SLIDES ONLY

NO SCRIPT PROVIDED

OPA
OFFICE OF PEOPLE ANALYTICS

Impact of CAT-ASVAB Form Equating Procedures on Individual Scores

Jeff Dahlke
Human Resources Research Organization

Briefing presented to the DACMPT
June 12, 2024

24-P-0547

Briefing Agenda

- Background Information
  - ASVAB Scale Maintenance
  - Purpose of the Present Research
  - Overview of CAT-ASVAB Equating Procedures
- Simulation Design
- Evaluation of Simulated Scores
- Conclusions
- Questions for the DAC

Background Information

OPA
OFFICE OF PEOPLE ANALYTICS

Importance of Maintaining Consistent Score Scales

- The Services rely on composite scores to make both selection and classification decisions
  - A large volume of applicants take the CAT-ASVAB, and small differences between (unequated) forms can potentially have a large impact on the number of qualified applicants
  - Composite cut scores should produce equal qualification rates across forms
- The ASVAB score scale allows policy makers to compare current applicant aptitude with past applicants, and to set target qualifications accordingly (Segall, 2004)
  - Appropriate application of qualification cut scores requires scores to have a consistent meaning over time and across forms
  - Score scales must not be allowed to vary as a function of the ASVAB form an applicant takes

NOTE: CAT-ASVAB *forms* are what might be called *pools* in other testing programs.
The CAT-ASVAB is an adaptive test, and use of the term *form* does not imply a conventional linear fixed item set.

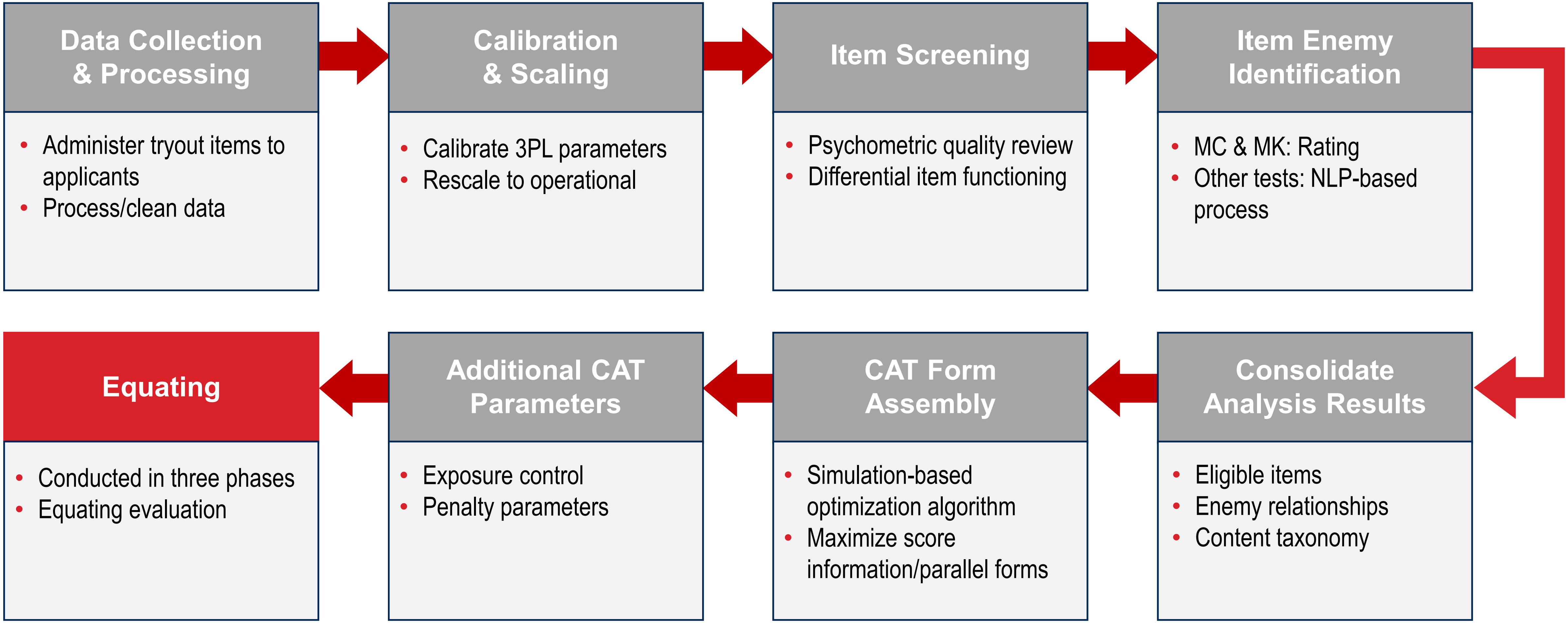
Overview of CAT-ASVAB Scale Maintenance Procedures

- 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

Purpose of the Present Research

- Following the August 2023 DACMPT meeting, the DACMPT recommended examining the potential bias that could arise at the individual level from using form-specific TCs to compute applicants' standard scores
  - The DACMPT suggested that a simulation study be designed to examine this
- HumRRO and DTAC coordinated to design a simulation that would address the DACMPT's recommendation
  - Simulation objective: Determine whether the equating process used to link examinees' scores to scaling norms could introduce bias at the individual level
  - Bias would be evident if simulated examinees with identical true latent ability levels receive systematically different scores on equated forms than on the reference form
- In this presentation, we provide an overview of the simulation's design, results, and implications

Process Overview for CAT-ASVAB Form Development



Data Collection & Processing

- Administer tryout items to applicants
- Process/clean data

Calibration & Scaling

- Calibrate 3PL parameters
- Rescale to operational

Item Screening

- Psychometric quality review
- Differential item functioning

Item Enemy Identification

- MC & MK: Rating
- Other tests: NLP-based process

Consolidate Analysis Results

- Eligible items
- Enemy relationships
- Content taxonomy

CAT Form Assembly

- Simulation-based optimization algorithm
- Maximize score information/parallel forms

Additional CAT Parameters

- Exposure control
- Penalty parameters

Equating

- Conducted in three phases
- Equating evaluation

OPA
OFFICE OF PEOPLE ANALYTICS

Equating Objective: Standard Score Equating

- **Equipercentile Objective:** ASVAB forms were originally equated to a reference form using equipercentile methods to produce ***equivalent composite distributions*** across alternate forms
  - When ASVAB transitioned to IRT scoring, new CAT-ASVAB forms continued to be equated to a reference form using a linear method that matches the mean and standard deviation of standard scores to a reference form
- IRT invariance assumptions take us most of the way toward ensuring that score distributions are scaled the same way across forms
  - The current equating approach relies more heavily on the invariance property of IRT than did the equipercentile approach, and aims to create equal distributions of scores across alternate forms
  - Equating serves as an “insurance policy” on top of IRT invariance assumptions
    - Guarantees that standard scores and the composite scores used in decision-making have the same means and standard deviations across forms
  - IRT invariance assumptions and the equating process work in tandem to ensure examinees’ standard scores are comparable across all ASVAB forms

CAT-ASVAB Equating: The Reference Form

- Changes to the ASVAB, like introducing new CAT forms, must be introduced in a deliberate, carefully planned manner to ensure the continuity of the interpretation of ASVAB scores
- Any given composite cut score should have the same meaning . . .
  - irrespective of which form is administered
  - as it did when standards were originally set
- The current ASVAB score scale was developed from a nationally representative sample collected during the PAY97 norming study (Moore et al., 2000)
  - Standard scores were normed to have population distribution with μ=50 and σ=10
- A reference form was included in the PAY97 study to initialize the scaling of scores
- Reference forms have subsequently been administered for special purposes only and serve to define the reference scale in equating studies

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
  - Conducted for each subtest (and for Auto and Shop [AS] and Verbal Expression [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

OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB Equating: Mechanics of the Process

- A set of pre-established reference form TCs exist 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

Research Questions for the Simulation

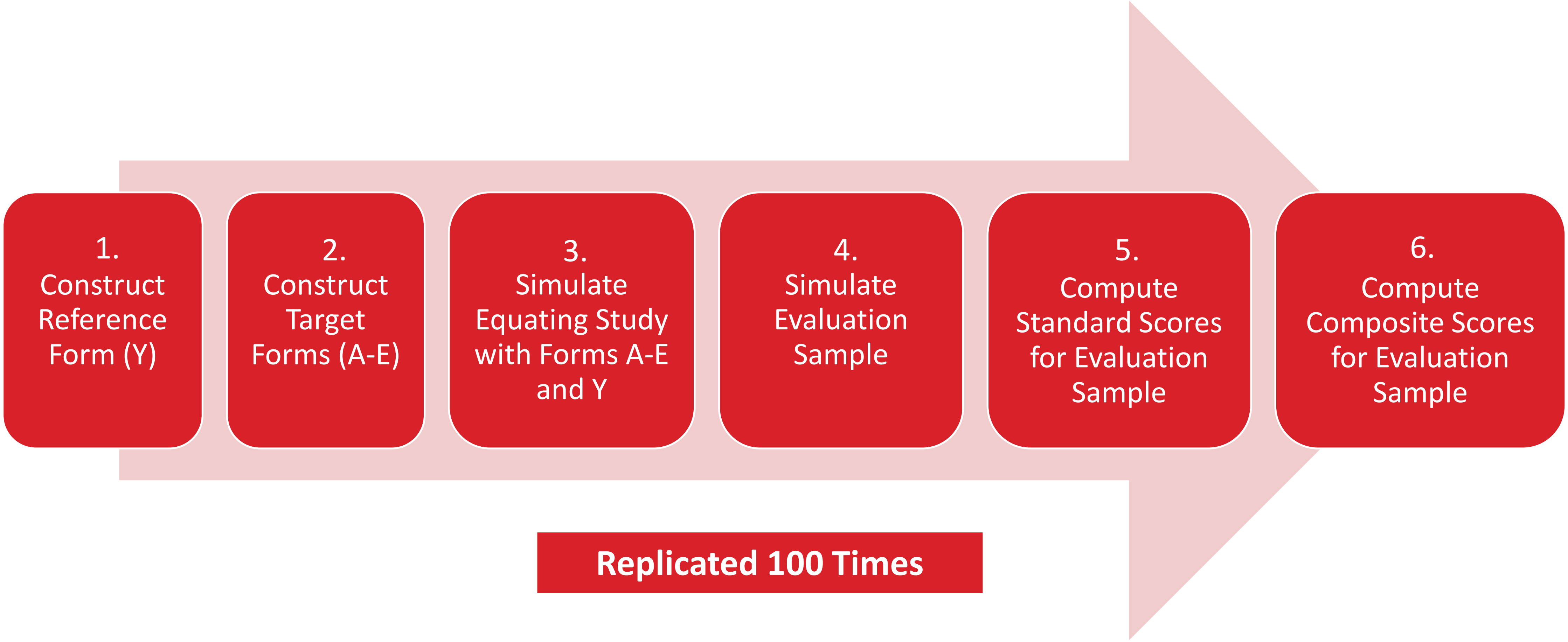
1. At the level of individual composites, what percentage of variance is attributable to imperfect estimation of the final TCs for new forms?
2. Given that the final TCs are computed separately per form, does the form that an examinee is randomly assigned produce person-level score bias (evaluated at the composite level), as a function of the form-specific TCs?
3. Given that the equating procedure is designed to maintain scales at the population level, does equating have differential impacts on individuals at different ability levels?

Simulation Design

Simulation Infrastructure and Scope

- Used the same simulation pipeline infrastructure as was described in the preceding presentation, “An Evaluation of Calibration Method and Sample Size on the Reliability of New CAT-ASVAB Forms” (Heinrich-Wallace, 2024)
- Simulated all CAT-ASVAB subtests except for Assembling Objects (AO) due to ongoing research evaluating the dimensionality of AO:
  - General Science (GS)
  - Arithmetic Reasoning (AR)
  - Word Knowledge (WK)
  - Paragraph Comprehension (PC)
  - Math Knowledge (MK)
  - Electronics Information (EI)
  - Auto Information (AI)
  - Shop Information (SI)
  - Mechanical Comprehension (MC)

Schematic Outline of Simulation Process



1. Construct Reference Form (Y)

2. Construct Target Forms (A-E)

3. Simulate Equating Study with Forms A-E and Y

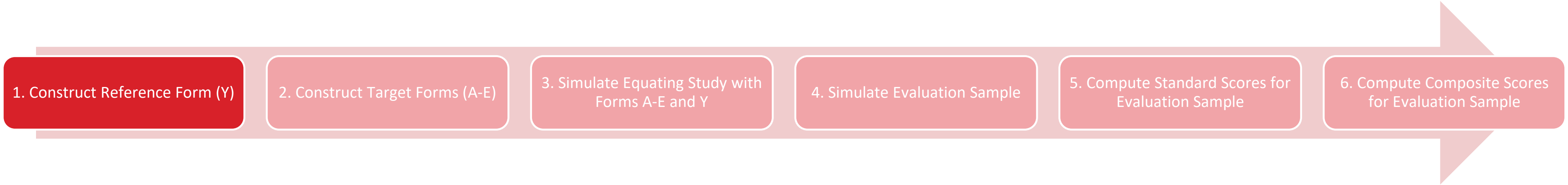
4. Simulate Evaluation Sample

5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

Replicated 100 Times

Step 1: Construct Reference Form (Y)



1. Construct Reference Form (Y)

2. Construct Target Forms (A-E)

3. Simulate Equating Study with Forms A-E and Y

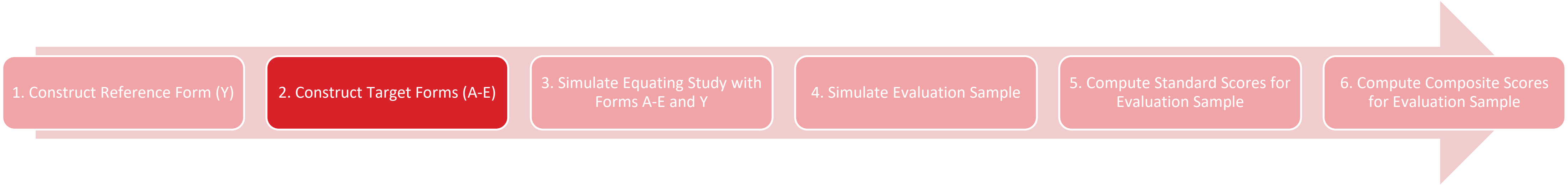
4. Simulate Evaluation Sample

5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

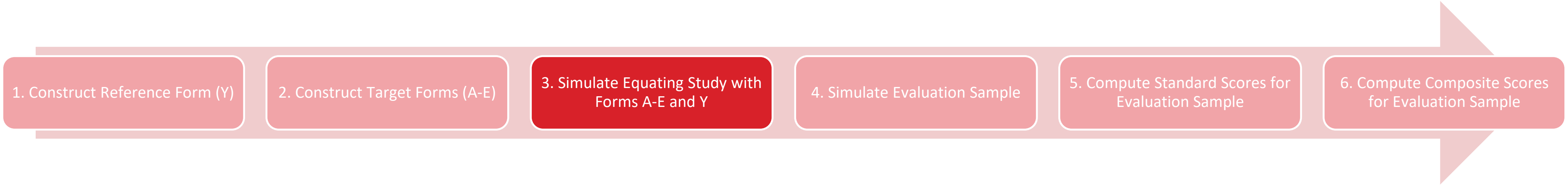
- Simulate two seed versions, each with 400 seed items (i.e., newly developed tryout items) per subtest
  - True item parameters simulated via the copula method
- Simulate responses to the seed items in each seed version
  - Simulated examinees (simulees) were drawn from a multivariate normal theta distribution
  - 32k simulees generated to reach sample-size objective
    - 1,200 simulees per item, with 15 seed items administered to each simulee for each subtest
- Calibrate the seed items and rescale their IRT item parameters
  - Target scale = pre-specified moments for latent distribution
- Construct 5 new forms from the calibrated seed items
- Select 1 of those forms as the reference form (Form Y)
- Simulate 10k additional examinees for Form Y
  - Use estimated thetas to initialize TCs, such that standard score distributions have means of 50 and SDs of 10

Step 2: Construct Target Forms (A-E)


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
6. Compute Composite Scores for Evaluation Sample

- Simulate two seed versions, each with 400 seed items per subtest
  - True item parameters simulated via the copula method
- Simulate responses to the seed items in each seed version
  - Seeded within CAT-ASVAB administration featuring the Step 1 forms
  - Simulees were drawn from a multivariate normal theta distribution
  - 32k simulees generated to reach sample-size objective
    - 1,200 simulees per item, with 15 seed items administered to each simulee for each subtest
- Calibrate the seed items and rescale their IRT item parameters
  - Target scale = moments estimated using simulated records from Step 1 forms
- Construct 5 new forms from the calibrated seed items (Forms A-E)

Step 3: Simulate Equating Study with Forms A-E and Y


1. Construct Reference Form (Y)

2. Construct Target Forms (A-E)

3. Simulate Equating Study with Forms A-E and Y

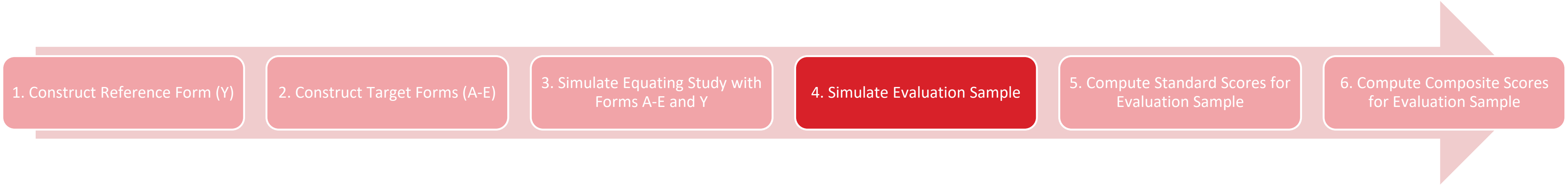
4. Simulate Evaluation Sample

5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

- Draw 60k simulees from a multivariate normal theta distribution
  - 10k for each form, matching operational sample size for final equating
- Assign 10k simulees to each form
- Administer each subtest to each simulee
- Compute the equated TCs for each standard score distribution on each form

Step 4: Simulate Evaluation Sample



1. Construct Reference Form (Y)

2. Construct Target Forms (A-E)

3. Simulate Equating Study with Forms A-E and Y

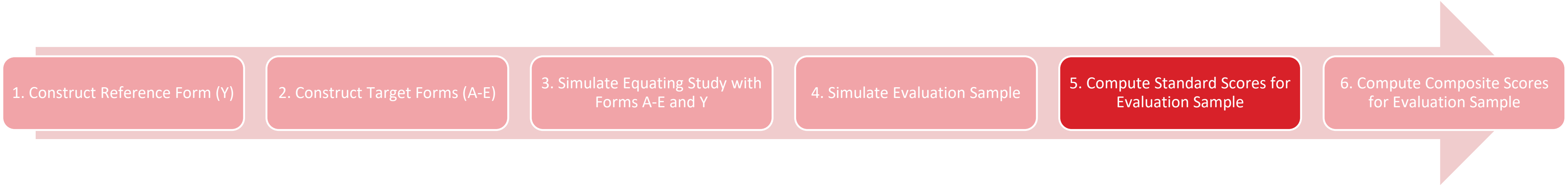
4. Simulate Evaluation Sample

5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

- Draw 10k simulees from a multivariate normal theta distribution
- Simulate records for each simulee on each form (Form Y and Forms A-E)
  - This has the effect of holding the true-score ability distribution constant across all 6 forms
  - The results for a given simulee across forms are matched on ability but are not dependent in any other way
- Compute theta estimates for each simulee on each form

Step 5: Compute Standard Scores for Evaluation Sample


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

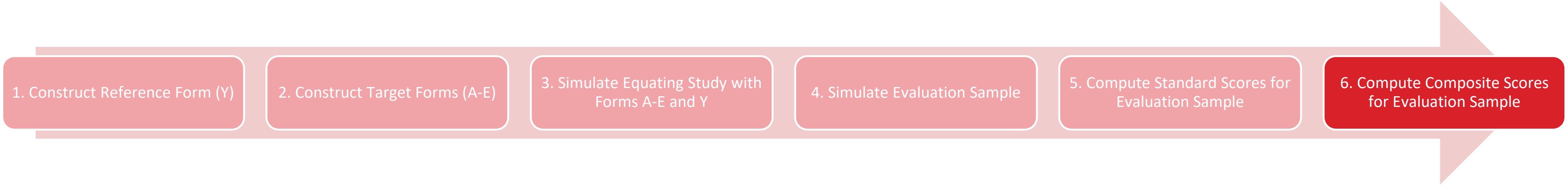
5. Compute Standard Scores for Evaluation Sample

6. Compute Composite Scores for Evaluation Sample

- Compute standard scores for simulees on all forms using three formulations:
  - $\boldsymbol{True}_{\boldsymbol{GeneratingTCs}}$: Generating theta scores transformed to standard scores using TCs derived from generating ability distributions
    - Represent true scores that are not impacted by measurement error or equating
  - $\boldsymbol{Obs}_{\boldsymbol{ProvisionalTCs}}$: Estimated theta scores transformed to standard scores using provisional TCs inherited from the reference form prior to equating new forms
    - Represent observed scores that are impacted by measurement error, but not equating
  - $\boldsymbol{Obs}_{\boldsymbol{FinalTCs}}$: Estimated theta scores transformed to standard scores using final form-specific TCs estimated via equating
    - Represent observed scores that are impacted by measurement error and equating

NOTE: For new forms, $Obs_{ProvisionalTCs}$ scores rely on IRT invariance assumptions. For reference forms, $Obs_{FinalTCs}$ scores are the same as the $Obs_{ProvisionalTCs}$ scores.

Step 6: Compute Composite Scores for Evaluation Sample



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
6. Compute Composite Scores for Evaluation Sample

- Use the standard scores from Step 5 to compute AFQT and Service composite scores
  - These are the primary focus of our evaluations, as they are the scores used to make decisions
  - The Service composites were limited to the composites that are evaluated during equating studies
    - Given that AO was not included in the simulation, the 2 composites typically evaluated during equating studies that include AO (Navy MEC2 and Navy OPS) were omitted from our design

Composites Included in the Simulation

Service	Composite	Computational Formula
All	AFQT	2(VE) + AR + MK
Air Force	Mechanical (M)	AR + 2(VE) + MC + AS
	Administrative (A)	VE + MK
	General (G)	VE + AR
	Electronic (E)	AR + MK + EI + GS
Army	Clerical (CL)	*
	Combat (CO)	*
	Electronics Repair (EL)	*
	Field Artillery (FA)	*
	General Maintenance (GM)	*
	General Technical (GT)	AR + VE
	Mechanical Maintenance (MM)	*
	Operators/Food (OF)	*
	Surveillance/Communication (SC)	*
	Skilled Technician (ST)	*
Marine Corps	Clerical (CL)	VE + MK
	Electrical (EL)	AR + MK + EI + GS
	General Technician (GT)	VE + AR + MC
	Mechanical (MM)	AR + MC + AS + EI
Navy	Administrative (ADM)	VE + MK
	Basic Electricity and Electronics (BEE)	GS + AR + 2(MK)
	Electronics (EL)	GS + AR + MK + EI
	Engineering (ENG)	AS + MK
	General Technician (GT)	VE + AR
	Hospitalman (HM)	VE + GS + MK
	Mechanical1 (MEC)	AR + AS + MC
	Nuclear (NUC)	VE + AR + MK + MC

Note. AFQT scores are subsequently transformed into a percentile metric.

* Computed as a non-integer weighted linear combination of standard scores for GS, AR, MK, EI, MC, AS, and VE.

OPA
OFFICE OF PEOPLE ANALYTICS

Evaluation of Simulated Scores

OFFICE OF PEOPLE ANALYTICS

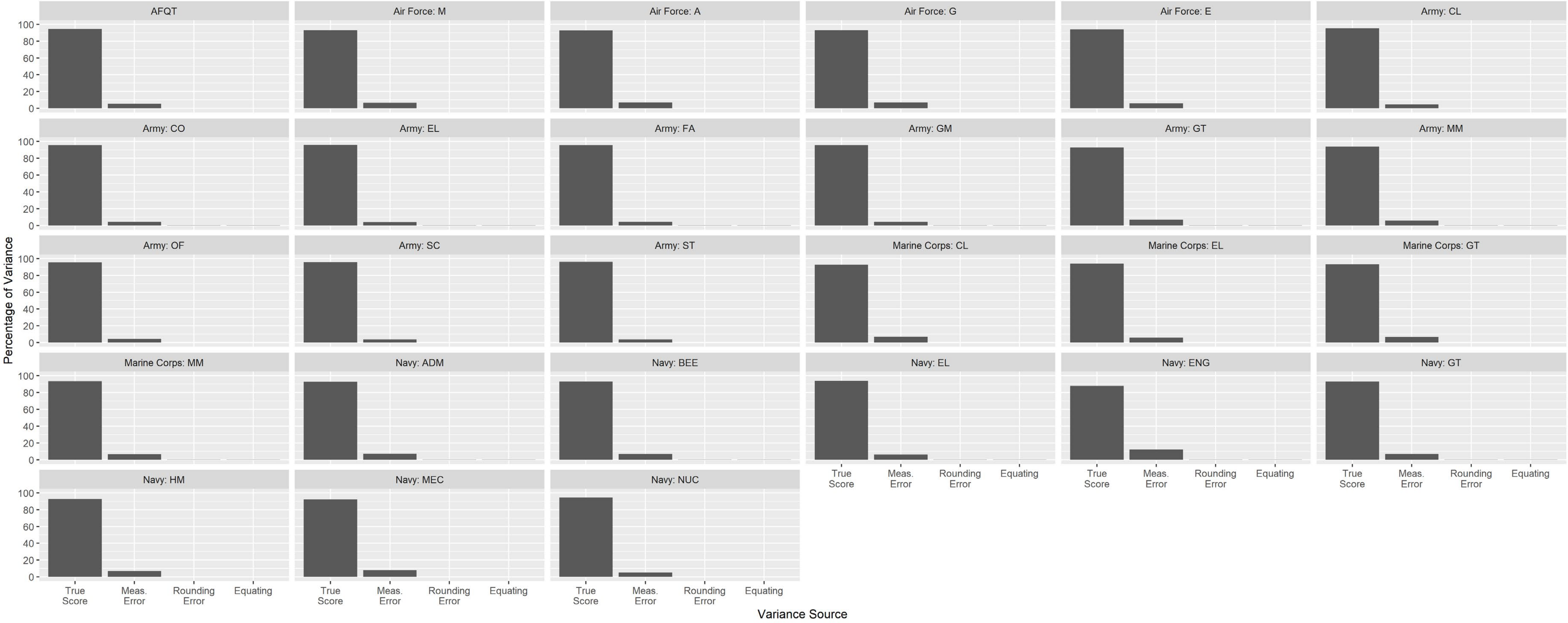
Research Questions for the Simulation

1. At the level of individual composites, what percentage of variance is attributable to imperfect estimation of the final TCs for new forms?
2. Given that the final TCs are computed separately per form, does the form that an examinee is randomly assigned produce person-level score bias (evaluated at the composite level), as a function of the form-specific TCs?
3. Given that the equating procedure is designed to maintain scales at the population level, does equating have differential impacts on individuals at different ability levels?

Evaluation of Equating Error Variance (RQ1)

- Decomposed observed variance into variance from four sources:
  - $\sigma^2_{Observed} = \sigma^2_{Reliable} + \sigma^2_{MeasurementError} + \sigma^2_{RoundingError} + \sigma^2_{EquatingError}$
    - ***Note*:** Rounding error represents irreducible error due to the rounding of standard scores to integers.
- Partitioned variance by estimating 3 regression models:
  - **Model 1**: Quantify amount of variance attributable to rounding
    - $\mathrm{round}(Obs_{FinalTCs_i}) = b_0 + b_1(\mathrm{round}(Obs_{FinalTCs_i}) - Obs_{FinalTCs_i}) + e_i$
  - **Model 2**: Quantify amount of variance attributable to rounding and true scores
    - $\mathrm{round}(Obs_{FinalTCs_i}) = b_0 + b_1(\mathrm{round}(Obs_{FinalTCs_i}) - Obs_{FinalTCs_i}) + b_2(True_{GeneratingTCs_i}) + e_i$
  - **Model 3**: Quantify amount of variance attributable to rounding, and true scores, and measurement error
    - $\mathrm{round}(Obs_{FinalTCs_i}) = b_0 + b_1(\mathrm{round}(Obs_{FinalTCs_i}) - Obs_{FinalTCs_i}) + b_2(True_{GeneratingTCs_i}) + b_3\left(\mathrm{round}(Obs_{ProvisionalTCs_i})\right) + e_i$
- Estimated variance allocations using R^2 values estimated for scores from the 5 target forms in each replication:
  - $Reliable\ (True\ Score)\ \% = (R^2_{Model2} - R^2_{Model1}) \times 100$
  - $Rounding\ Error\ \% = R^2_{Model1} \times 100$
  - $Measurement\ Error\ \% = (R^2_{Model3} - R^2_{Model2}) \times 100$
  - $Equating\ Error\ \% = (1 - R^2_{Model3}) \times 100$

Evaluation of Equating Error Variance (RQ1)



Evaluation of Equating Error Variance (RQ1)

- Equating error represents a very small share of observed variance
  - Less than 0.1% of variance for 24/27 composites
  - Contribution to scores is similar in magnitude to rounding error

Service	Composite	Mean (SD)			
		Reliable %	Meas. Error %	Rounding Error %	Equating %
All	AFQT	94.67 (0.17)	5.22 (0.17)	0.04 (0.02)	0.07 (0.01)
Air Force	Mechanical (M)	93.18 (0.21)	6.72 (0.21)	0.03 (0.01)	0.07 (0.01)
	Administrative (A)	92.83 (0.27)	7.01 (0.27)	0.05 (0.02)	0.11 (0.02)
	General (G)	92.97 (0.26)	6.91 (0.26)	0.03 (0.01)	0.08 (0.01)
	Electronic (E)	93.84 (0.28)	6.06 (0.27)	0.03 (0.01)	0.07 (0.01)
Army	Clerical (CL)	95.24 (0.14)	4.62 (0.14)	0.07 (0.03)	0.07 (0.00)
	Combat (CO)	95.64 (0.15)	4.23 (0.15)	0.06 (0.02)	0.07 (0.01)
	Electronics Repair (EL)	96.02 (0.13)	3.85 (0.13)	0.06 (0.03)	0.06 (0.00)
	Field Artillery (FA)	95.64 (0.15)	4.23 (0.14)	0.06 (0.02)	0.06 (0.00)
	General Maintenance (GM)	95.60 (0.17)	4.27 (0.17)	0.06 (0.02)	0.07 (0.01)
	General Technical (GT)	92.89 (0.22)	6.76 (0.22)	0.27 (0.06)	0.07 (0.01)
	Mechanical Maintenance (MM)	93.97 (0.35)	5.89 (0.35)	0.06 (0.02)	0.07 (0.01)
	Operators/Food (OF)	95.58 (0.16)	4.30 (0.16)	0.06 (0.02)	0.06 (0.01)
	Surveillance/Communication (SC)	95.92 (0.12)	3.95 (0.12)	0.07 (0.03)	0.07 (0.00)
	Skilled Technician (ST)	96.06 (0.11)	3.81 (0.11)	0.06 (0.02)	0.06 (0.00)
Marine Corps	Clerical (CL)	92.89 (0.29)	7.00 (0.29)	0.00 (0.01)	0.10 (0.01)
	Electrical (EL)	93.85 (0.25)	6.04 (0.25)	0.03 (0.02)	0.08 (0.01)
	General Technician (GT)	93.37 (0.24)	6.50 (0.24)	0.07 (0.03)	0.07 (0.01)
	Mechanical (MM)	93.40 (0.36)	6.46 (0.36)	0.07 (0.03)	0.07 (0.01)
Navy	Administrative (ADM)	92.82 (0.30)	7.01 (0.29)	0.09 (0.04)	0.08 (0.01)
	Basic Electricity and Electronics (BEE)	93.16 (0.34)	6.72 (0.34)	0.06 (0.04)	0.06 (0.01)
	Electronics (EL)	93.84 (0.26)	6.06 (0.25)	0.05 (0.03)	0.05 (0.01)
	Engineering (ENG)	87.78 (0.53)	12.01 (0.54)	0.09 (0.04)	0.11 (0.03)
	General Technician (GT)	93.09 (0.22)	6.79 (0.22)	0.06 (0.03)	0.06 (0.01)
	Hospitalman (HM)	92.86 (0.35)	7.01 (0.35)	0.06 (0.03)	0.06 (0.01)
	Mechanical1 (MEC)	92.10 (0.33)	7.80 (0.34)	0.04 (0.02)	0.05 (0.01)
	Nuclear (NUC)	94.71 (0.18)	5.21 (0.18)	0.04 (0.02)	0.04 (0.00)

Evaluation of Overall Score Bias (RQ2)

- Does using new forms' final TCs to compute standard scores produce systematically different score estimates compared to the scores examinees would achieve if they were assigned to the reference form?
  - In each replication of the simulation design, the same distribution of generating thetas was used to simulate records for each form (Y and A-E)
  - Allows for evaluation of scores between forms for simulees of identical ability
- Evaluated bias for each combination of composite × form × replication
  - $Bias = \sum_{i=1}^{N} \frac{\left(x_{NewForm_i} - x_{ReferenceForm_i}\right)}{N}$
    - 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)
- For each evaluation contrast for each composite, we summarized distributions of estimates for 5 new forms across 100 replications (500 estimates per composite)

Evaluation of Overall Score Bias (RQ2)

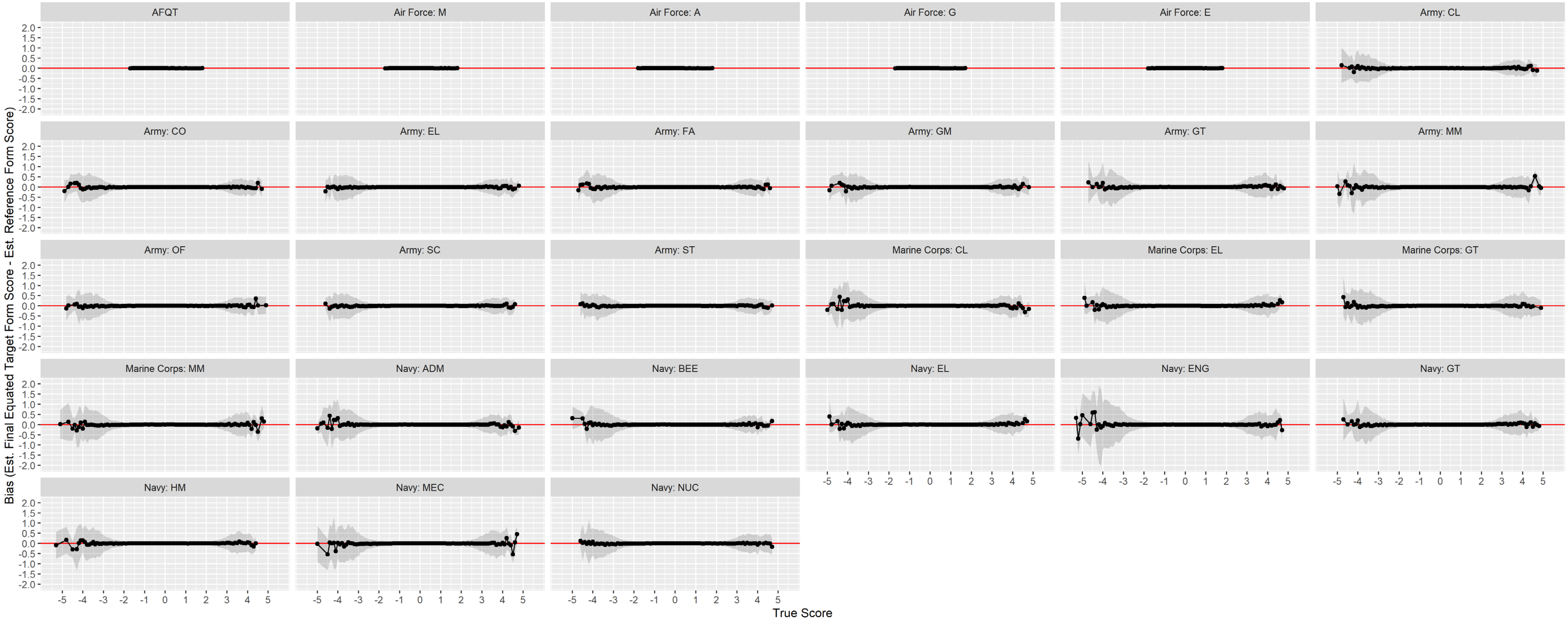
- Across all composites, the average bias was <0.005 in absolute value
  - Point estimates were all 0.00
  - SDs of estimates were all 0.00 – 0.01
- No evidence of systematic bias as a function of simulees being assigned to an equated new form vs. the reference form

Service	Composite	Mean (SD)
All	AFQT	0.00 (0.01)
Air Force	Mechanical (M)	0.00 (0.01)
	Administrative (A)	0.00 (0.01)
	General (G)	0.00 (0.01)
	Electronic (E)	0.00 (0.01)
Army	Clerical (CL)	0.00 (0.01)
	Combat (CO)	0.00 (0.01)
	Electronics Repair (EL)	0.00 (0.01)
	Field Artillery (FA)	0.00 (0.01)
	General Maintenance (GM)	0.00 (0.01)
	General Technical (GT)	0.00 (0.01)
	Mechanical Maintenance (MM)	0.00 (0.01)
	Operators/Food (OF)	0.00 (0.01)
	Surveillance/Communication (SC)	0.00 (0.00)
	Skilled Technician (ST)	0.00 (0.00)
Marine Corps	Clerical (CL)	0.00 (0.01)
	Electrical (EL)	0.00 (0.01)
	General Technician (GT)	0.00 (0.01)
	Mechanical (MM)	0.00 (0.01)
Navy	Administrative (ADM)	0.00 (0.01)
	Basic Electricity and Electronics (BEE)	0.00 (0.01)
	Electronics (EL)	0.00 (0.01)
	Engineering (ENG)	0.00 (0.01)
	General Technician (GT)	0.00 (0.01)
	Hospitalman (HM)	0.00 (0.01)
	Mechanical1 (MEC)	0.00 (0.01)
	Nuclear (NUC)	0.00 (0.01)

Evaluation of Conditional Score Bias (RQ3)

- We replicated our bias analyses for segments of the true-score ability continuum to evaluate conditional effects
- 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

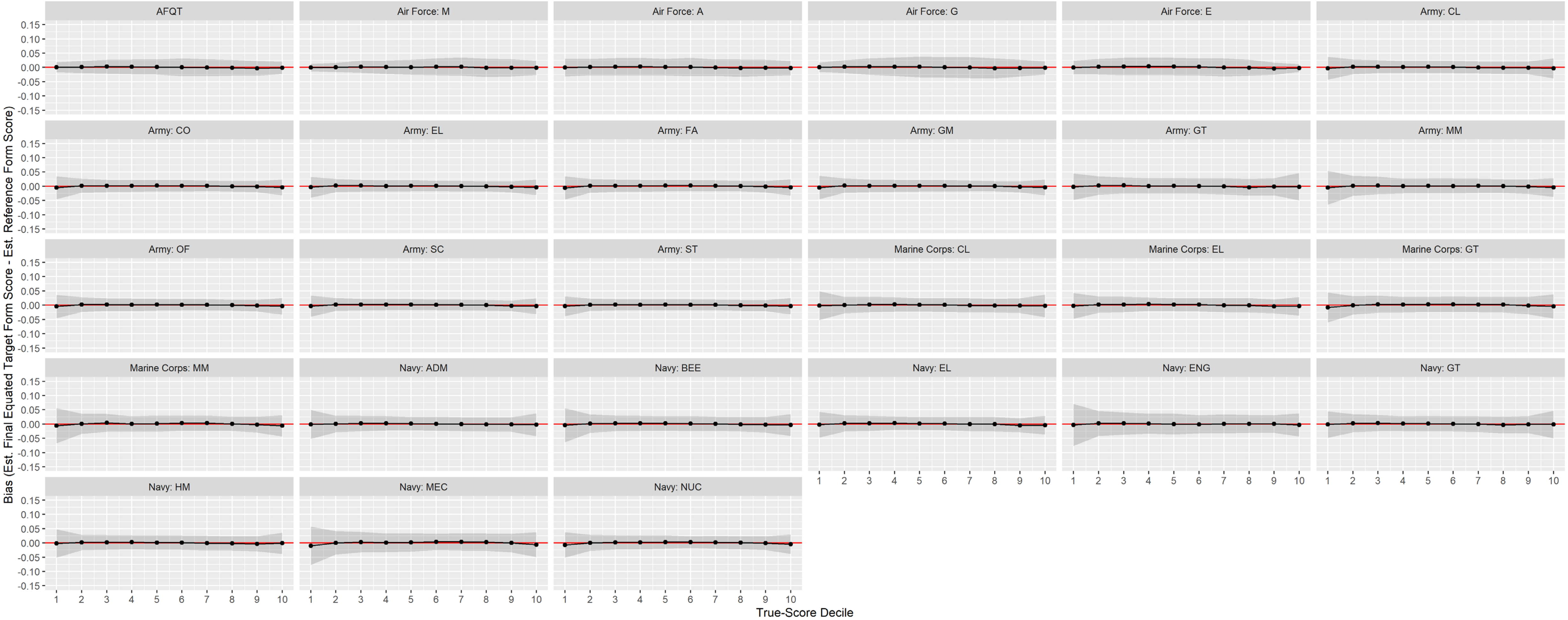
Evaluation of Conditional Score Bias (RQ3): By True-Score z Score



- Error ribbons represent 95% confidence intervals.
- All conditional mean bias estimates are < .7 in absolute magnitude.
- Non-zero mean bias estimates are primarily a function of sampling error for uncommonly attained scores at the tails of distributions.

OFFICE OF PEOPLE ANALYTICS

Evaluation of Conditional Score Bias (RQ3): By True-Score Decile



- Error ribbons represent 95% confidence intervals.
- All conditional mean bias estimates are empirically indistinguishable from zero.

Conclusions

OFFICE OF PEOPLE ANALYTICS

Conclusions

- Equating is responsible for a very small proportion of observed-score variance and does not systematically bias estimated scores
    - Bias is not evident at the level of complete score distributions nor at the level of specific scores or deciles
- The results of our simulation indicate that the equating process serves its intended function without detrimental impacts on examinees’ scores
    - Equating provides an added layer of scale-continuity assurance on top of IRT invariance assumptions
    - Equating ensures that standard scores have the same mean and SD across forms
        - Allows CAT-ASVAB forms to be used interchangeably
        - Supports the comparability of composite scores and the consistency of qualification rates across forms
    - The use of equating to maintain the means and SDs of standard score distributions does not introduce bias to individual-level scores

Questions for the DAC

Questions for the DAC

- Do the results of this simulation sufficiently address the DAC's recommendation (from August 2023) to evaluate whether equating introduces bias into individual scores?
- Does the DAC have feedback on potentially reducing the sample size for equating or eliminating a phase from the three-phase design?
  - Current (cumulative) sample size goals:
    - Phase 1: 500 per form
    - Phase 2: 1,500 per form
    - Phase 3: 10,000 per form

Thank You!

For more information please contact:

Jeff Dahlke
jdahlke@humrro.org

References

- 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.
- Moore, W., Pedlow, S., Krishnamurty, P., & Wolter, K. (2000). *National Longitudinal Survey of Youth 1997 (NLSY97)*. National Opinion Research Center.
- Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale*. Defense Manpower Data Center.

OPA

OFFICE OF PEOPLE ANALYTICS

Supplemental Slides

Equating Study Design

OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB Equating: Refinement 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
  - Phase sample sizes are cumulative such that they include all individuals from the previous phase
  - Intent of phased design is 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)
    - The first two phases of TC estimation pool data across new forms to estimate one set of TCs that is shared across 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

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

Estimation of Standard Scores for Generating Ability Distributions

OFFICE OF PEOPLE ANALYTICS

Procedure for Computing True Subtest Standard Scores

- Convert generating thetas to true standard scores for each subtest:
  - $Standard\ Score_{Subtest} = \frac{(\theta - \mu_\theta)}{\sigma_\theta} \times 10 + 50$
    - Here, μ_θ and σ_θ represent the mean and SD parameters associated with a given subtest’s generating ability distribution.
  - These values are fixed; they do not vary across forms or simulation replications, and each simulee has one set of true standard scores across forms.
  - These standard score values represent scores from a population with μ=50 and σ=10.

Procedure for Computing True AS Standard Scores

- Convert thetas for AI and SI to standard scores for the AS composite, where AI and SI are equally weighted:
  - $Standard\ Score_{AS} = \frac{\left(\frac{\theta_{AI} - \mu_{\theta_{AI}}}{\sigma_{\theta_{AI}}}\right) + \left(\frac{\theta_{SI} - \mu_{\theta_{SI}}}{\sigma_{\theta_{SI}}}\right)}{\sqrt{2 + 2 \times \frac{\sigma_{\theta_{AI}, \theta_{SI}}}{\sigma_{\theta_{AI}} \times \sigma_{\theta_{SI}}}}} \times 10 + 50$
- Where:
  - $\mu_{\theta_{AI}}$ and $\sigma_{\theta_{AI}}$ represent the mean and SD parameters associated with the generating ability distribution for AI.
  - $\mu_{\theta_{SI}}$ and $\sigma_{\theta_{SI}}$ represent the mean and SD parameters associated with the generating ability distribution for SI.
  - $\sigma_{\theta_{AI}, \theta_{SI}}$ represents the covariance parameter between AI and SI.

OPA
OFFICE OF PEOPLE ANALYTICS

Procedure for Computing True VE Standard Scores

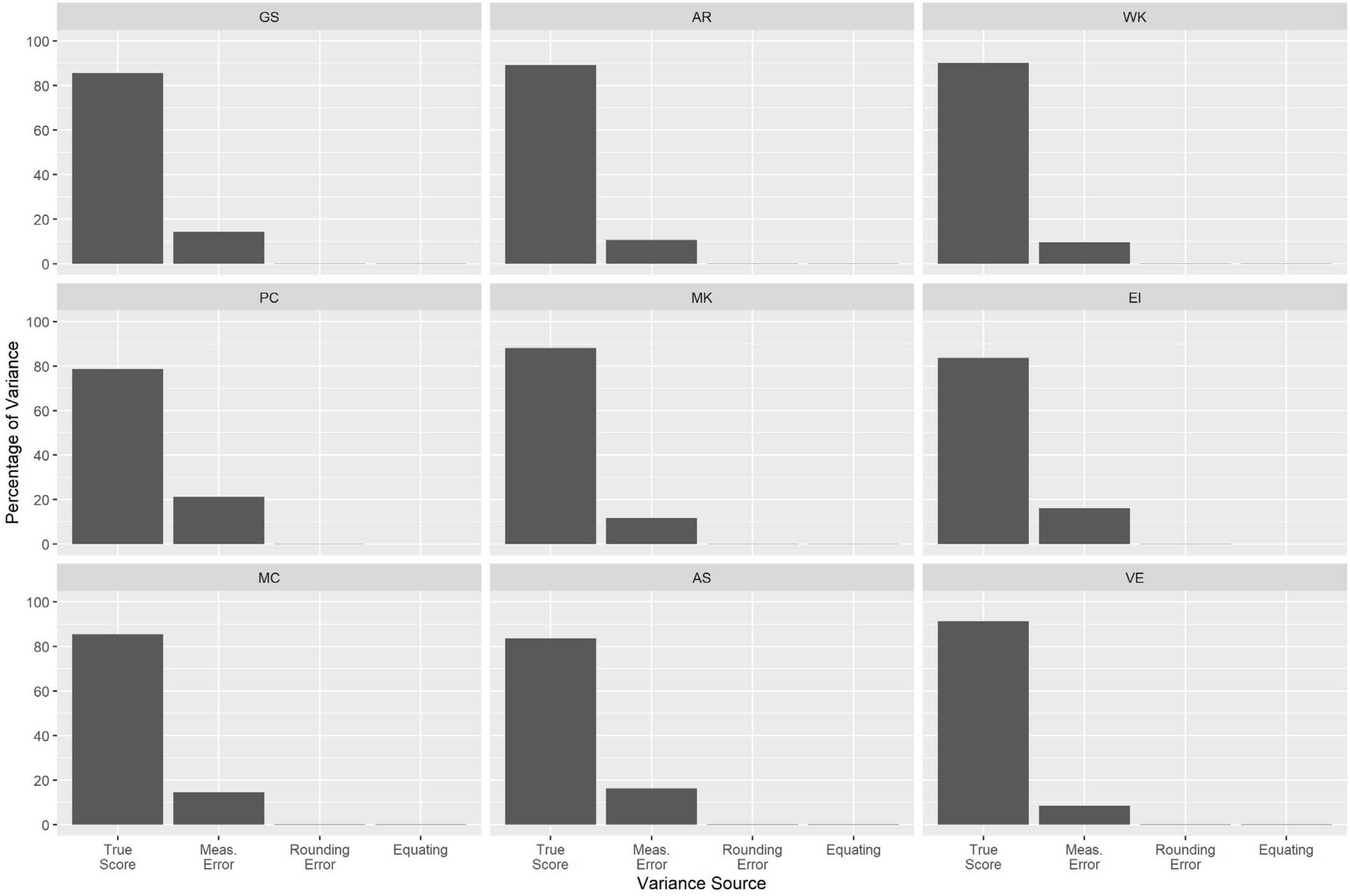
- Convert generating thetas for WK and PC to true standard scores for the VE composite, such that WK receives twice the weight of PC:

  - $$Standard\ Score_{VE} = \frac{\frac{2}{3}\times\left(\frac{\theta_{WK}-\mu_{\theta_{WK}}}{\sigma_{\theta_{WK}}}\right)+\frac{1}{3}\times\left(\frac{\theta_{PC}-\mu_{\theta_{PC}}}{\sigma_{\theta_{PC}}}\right)}{\sqrt{\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2+2\times\frac{2}{3}\times\frac{1}{3}\times\frac{\sigma_{\theta_{WK},\theta_{PC}}}{\sigma_{\theta_{WK}}\times\sigma_{\theta_{PC}}}}} \times 10 + 50$$

  - Where:
    - $\mu_{\theta_{WK}}$ and $\sigma_{\theta_{WK}}$ represent the mean and SD parameters associated with the generating ability distribution for WK.
    - $\mu_{\theta_{PC}}$ and $\sigma_{\theta_{PC}}$ represent the mean and SD parameters associated with the generating ability distribution for PC.
    - $\sigma_{\theta_{WK},\theta_{PC}}$ represents the covariance parameter between WK and PC.

Simulation Evaluation for Standard Scores

Evaluation of Equating Error Variance (RQ1)

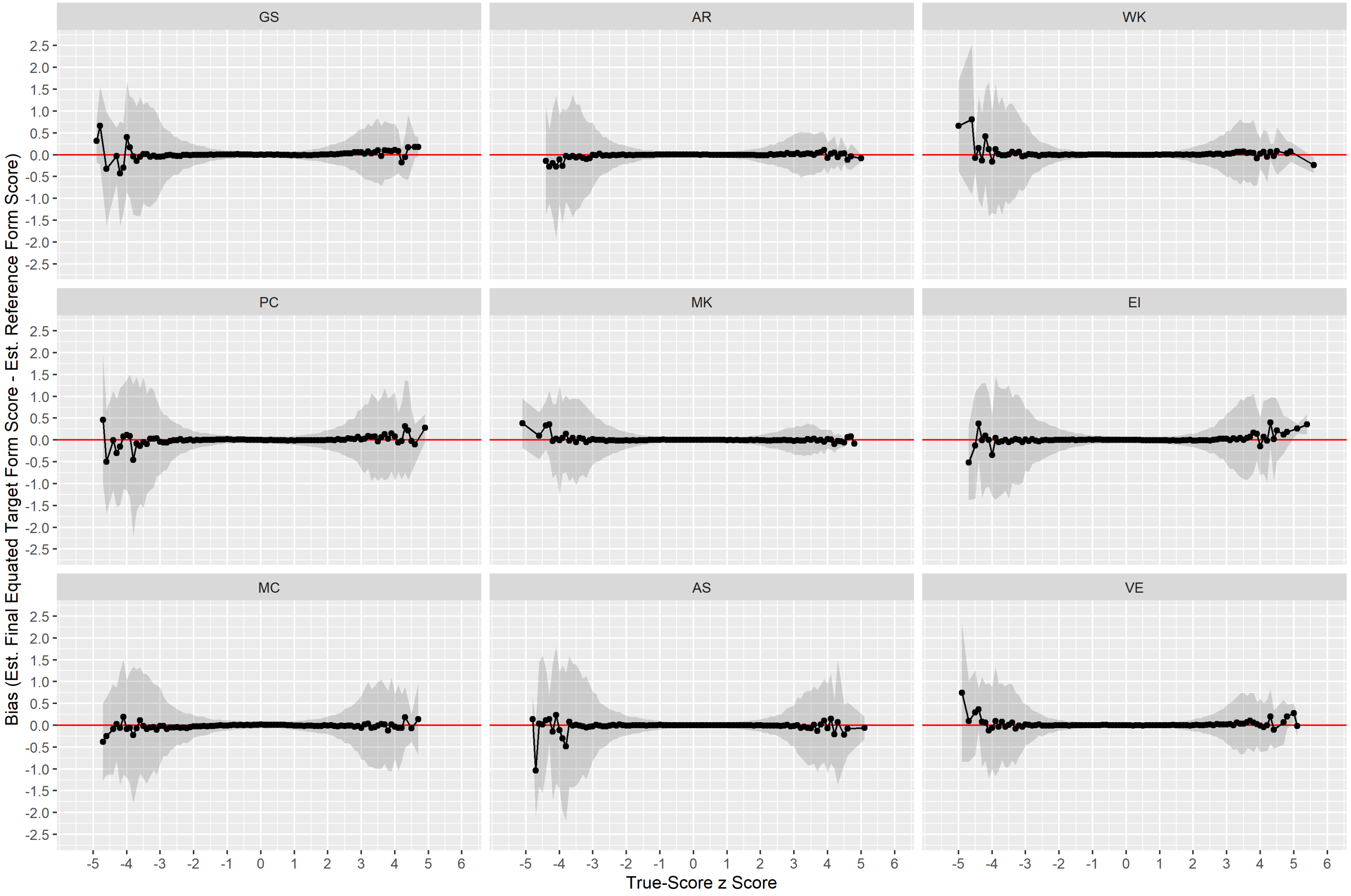


Standard Score	Mean (SD)			
	Reliable %	Meas. Error %	Rounding Error %	Equating %
GS	85.55 (1.13)	14.29 (1.12)	0.08 (0.04)	0.08 (0.01)
AR	89.24 (0.41)	10.61 (0.41)	0.08 (0.04)	0.07 (0.01)
WK	90.13 (0.49)	9.68 (0.48)	0.11 (0.05)	0.08 (0.02)
PC	78.58 (0.75)	21.18 (0.71)	0.14 (0.16)	0.10 (0.01)
MK	88.03 (0.85)	11.70 (0.85)	0.14 (0.07)	0.12 (0.02)
EI	83.69 (1.82)	16.10 (1.81)	0.11 (0.04)	0.10 (0.02)
MC	85.39 (0.92)	14.48 (0.92)	0.07 (0.03)	0.06 (0.01)
AS	83.52 (1.13)	16.26 (1.14)	0.10 (0.03)	0.12 (0.05)
VE	91.34 (0.31)	8.44 (0.30)	0.11 (0.03)	0.11 (0.01)

Evaluation of Overall Score Bias (RQ2)

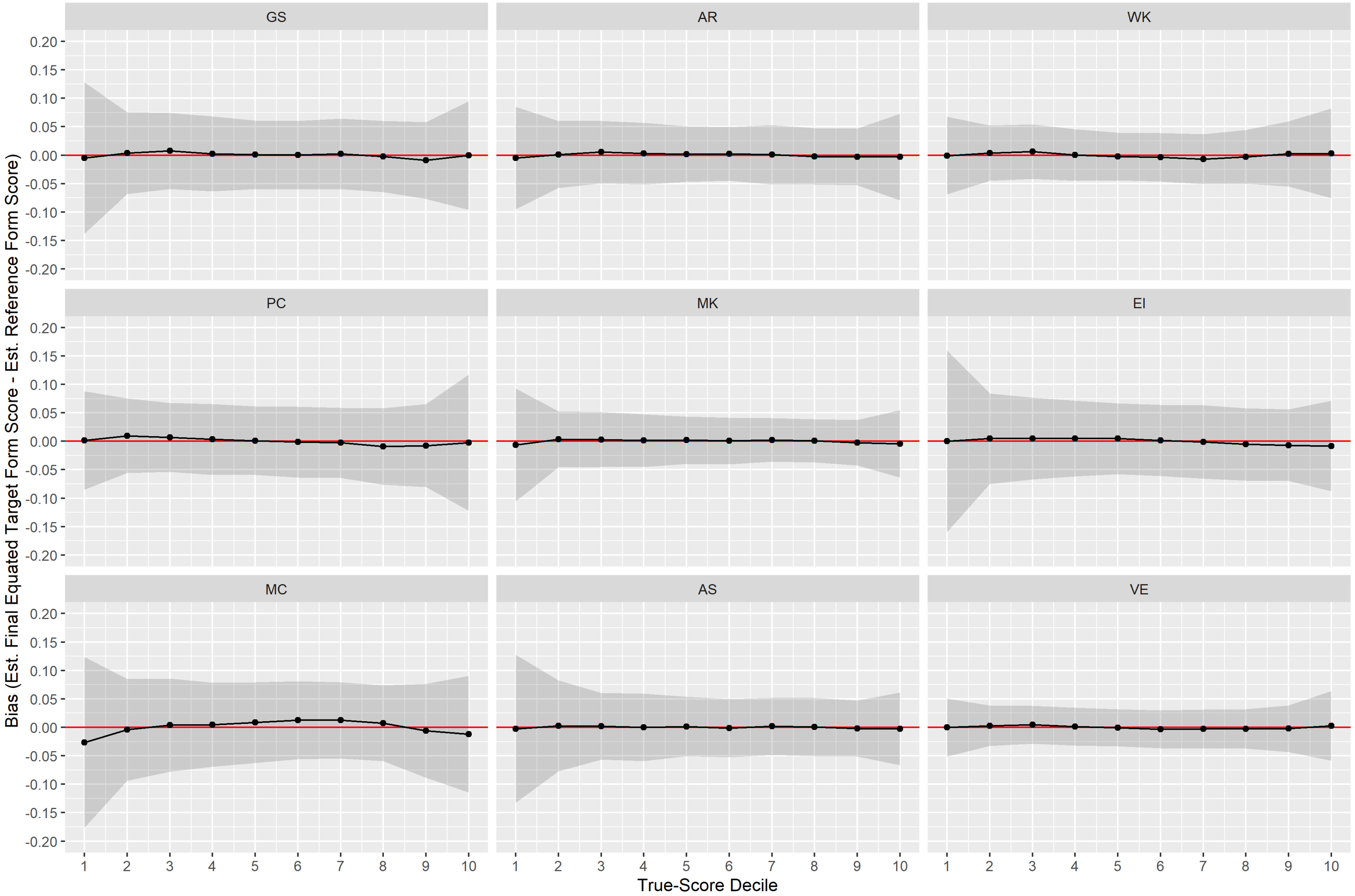
Standard Score	Mean (SD)
GS	0.00 (0.01)
AR	0.00 (0.01)
WK	-0.00 (0.01)
PC	0.00 (0.01)
MK	0.00 (0.01)
EI	-0.00 (0.01)
MC	0.00 (0.01)
AS	-0.00 (0.01)
VE	0.00 (0.01)

Evaluation of Conditional Score Bias (RQ3): By True-Score z Score



- Error ribbons represent 95% confidence intervals.

Evaluation of Conditional Score Bias (RQ3): By True-Score Decile



- Error ribbons represent 95% confidence intervals.
- All conditional mean bias estimates are empirically indistinguishable from zero.