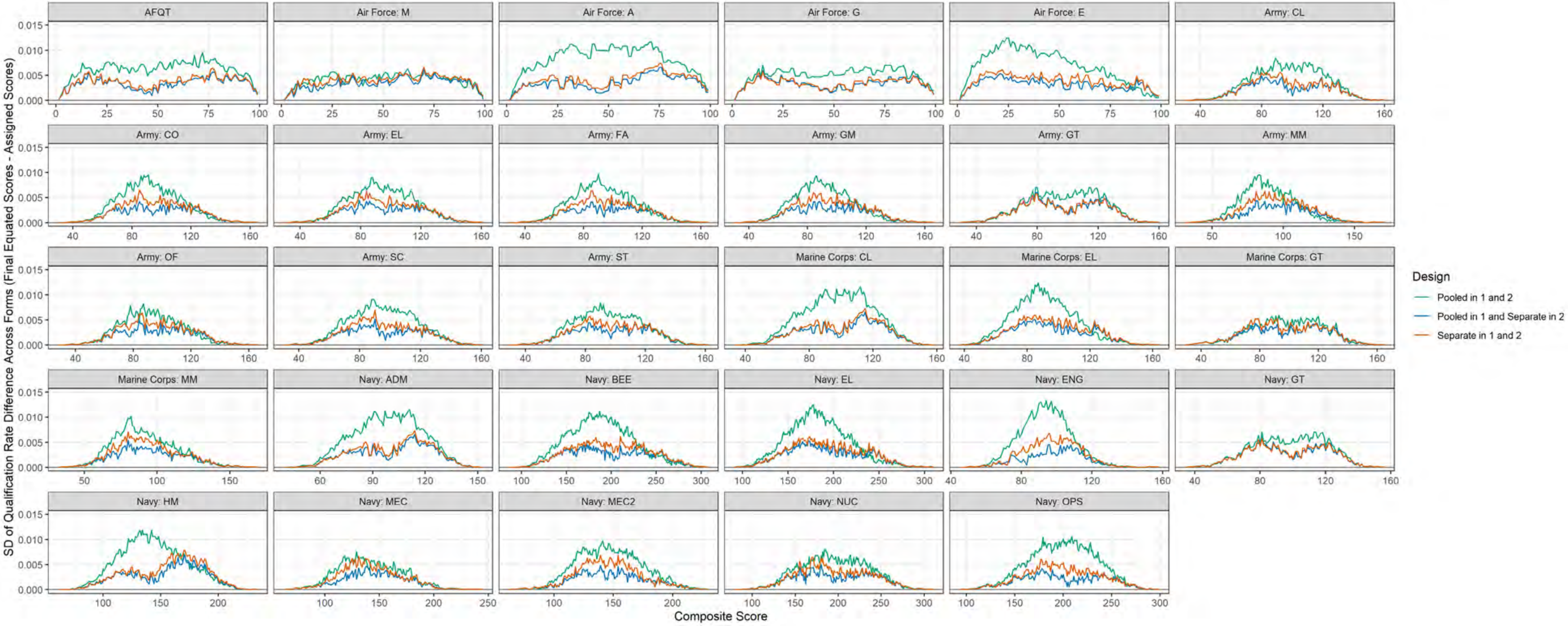
- Allows a substantial reduction in the duration of an equating study
- Has minimal impact on the overall quality of examinees’ scores

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)

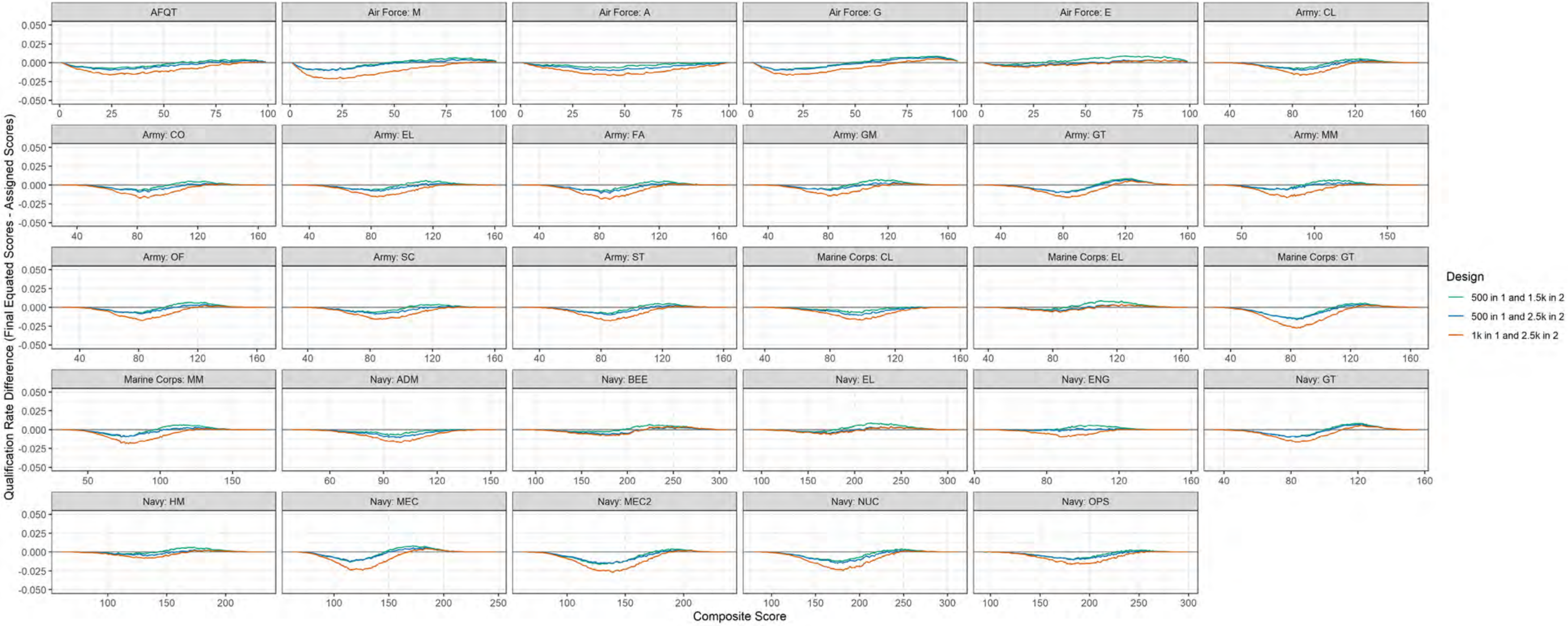


OFFICE OF PEOPLE ANALYTICS

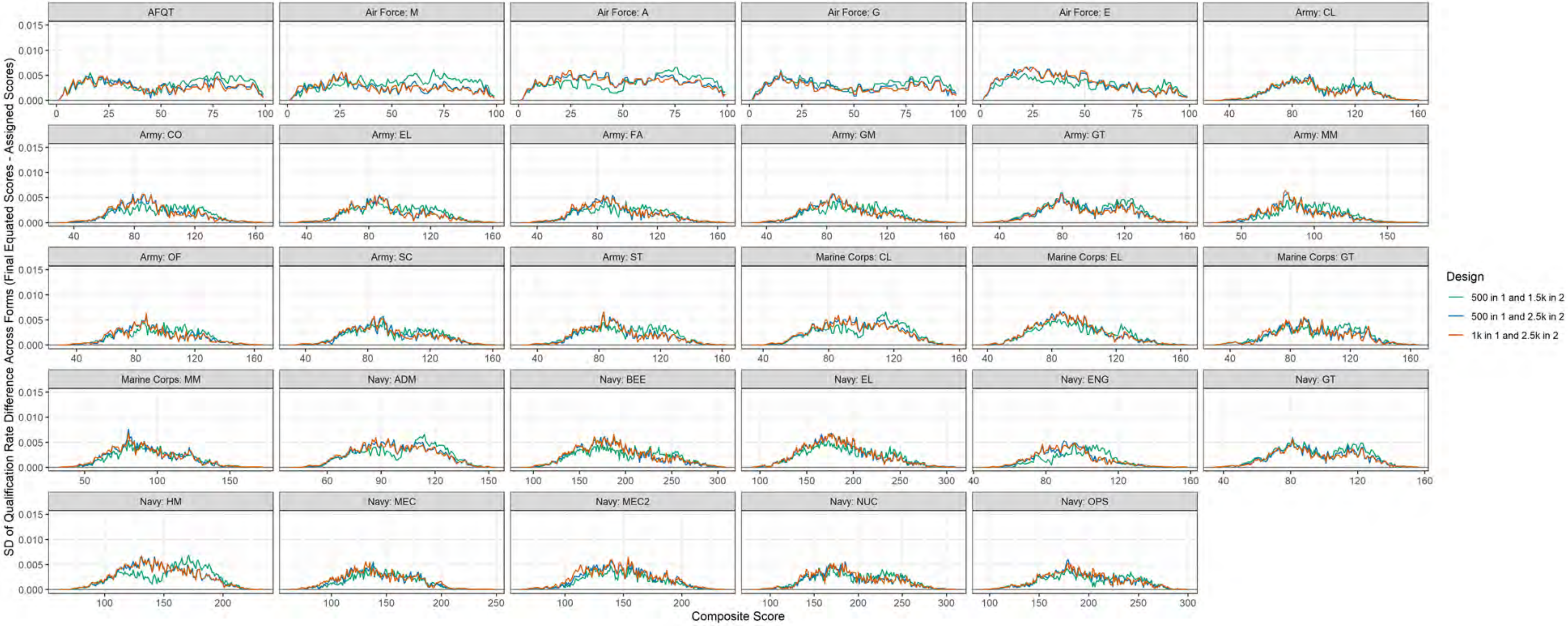
Evaluation 2 Winner: Pooled Equating in Phase 1 with Separate Equating in Phase 2

- 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

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



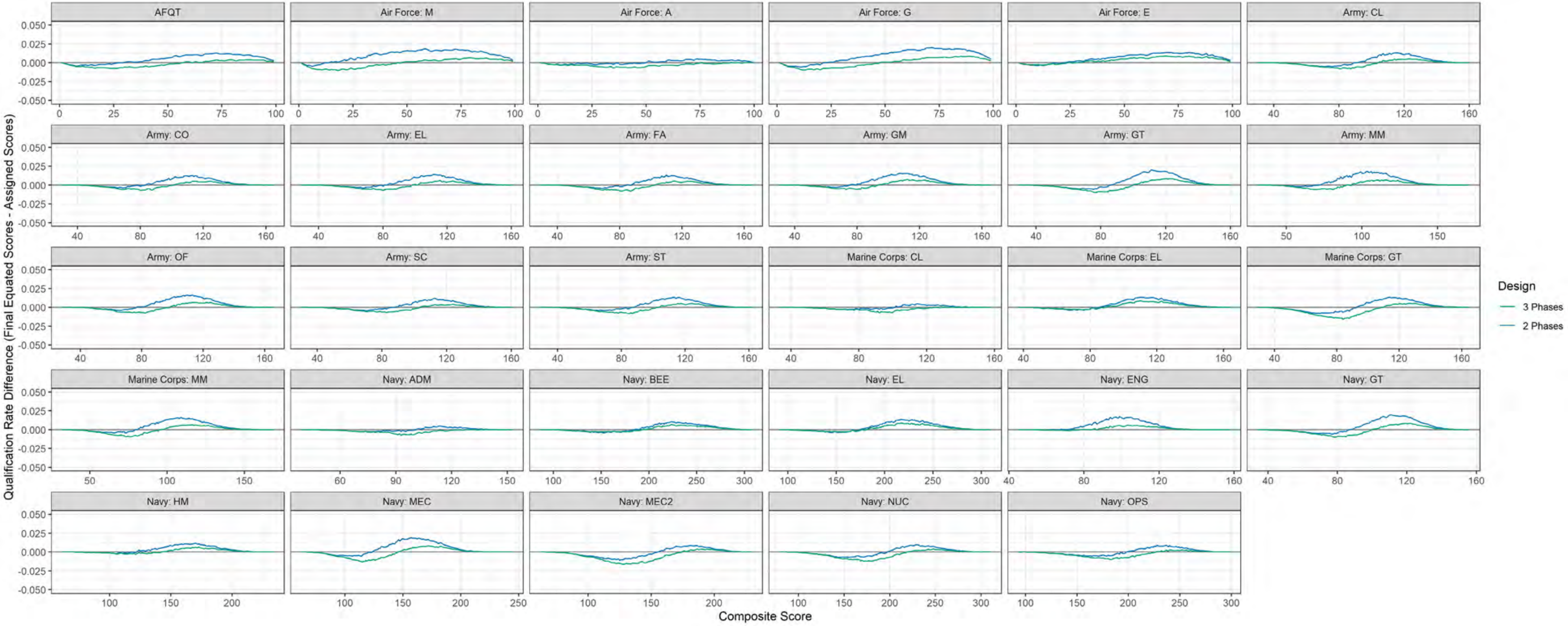
Evaluation 3: Using Existing Early-Phase *N*s vs. Increased *N*s (Standard Deviations of Qualification Rate Differences Across Forms)



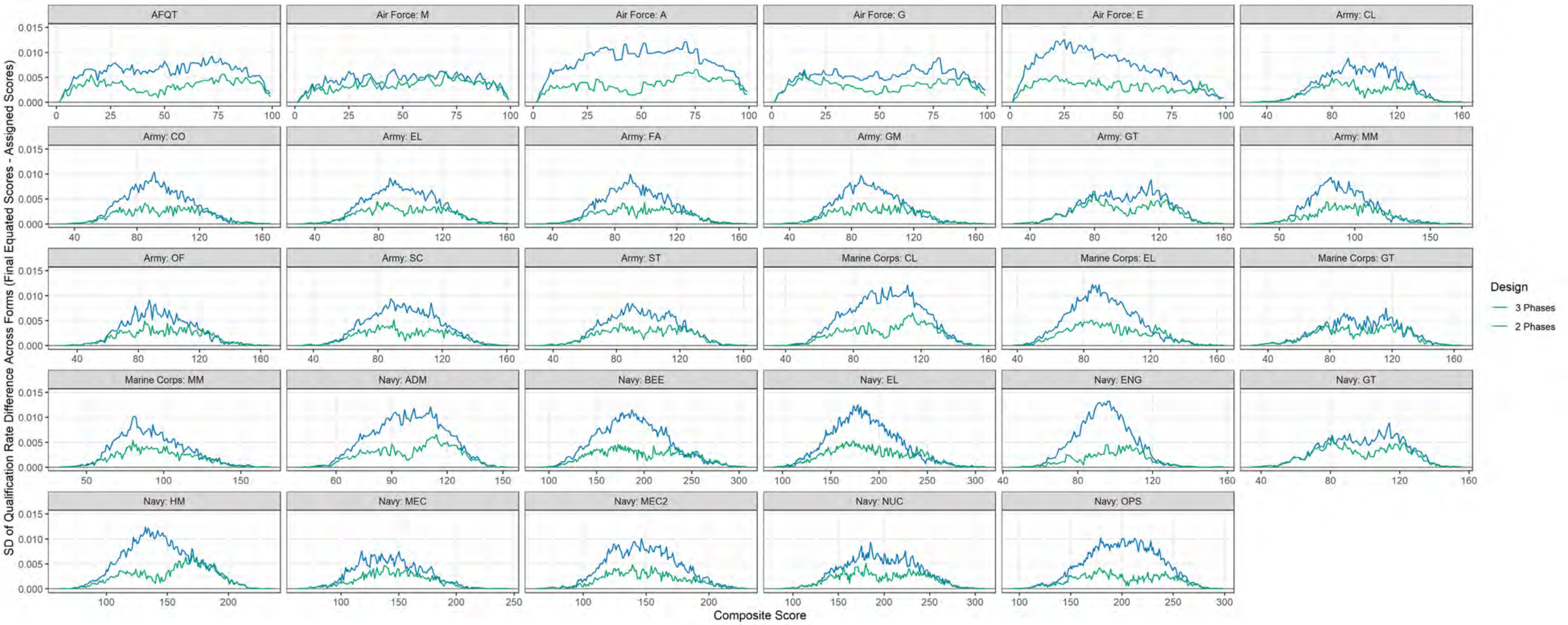
Evaluation 3 Winner: *N* = 500 per Form in Phase 1 and *N* = 1,500 per Form in Phase 2

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

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



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



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

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

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)

Thank You!

For more information,
please contact:

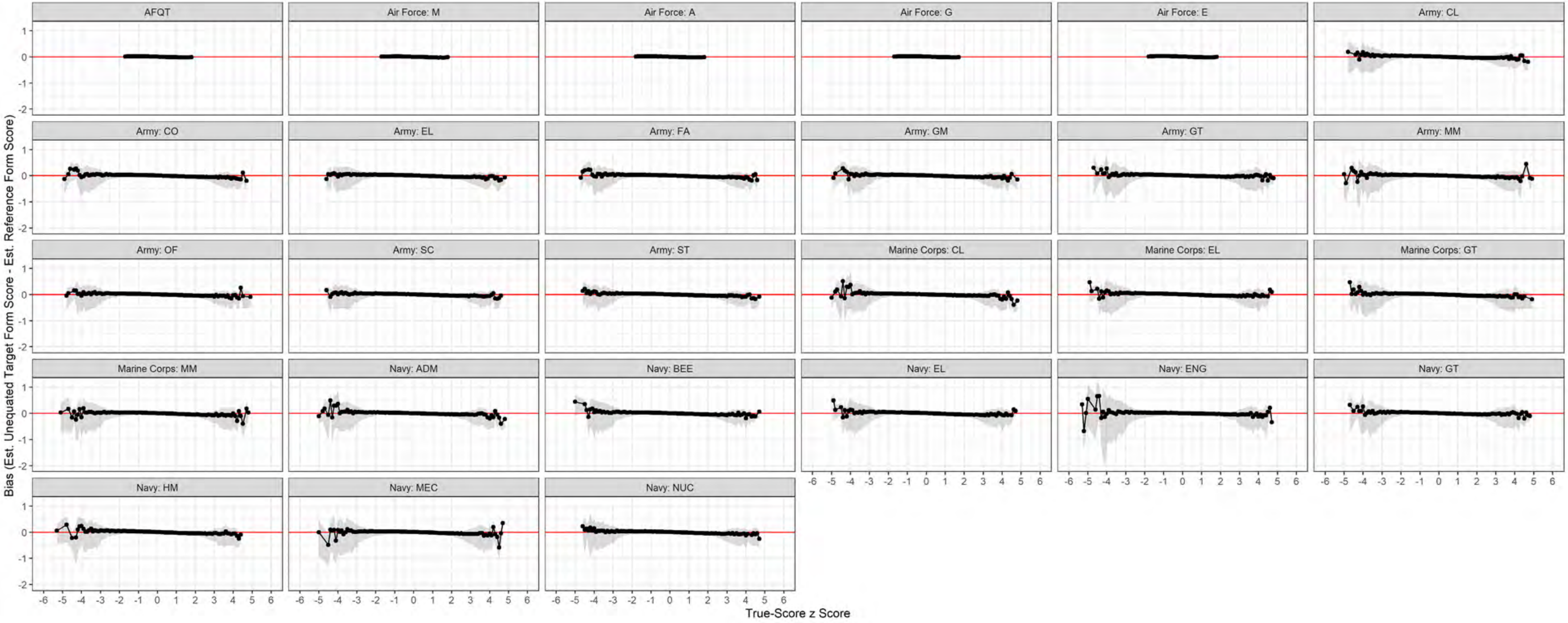
Jeff Dahlke
jdahlke@humrro.org
jeffrey.a.dahlke.ctr@mail.mil

OPA
OFFICE OF PEOPLE ANALYTICS

Supplemental Slides: Simulation-Based Evaluation of Unequated Scores

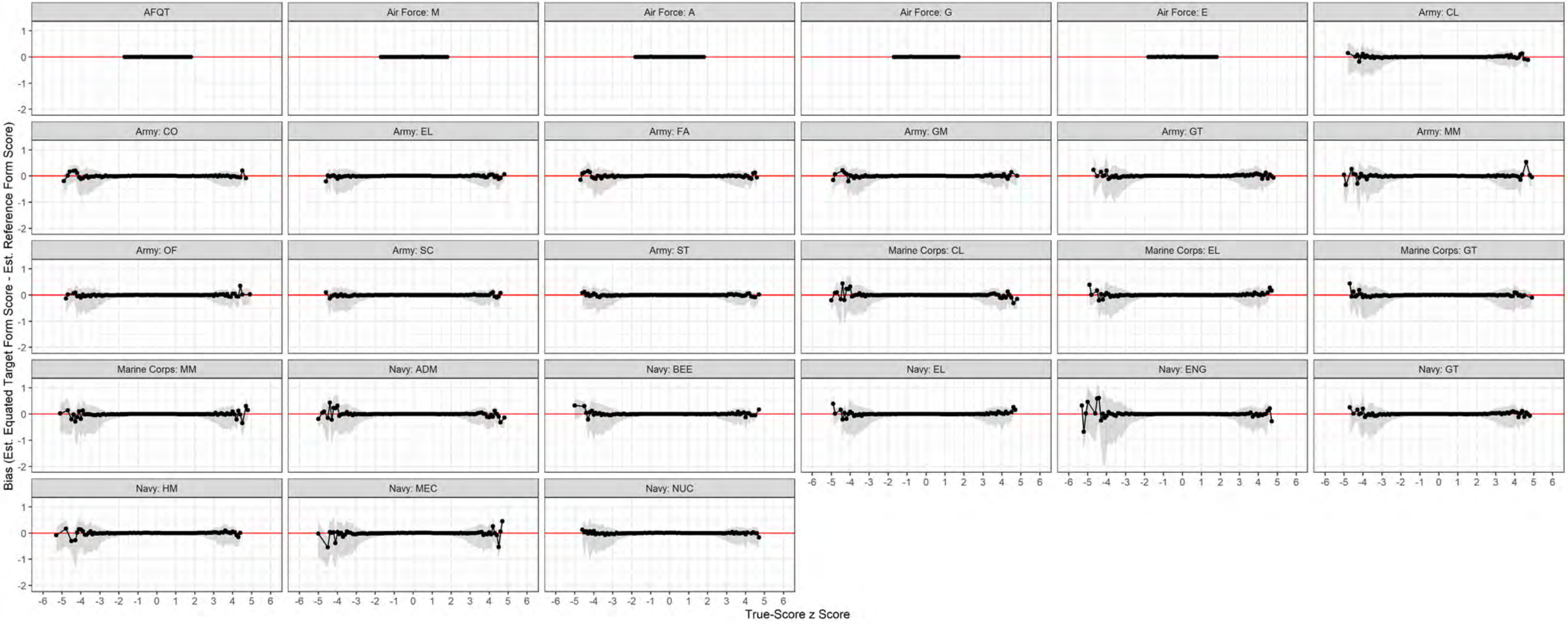
OPA
OFFICE OF PEOPLE ANALYTICS

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



Note. Error ribbons represent 95% confidence intervals.

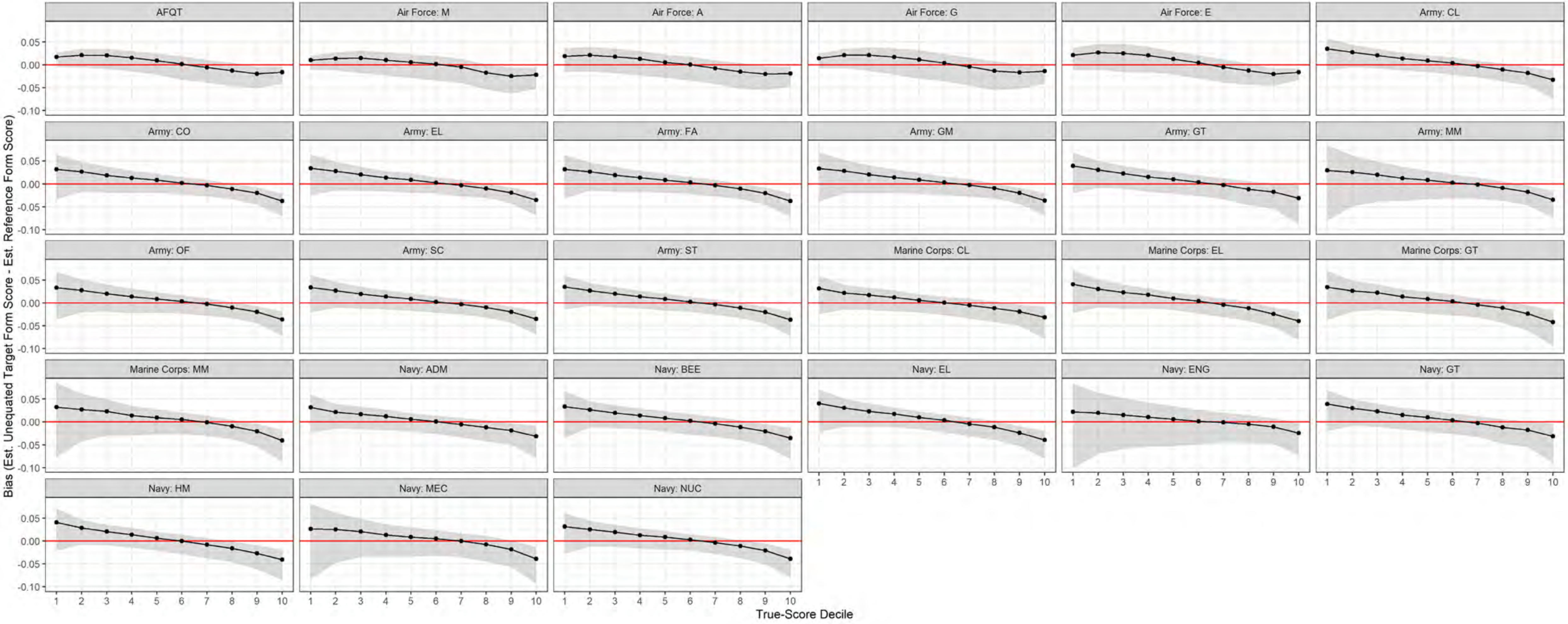
Evaluation of Equated Composite Score Bias by True-Score *z* Score



OFFICE OF PEOPLE ANALYTICS

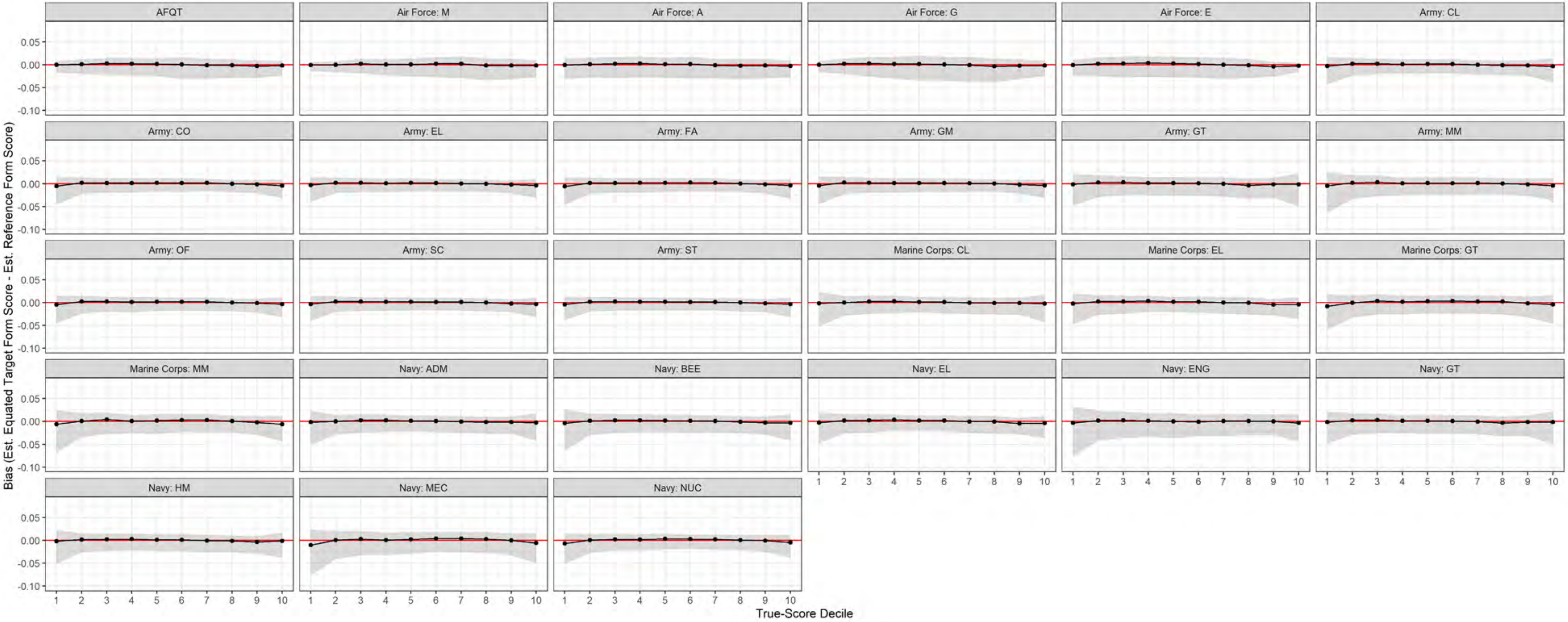
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Unequated Composite Score Bias by True-Score Decile



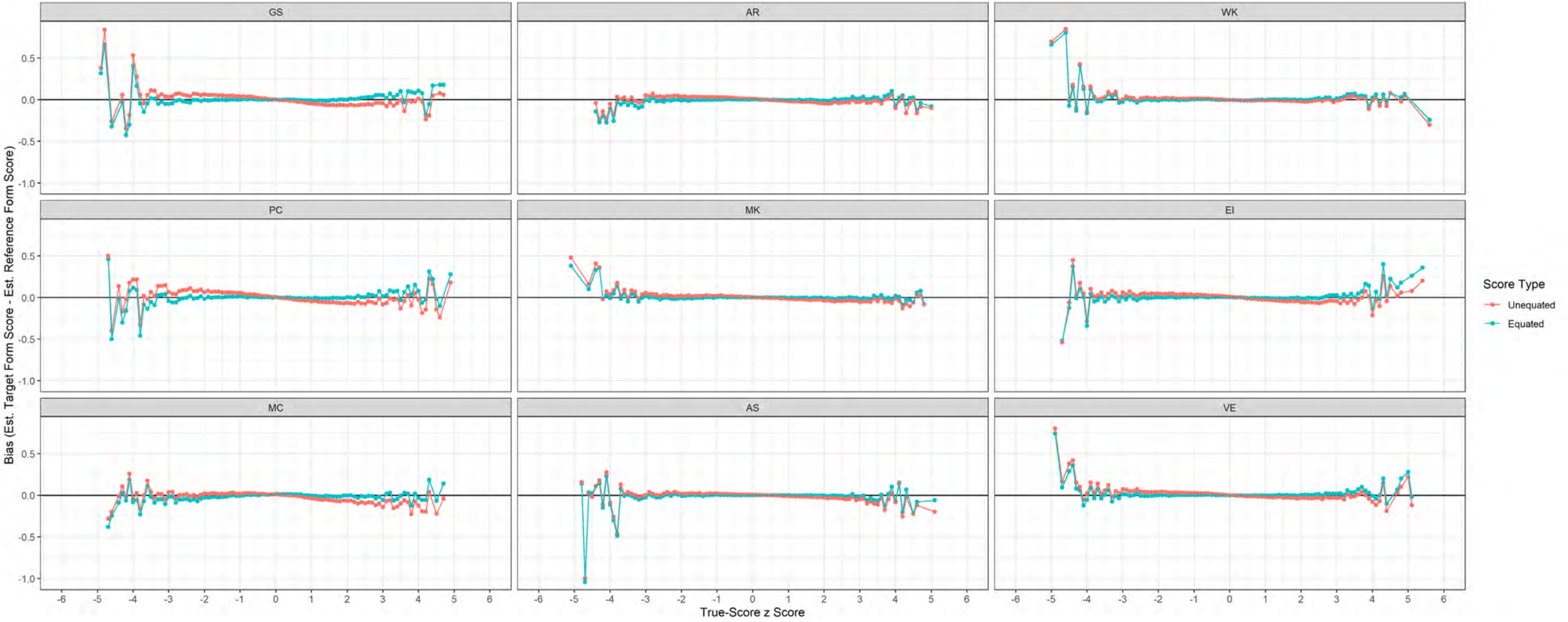
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Equated Composite Score Bias by True-Score Decile



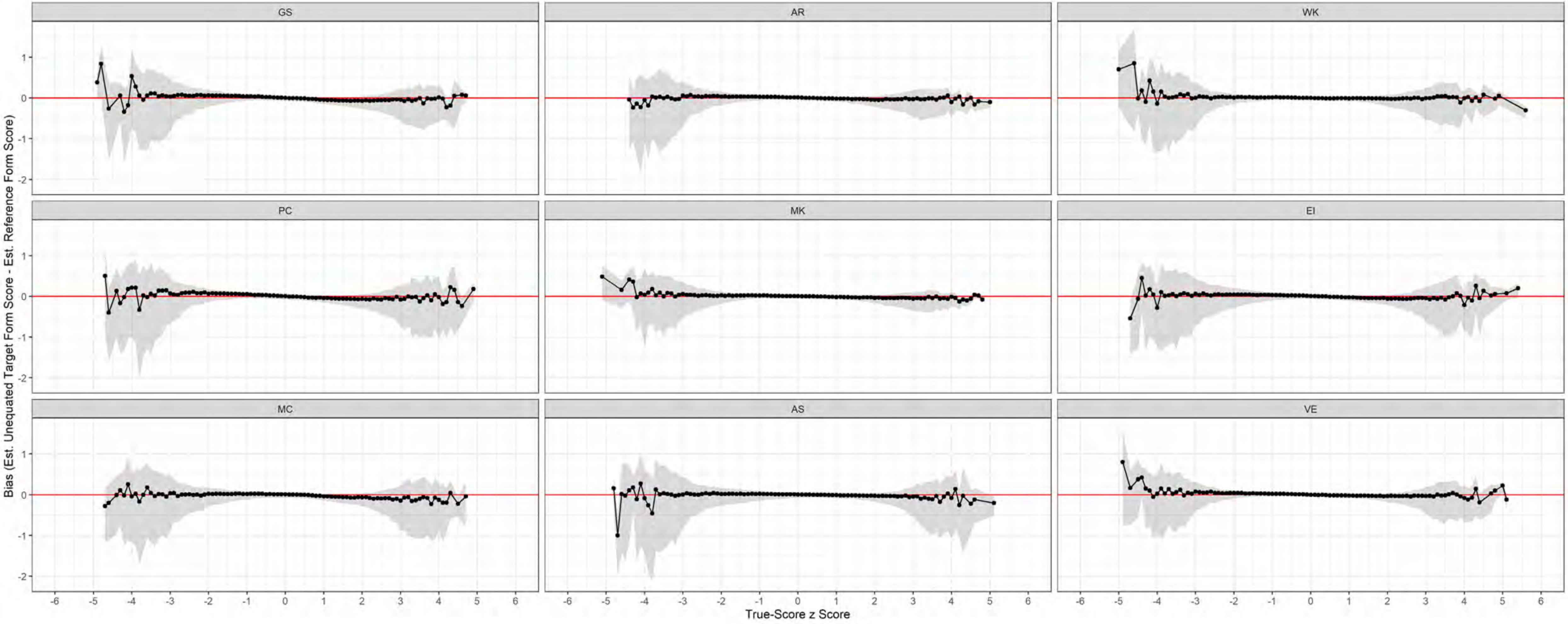
Note. Error ribbons represent 95% confidence intervals.

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



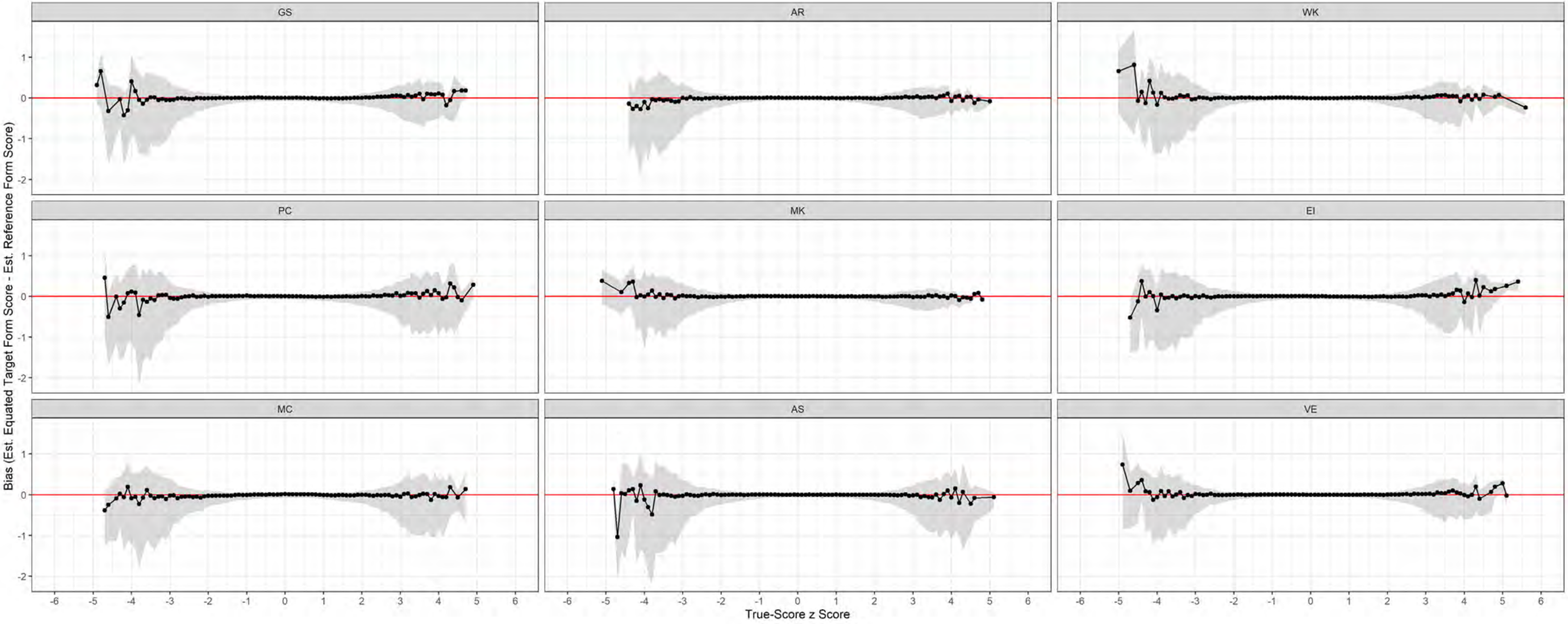
OFFICE OF PEOPLE ANALYTICS

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



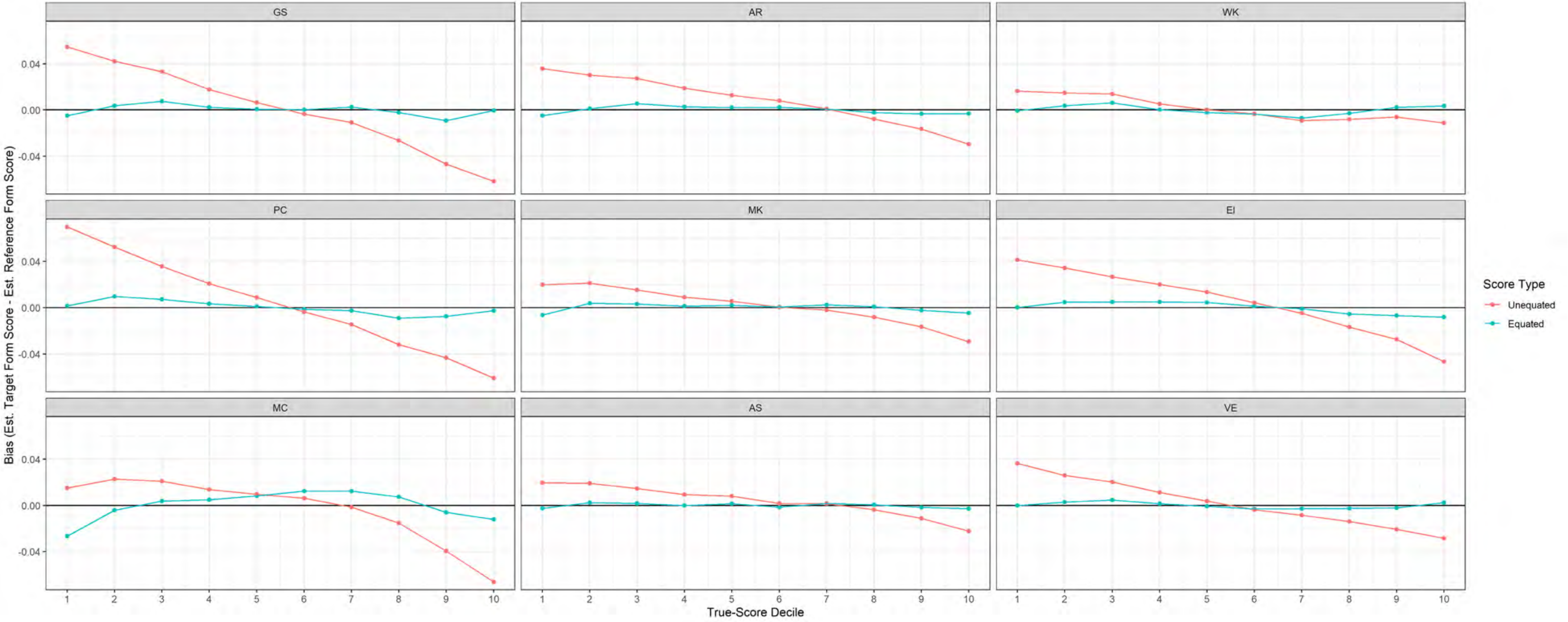
Note. Error ribbons represent 95% confidence intervals.

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



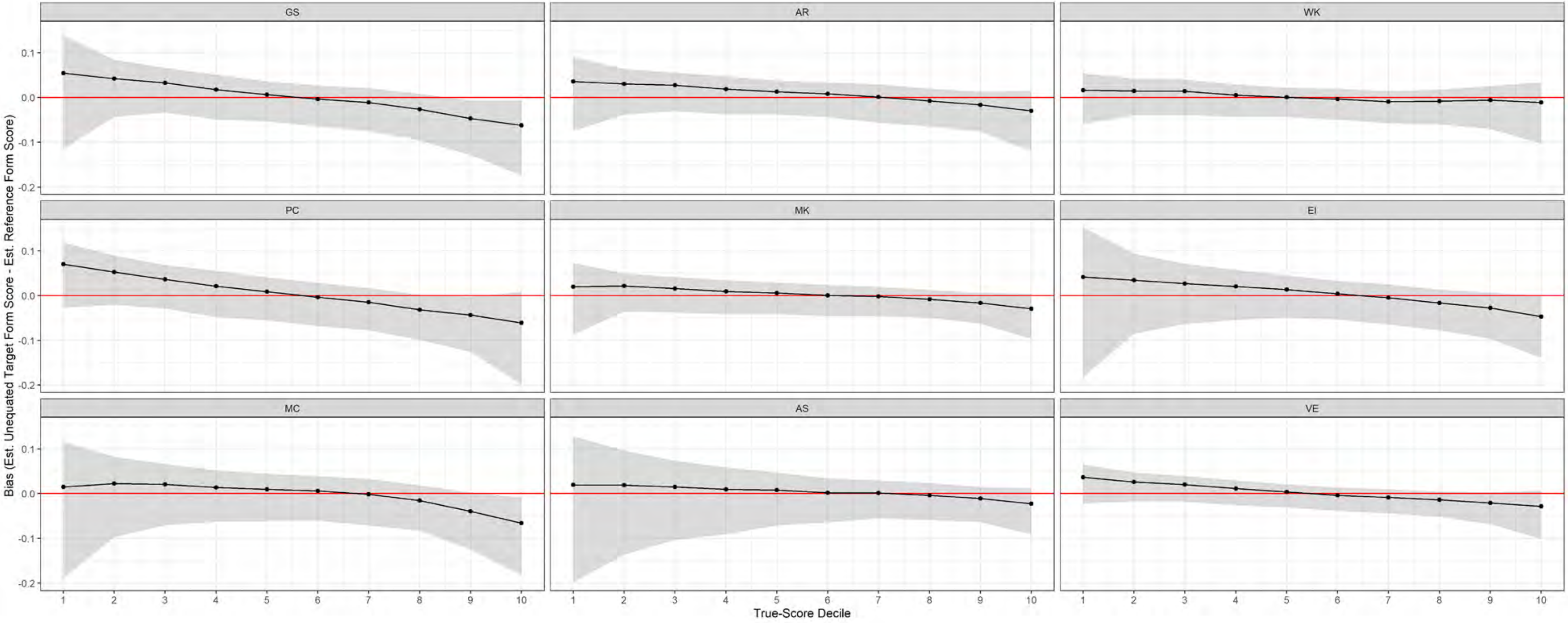
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Standard Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

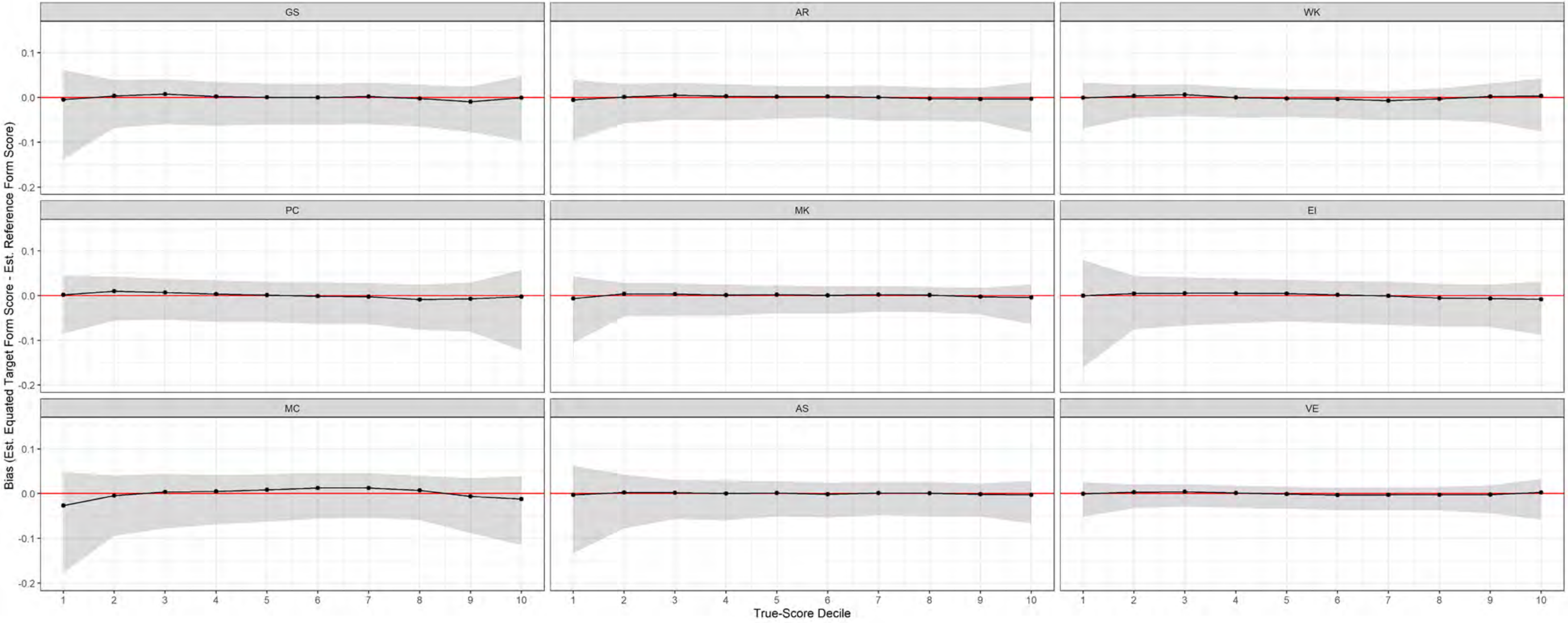
Evaluation of Unequated Standard Score Bias by True-Score Decile



OPA
OFFICE OF PEOPLE ANALYTICS

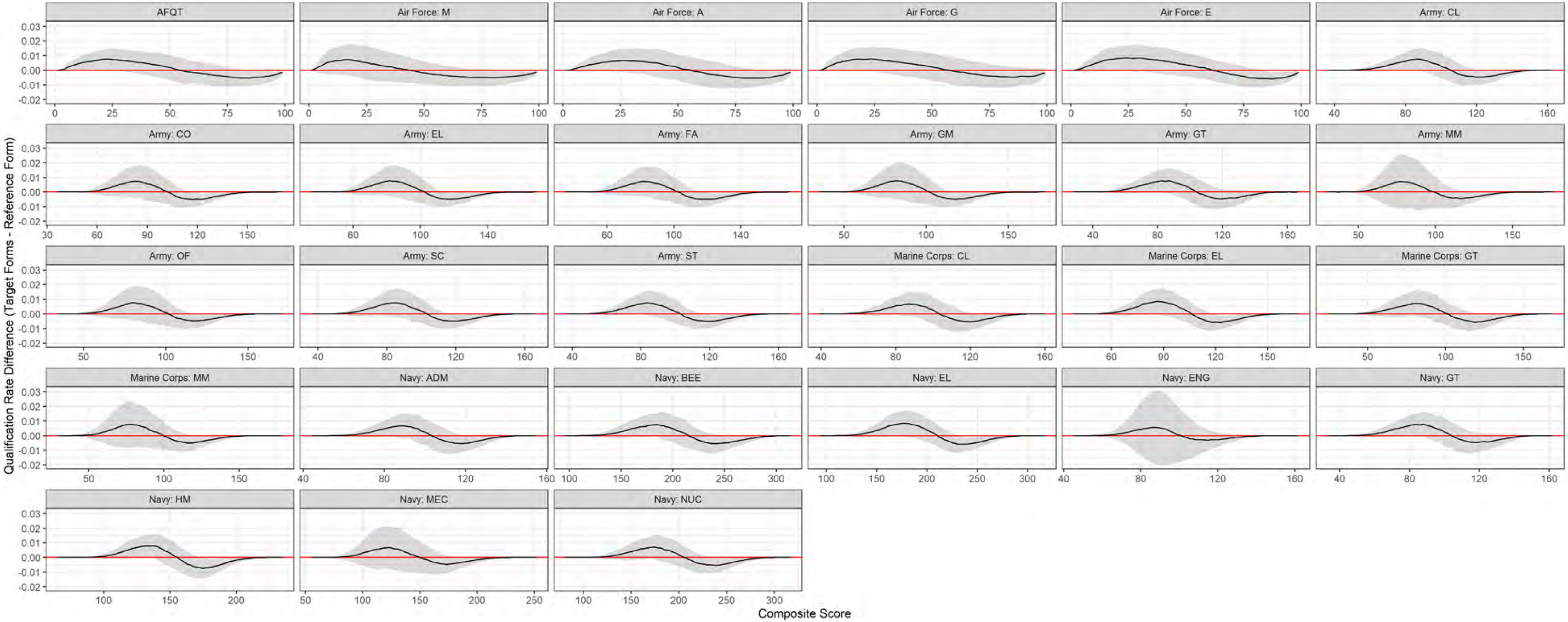
Note. Error ribbons represent 95% confidence intervals.

Evaluation of Equated Standard Score Bias by True-Score Decile



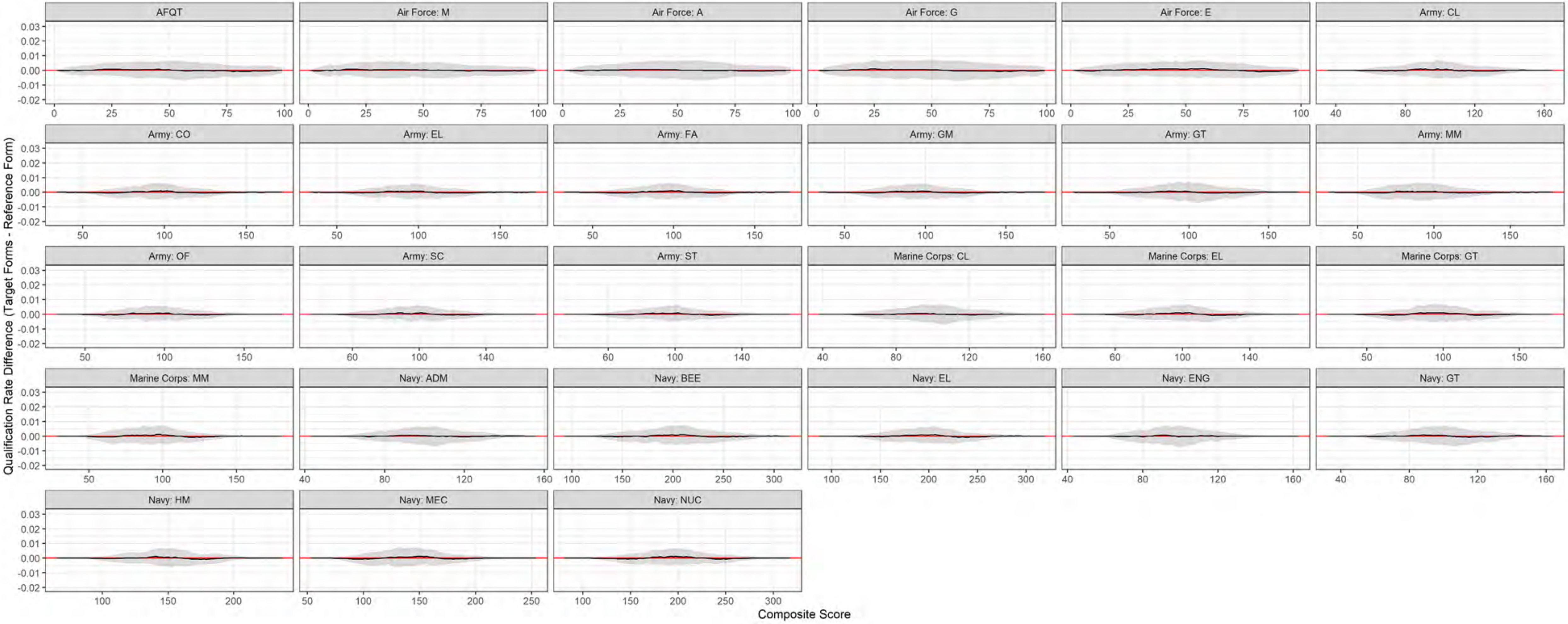
Note. Error ribbons represent 95% confidence intervals.

Qualification Rate Differences for Unequated Composite Scores



Note. Error ribbons represent 95% confidence intervals.

Qualification Rate Differences for Equated Composite Scores

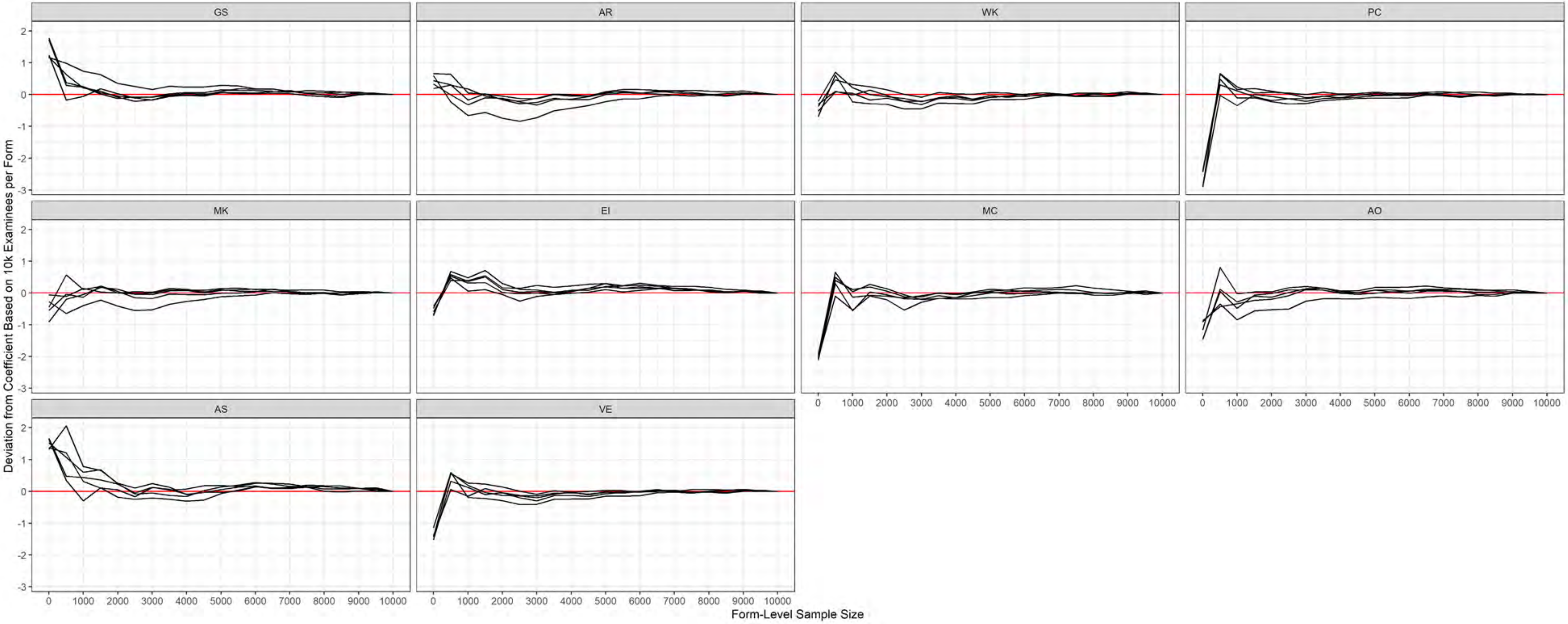


Note. Error ribbons represent 95% confidence intervals.

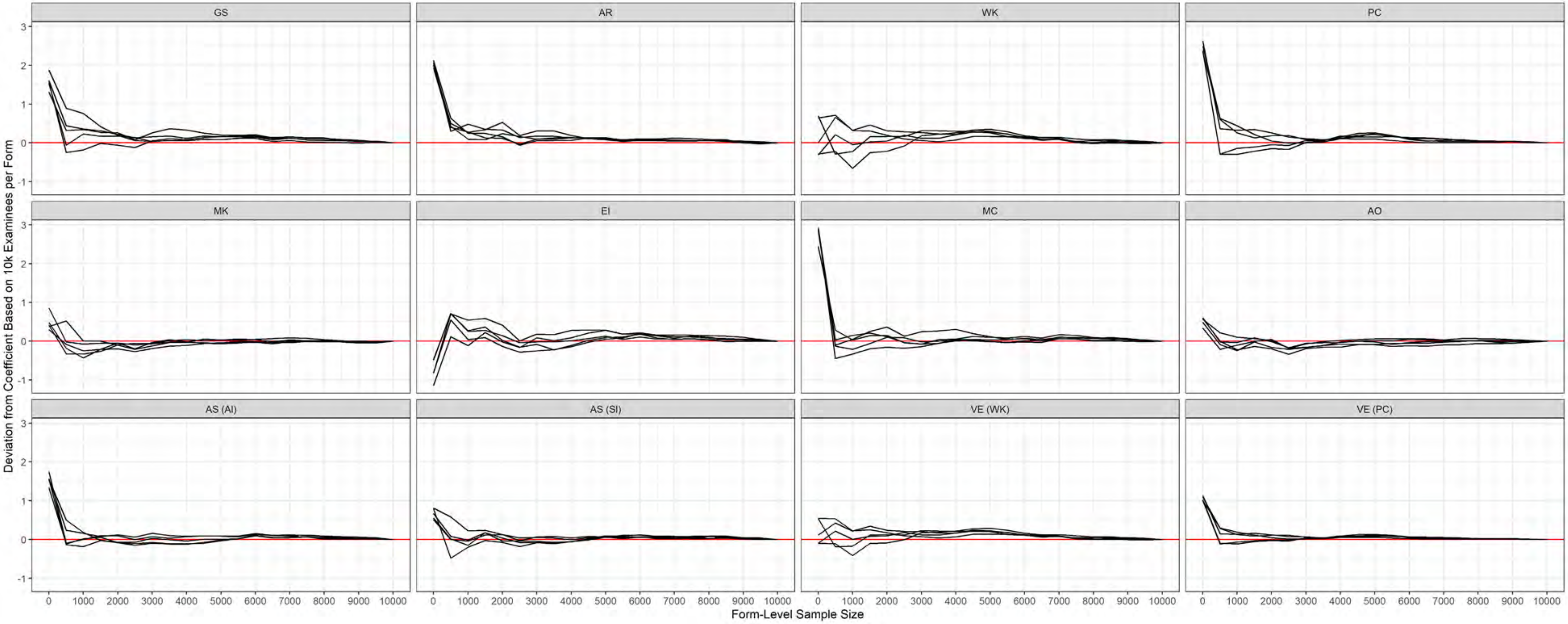
Supplemental Slides:
Plots of TC Convergence and Sampling Error per TC Coefficient

OPA
OFFICE OF PEOPLE ANALYTICS

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

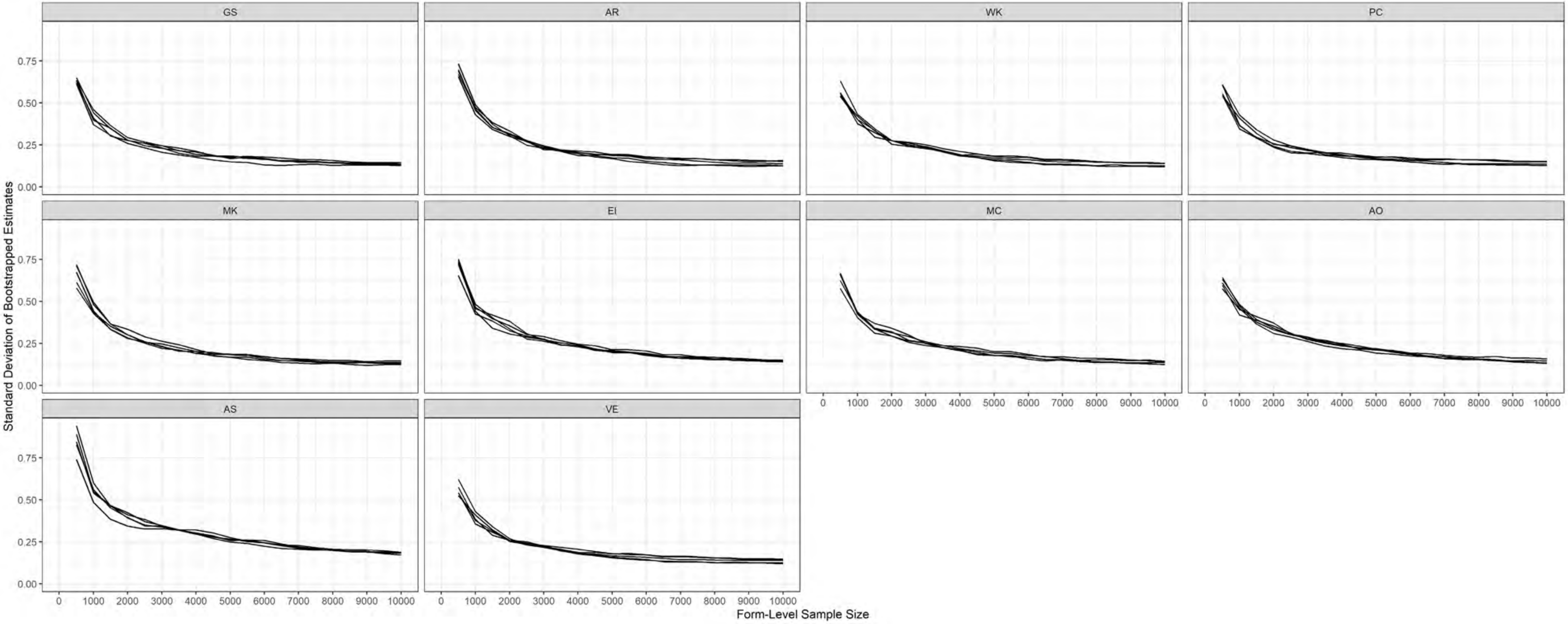


TC Convergence with N_{Form}= 10k Solution for Slope Coefficients



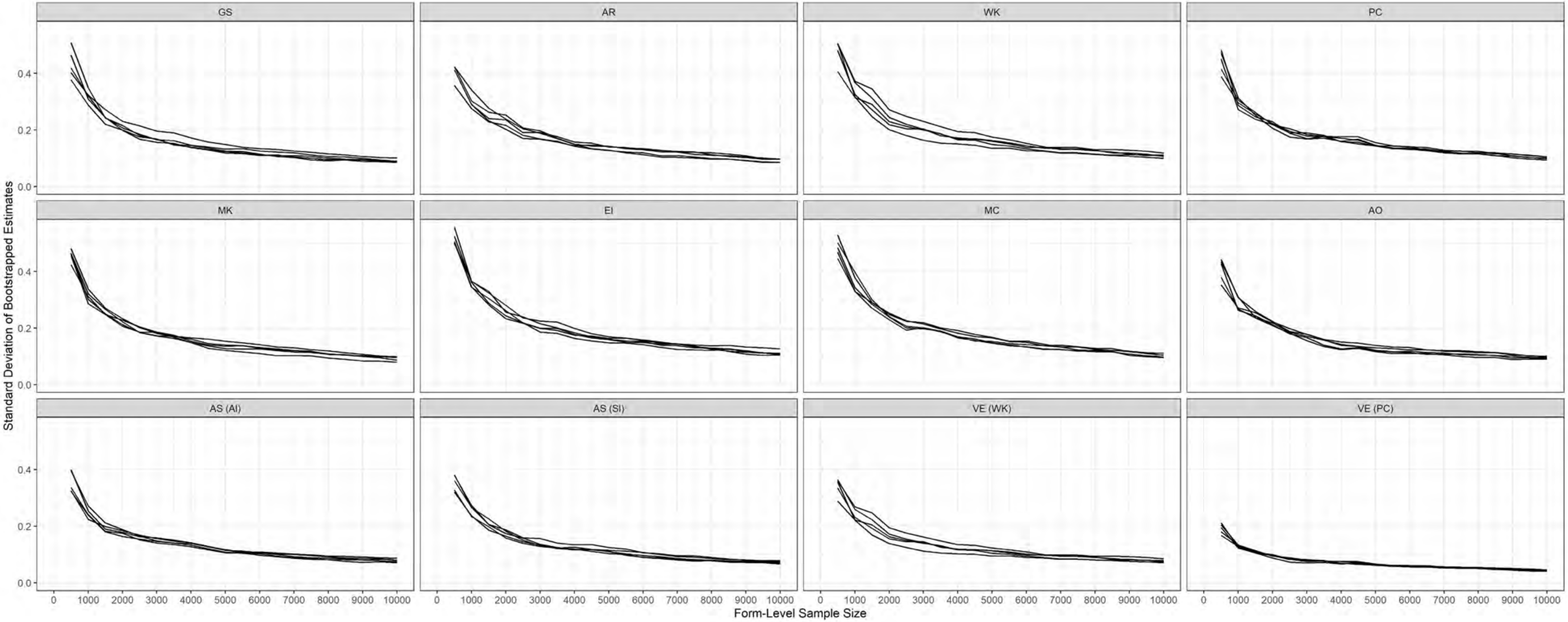
OFFICE OF PEOPLE ANALYTICS

Bootstrapped Standard Errors for Intercept Coefficients



OFFICE OF PEOPLE ANALYTICS

Bootstrapped Standard Errors for Slope Coefficients



OFFICE OF PEOPLE ANALYTICS