CAT-ASVAB Pool +
P&P- ASVAB Form Development

Presenter: Matthew Trippe, HumRRO

December 15, 2022

HumRRO Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314-1578 | Phone: 703.549.3611 | www.humrro.org

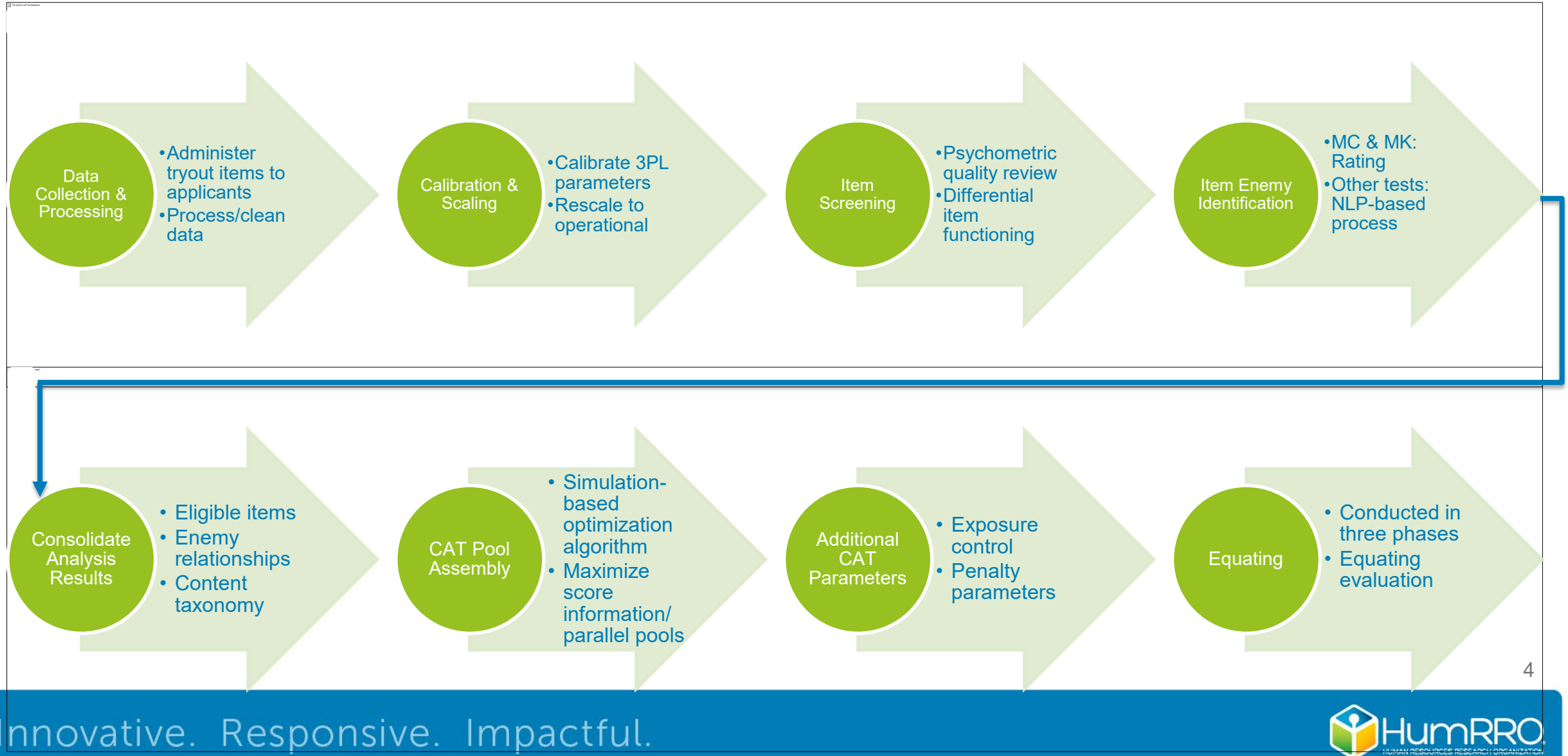
Overview

- Development of CAT-ASVAB Item Pools
  - Process overview
  - Summary/examples of key processes
  - Current status
  - Next steps

- Development of P&P-ASVAB Forms
  - Process overview
  - Summary/examples of key processes
  - Summary of technical challenges & solutions
    - Auto & Shop (AS)
    - Paragraph Comprehension (PC)
  - Current status
  - Next steps

CAT-ASVAB Pool (Form) Development

Process Overview



Tryout Item Data Collection

Annual Pool Development Targets*

Admin Order	Subtest	Pools	Notes
1	General Science (GS)	4	Non-AFQT, moderate threat of compromise
2	Arithmetic Reasoning (AR)	4	AFQT, moderate threat of compromise
3	Word Knowledge (WK)	8	AFQT, greatest threat of compromise
4	Paragraph Comprehension (PC)	4	AFQT, moderate threat of compromise
5	Math Knowledge (MK)	4	AFQT, moderate threat of compromise
6	Electronics Information (EI)	2	Non-AFQT, lower threat of compromise
7	Automotive Information (AI)	2	
8	Shop Information (SI)	2	
9	Mechanical Comprehension (MC)	2	
10	Assembling Objects (AO)	2	

Current Tryout Seeding Design

Group	Seed Sequence	Admin Proportion
1	15 GS, 15 AR	0.29
2	15 WK, 15 PC	0.29
3	15 WK, 15 MK, 15 EI, 15 AI	0.14
4	15 WK, 15 MK, 15 AP	0.14
5	15 SI, 15 MC, 15 AC	0.14

AP = AO Puzzles
AC = AO Connections

*These original targets have been modified per slide 6

Tryout Item Data Collection (cont.)

- Tryout items are developed in “series” of 100 items per test
- Convention of 200 tryout items required to develop one CAT pool
- Tryout series are administered in “seed versions”
  - Current seed version administration configuration:
    - AI, AO, EI, SI, MC : Two series or 200 items
    - AR, GS, MK, PC: Four series or 400 items
    - WK: eight series or 800 items

- Original annual pool development targets (slide 5) are proving to be too aggressive to support from several perspectives
  - Data collection
  - Psychometric team demands
  - Information Technology team demands

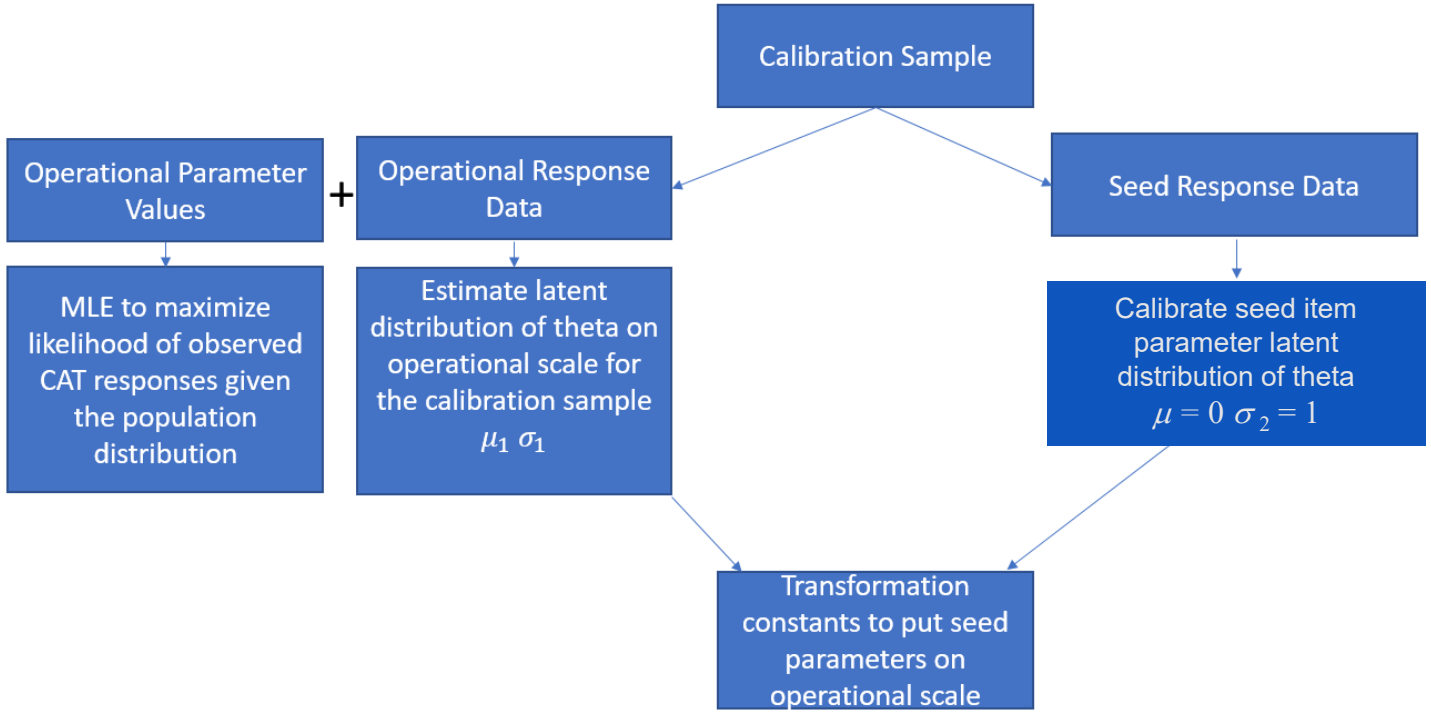
- We are in the process of revising item development and seeding design to be compatible with a “flat” target of 4–5 CAT pools every two years
  - CAT Pools 5–9 operational: 2008–2022
  - CAT Pools 11–15 operational: expected 2023–2025

Item Parameter Calibration

- CAT-ASVAB based on Three-Parameter Logistic model (3PL)
- DTAC simulation studies of calibration process suggest item-level sample size ≥ 1,000 is desirable for optimal parameter recovery
  - Target item-level sample size of 1,200
  - Accounts for some data loss associated with data cleaning (e.g., removal of corrupt or invalid records)
  - Achieving target depends on (variable) testing volumes, but generally requires ~8 months of data collection
- Each test calibrated separately using BILOG-MG
- DTAC simulations find that parameter recovery is improved as the number of seed items administered to each examinee increases
  - Parameter recovery found to be relatively poor when 10 or fewer seed items administered
  - Each examinee responds to 15 randomly administered tryout items per test according to seed design (slide 5)
- Tryout items calibrated in seed versions
  - 200, 400, or 800 items per calibration
- Sparse response data matrix
  - AI, AO, EI, SI, MC: ~16,000 examinees
  - AR, GS, MK, PC : ~32,000 examinees
  - WK: ~64,000 examinees

Item Parameter Rescaling

- Calibrate seed parameter values using seed response data. Latent distribution of theta is fixed to BILOG defaults (0,1)
- Use operational responses from calibration sample + operational parameter values to estimate latent distribution of theta on the operational scale for the calibration sample
- Compute transformation constants to put seed parameters on the operational scale



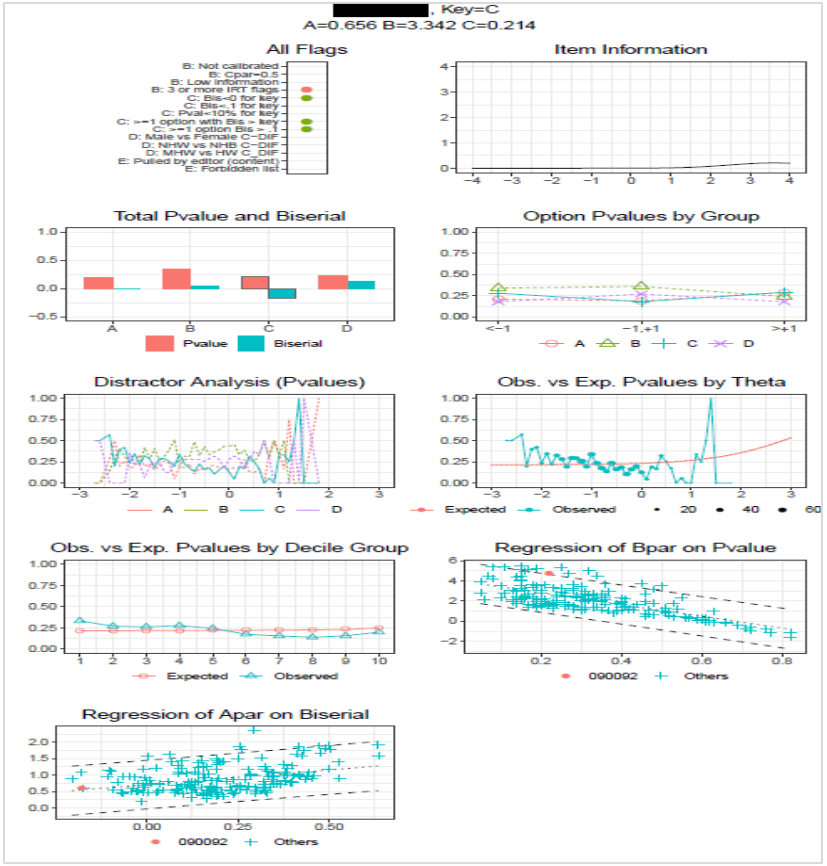
$$A = \frac{\sigma_1}{\sigma_2} \qquad B = \mu_1 - (A \times \mu_2)$$

Empirical Item Screening

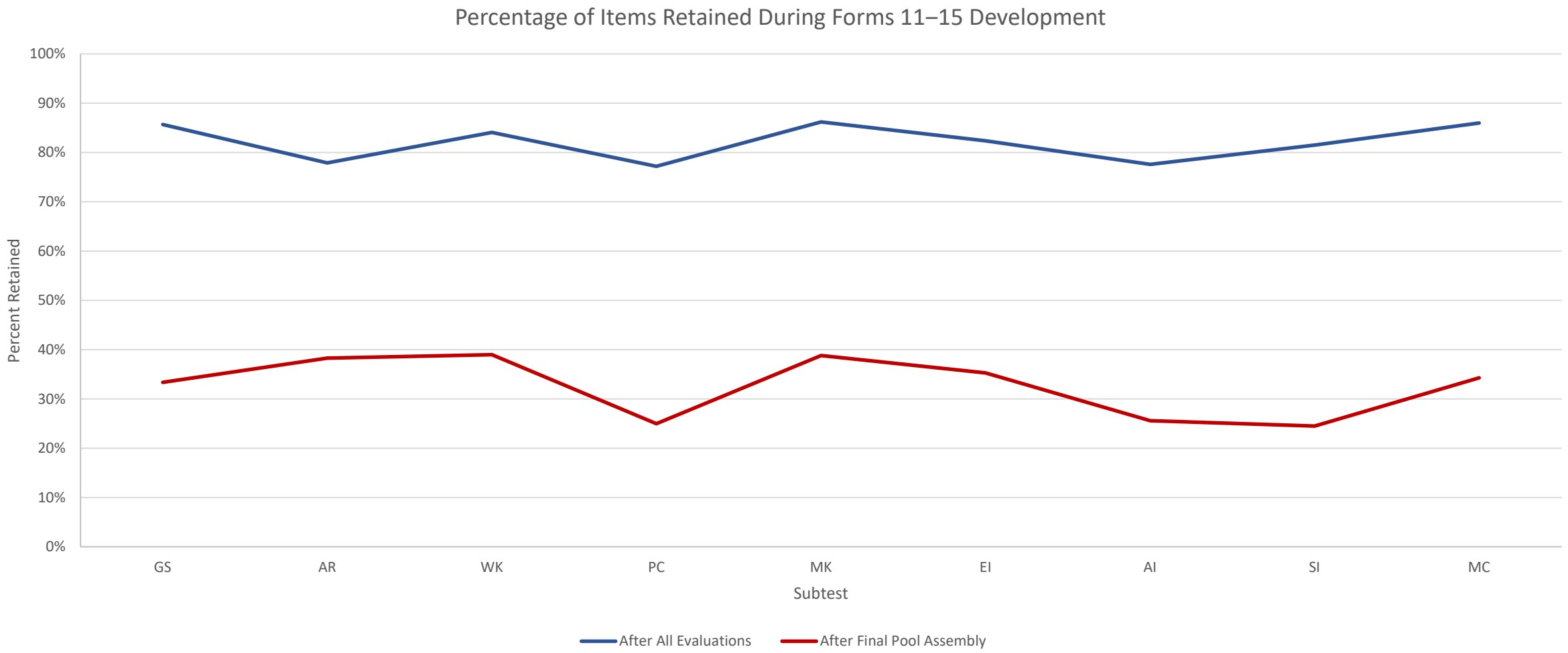
Psychometric Quality Analyses (per item)

- Item information
- Item-model fit
  - Eight fit indices
- Distractor analysis
  - Content review
- Differential item functioning (DIF)
  - Non-Hispanic White vs. Hispanic White
  - Non-Hispanic White vs. Non-Hispanic Black
  - Male vs. Female
  - See 12JUL22 MAPWG briefing for details
- Screening Rubric
  - Many items automatically eligible for operational status
  - Some items automatically ineligible for operational status
  - Several require psychometric/content SME to determine eligibility

"One Pager" Visual Summary (per item)



Empirical Item Screening (cont.)



Item Enemy Analysis

Math Knowledge (MK) & Mechanical Comprehension (MC)

- Pommerich & Segall (2008) evaluated local dependence (LD) in CAT
  - LD in item parameters has minimal effect on precision
  - LD in item responses has substantial effect on precision
- Mitigating LD requires identification of item enemy groups
- Items likely to trigger LD if administered to the same person
  - Two or more items that measure similar or highly related content
- Before assembling forms 5–9, DTAC developed a content framework for identifying enemy groups for tests where LD is of particular concern
  - MC: 95 content areas; 111 content areas as of 2022
  - MK: 155 content areas; 212 content areas as of 2022
- CAT-ASVAB ensures an applicant is administered no more than one item in an enemy group

All Other Tests

- No direct empirical evidence of local dependence affecting item responses in other tests, but we know some items assess similar content
- Existing enemy documentation is limited
  - Item developers cannot know which series will be considered together for pool assembly in the future
  - Definition of enemy is not necessarily based on local dependence
- Evaluating the degree of content similarity among a matrix of >1,000 items per test is a challenging task
- HumRRO has developed methods to optimize human/SME labor & Machine Learning/Natural Language Processing roles

Item Enemy Analysis (cont.)

Process for WK

- Extract the focal word from item stems and the corresponding keyed responses
- Match focal/keyed words from items to a taxonomy of words that relates word forms to root words
  - E.g., in this taxonomy, “deceive,” “deceit,” and “deception” all have the same root word
- Compare focal/keyed root words across items and identify item pairs that assess knowledge of the same word(s)
- Compile pair-wise relations and construct discrete enemy groups

Process for AI, AR, EI, GS, PC, and SI

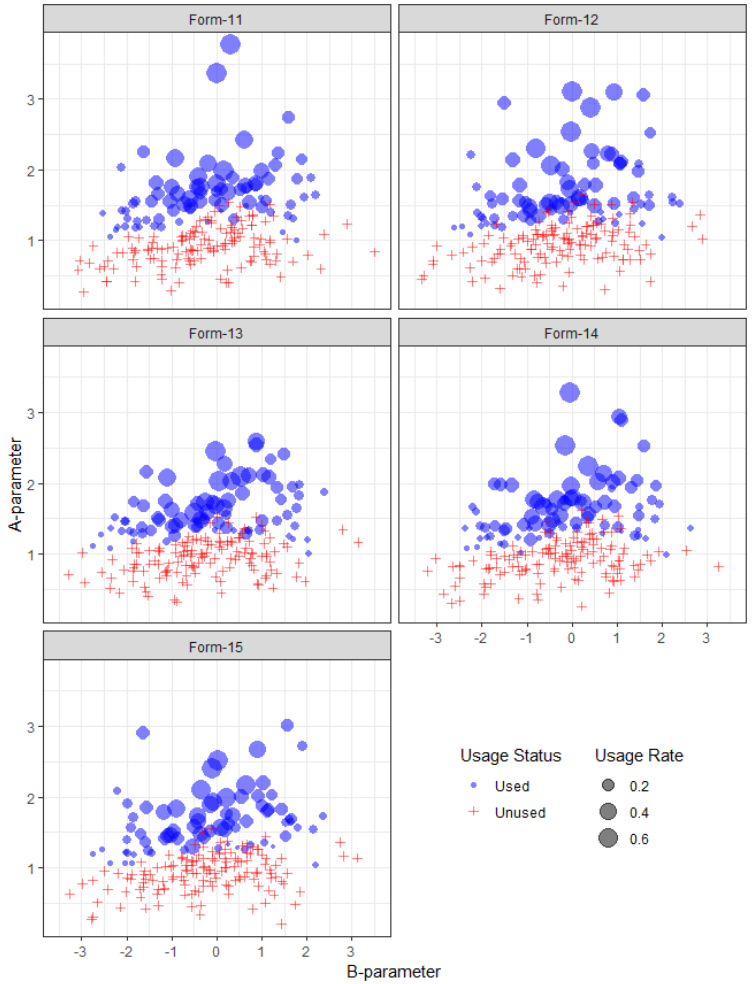
- Compute similarity among item pairs based on quantitative embeddings of item text
- Establish a similarity threshold for potential enemies via subtest-specific bookmarking activity using local dependence focused operational definition of “enemy” specific to each test
- Identify the item pairs above the threshold and review them to eliminate false positives
- Compile pair-wise relations and construct discrete enemy groups

CAT-ASVAB Pool Assembly

- CAT Pools
  - CAT administration is based on pools from which a *potentially* unique set of items is administered to each examinee
  - Pools need to contain items from the full range of content and difficulty
  - Pools need to contain sufficient information/score precision across the full range of ability
- Pool assembly goals
  - For each test, assign each item to one of five pools (e.g., 11, 12, 13, 14, 15)
  - Maximize conditional precision levels of each pool
  - Constrain conditional precision levels to be comparable across pools
  - Account for enemy items—distribute them evenly across pools
  - Account for content taxonomies where applicable (GS, AO)

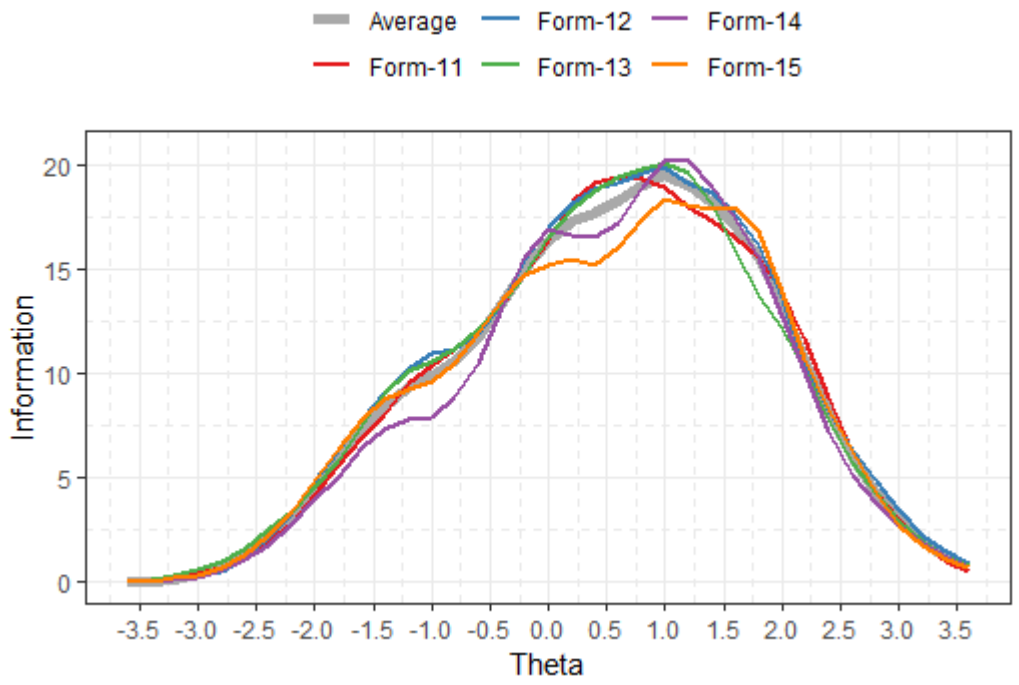
CAT Pool Assembly Example: WK Item Assignment

- Divide eligible items (~1000) into five candidate pools with ~200 items
  - Items appear in only one pool
  - Total information approx. equal across candidate pools
  - Items from the same enemy group constrained to be distributed evenly across candidate pools
- Estimate exposure control parameters via simulation (see slide 16)
- Compute score information functions (SIF) for each candidate pool via large (n>60k) CAT simulation
  - Administer most informative item for given simulee while controlling exposure
  - Items administered at least once assigned to final pool
- "Greedy" algorithm that assigns only the most informative items to pools
- Many eligible items are not assigned to a pool (see slide 10)
  - Attempt to re-use in future pool assembly

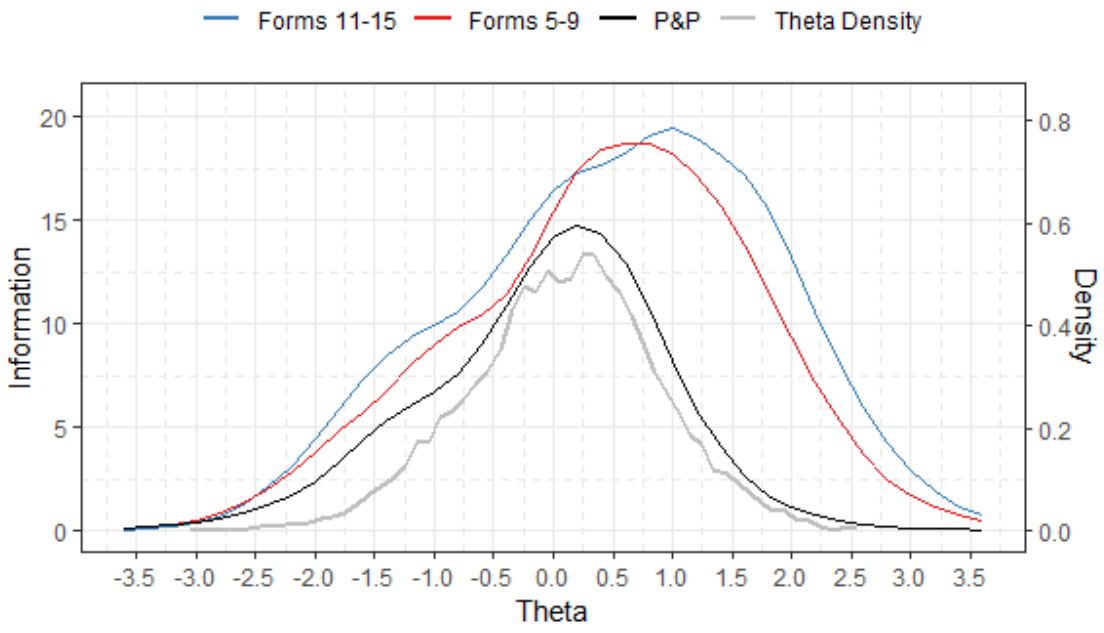


CAT Pool Assembly Example: WK Score Information

Word Knowledge (WK) CAT Pools 11–15



WK CAT Pools 11–15 vs. CAT Pools 5–9 vs. P&P

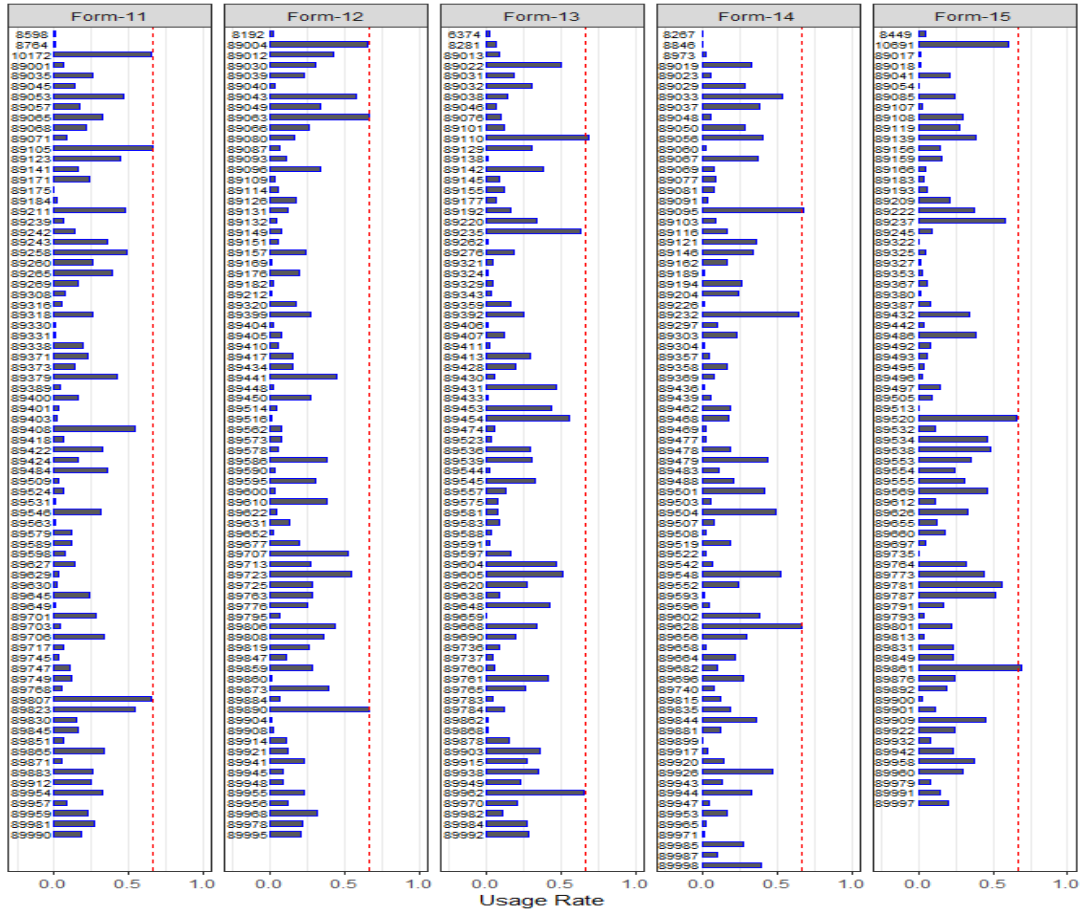


Additional CAT Parameters: Exposure Control

Exposure Control

- Sympson-Hetter exposure control applied to item selection
  - *r* = 2/3, max exposure
  - *P(S)* = probability of selection
    - *P(S) = NS/NE*
  - *P(A)* = probability of administration
    - *P(A) = NA/NE*
  - NE = total examinees
- Multi-step simulation:
  - K_i = 1, initial value for all items
- Iterate until max P(A) ~ *r*
  - Select most informative item
  - Generate random *x* from uniform distribution (0,1)
  - If $x \leq K_i$, then administer item
  - If $P(S) > r$, then $K_i = r/P(S)$
  - If $P(S) \leq r$, then $K_i = 1.0$
- Overall exposure rate of 1/6 in Enlistment Testing Program (ETP)
  - Four operational pools
  - 2/3 (rate) x 1/4 (pools) = 1/6

Usage Rate Example: WK

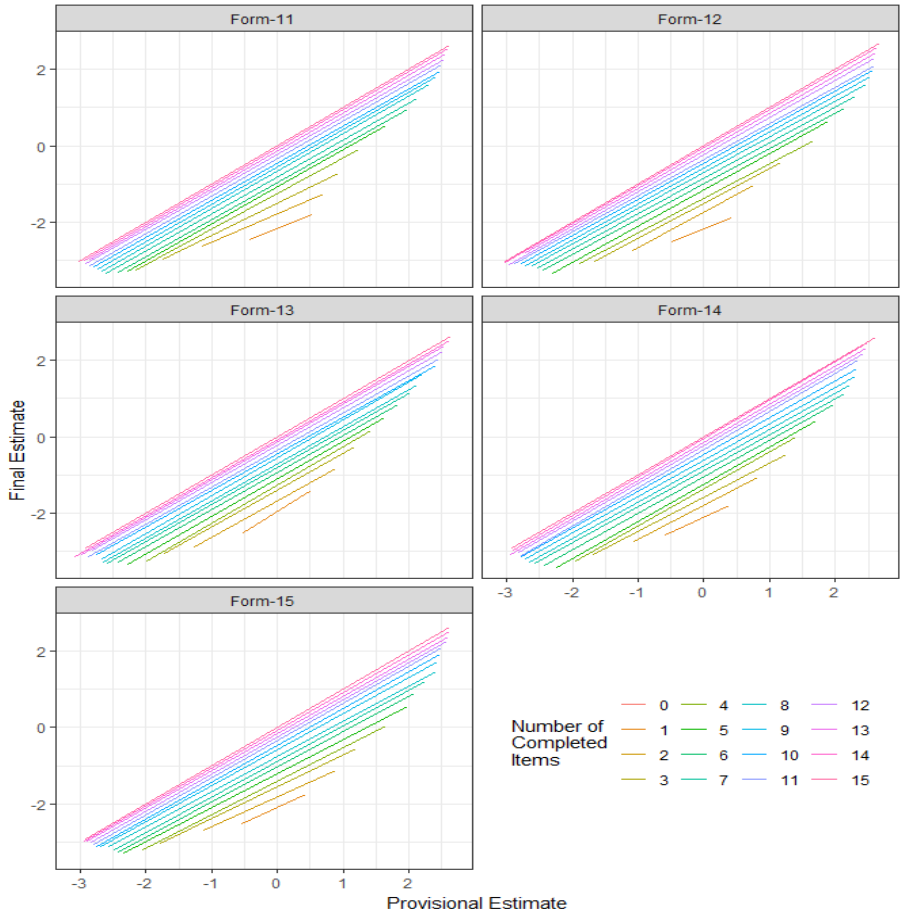


Additional CAT Parameters: Penalty Parameters

Penalty for Incomplete CAT

- CAT-ASVAB scores (Bayes Modal Estimator) contain bias that draws estimate toward mean of prior
- Bias is larger in shorter tests like CAT-ASVAB
- Low-ability examinees could potentially exploit this by answering the minimum number of questions allowed
- Simulation-based penalty procedure assigns a final score that is equivalent to the expected score obtained by random guessing on the unanswered questions
- Penalty functions are regression equations

Penalty Function Example: WK



CAT-ASVAB Equating Study

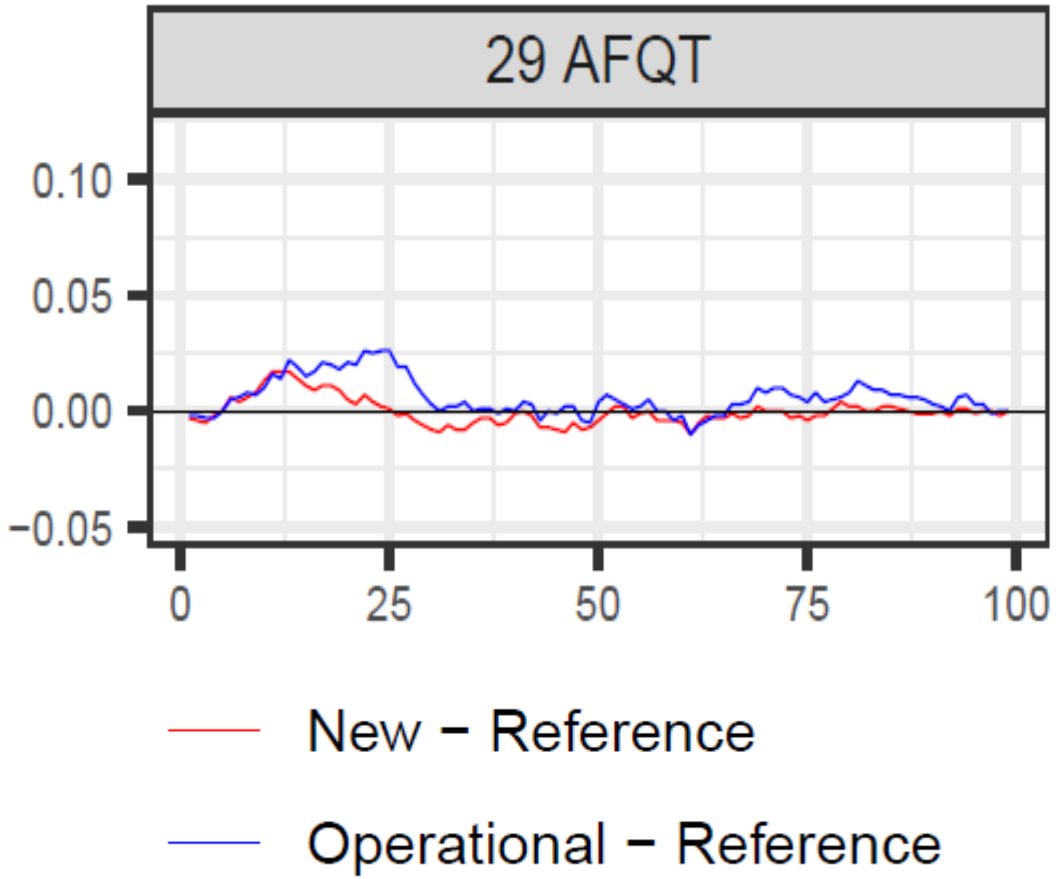
Equating Study Design

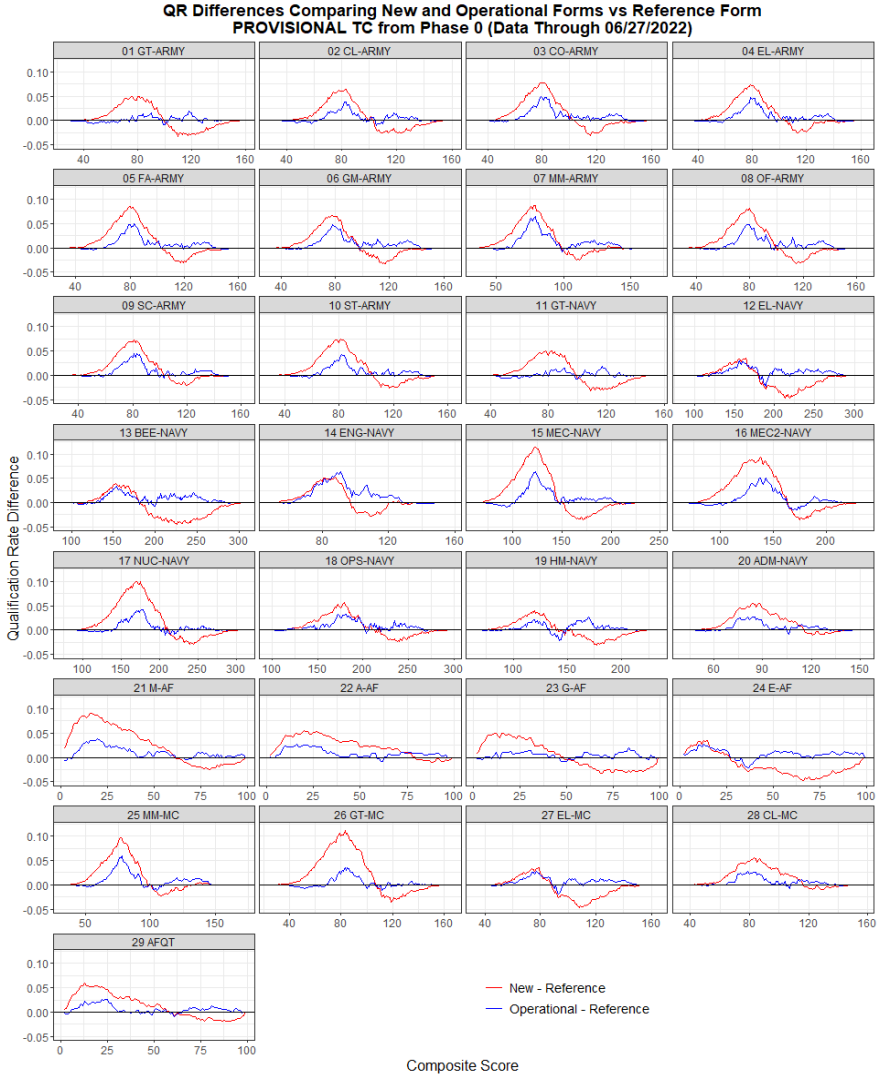
- Equating is implemented in three phases of operational administration of new pools to military applicants
- Each phase includes progressively larger sample size
- Intent of phased design is to maximize accuracy of reported operational scores
- Random groups design
- Each applicant is assigned to a single pool with 1/7 assignment probability
  - The reference form 4, administered only during equating studies
  - An operational form
  - A new form (11–15)
- Evaluate differences in qualification composite cumulative distribution functions (CDFs) between reference form 4 and new pools

Three Phases

Form #	Description	Assignment Probability	Phase I Target n 10JUN22	Phase II Target n 27JUN22	Phase III Target n ~DEC22
4	Reference	1/7	500	1,500	10,000
5	Operational	1/7	500	1,500	10,000
11	New	1/7	500	1,500	10,000
12	New	1/7	500	1,500	10,000
13	New	1/7	500	1,500	10,000
14	New	1/7	500	1,500	10,000
15	New	1/7	500	1,500	10,000
Total		1	3,500	10,500	70,000

CAT-ASVAB Equating: Qualification Rate Differences





CAT-ASVAB Pool Development Status

Current Status

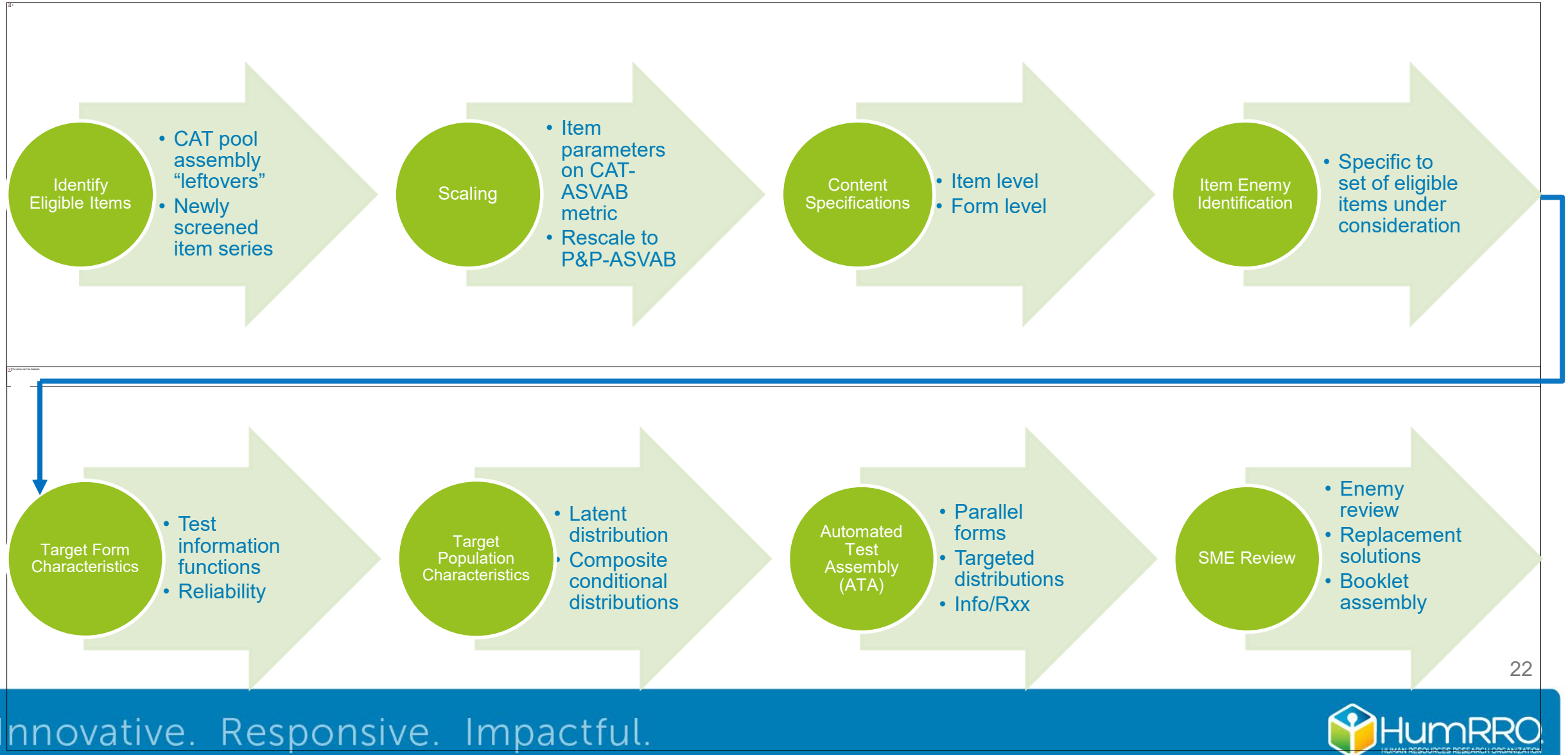
- CAT-ASVAB Pools 11–15
  - Administered to applicants in May 2022 as part of equating study
  - Equating phases 1 & 2 complete
  - Phase 3 target sample size projected to be achieved in mid-December 2022
- CAT-ASVAB Pools 16–20
  - Developed modern computing workflow for pool assembly
  - Run in parallel with original Fortran-based processes
  - Series processed since assembling 11–15
    - WK: 28 series
    - AFQT + GS: 12 series
    - Technical: 4 series

Next Steps

- Complete phase 3 equating for 11–15
  - Final transformation constants
  - Thorough evaluation/analysis
- Begin developing CAT Pools 16–20
  - Use eligible items from:
    - New series processed since 11–15
    - Items not assigned to a pool during 11–15 assembly
    - Items not assigned to P&P-ASVAB

P&P-ASVAB Form Development

Process Overview



P&P-ASVAB Form Development Goals

- Develop new P&P-ASVAB forms to replace existing forms used in the Career Exploration (CEP) and Enlistment Testing Program (ETP)
- CEP has four forms (23A, 23B, 24A, 24B), where A and B versions include the same items reordered
- ETP has four forms (25A, 25B, 26A, 26B), where:
  - A and B versions contain unique items for AFQT tests
  - A and B versions include the same items reordered for non-AFQT tests
- Development of new P&P-ASVAB forms for CEP & ETP has largely been discontinued
- One last wave of development

Eligible Items & Scaling

- Item development is focused on CAT-ASVAB
  - P&P-ASVAB development draws from same resources of eligible items
- Eligible items include:
  - Eligible for CAT-ASVAB Pools 5–9 but not assigned
  - Eligible for CAT-ASVAB Pools 11–15 but not assigned
  - Eligible items from item series processed since development of CAT-ASVAB Pools 11–15

- P&P-ASVAB and CAT-ASVAB are on separate scales
  - DTAC previously conducted “anchoring” study to link P&P-ASVAB scale to CAT-ASVAB scale
  - Latent mean and standard deviations from that study were used to apply linking constants in reverse to place item parameters scaled to CAT-ASVAB (per slide 8) on P&P-ASVAB scale

$$A = SD_{CAT}/SD_{PP} \;\; ; \;\; B = MEAN_{CAT} - MEAN_{PP} * A$$

$$a_{PP} = a_{CAT} * A \;\; ; \;\; b_{PP} = b_{CAT}/A - B/A$$

Content Specifications

Form-Level Blueprint

- Each P&P-ASVAB test has a content blueprint specifying:
  - Number of items
  - Sub-content distribution
    - E.g., AR: whole numbers, rational numbers
    - E.g., GS: life science, physical science
- CEP and ETP blueprints are the same

Test Length

	CAT-ASVAB		P&P-ASVAB	
	Items	Minutes*	Items	Minutes
GS	15	12/25	25	11
AR	15	55/113	30	36
WK	15	9/18	35	11
PC	10	27/75	15	13
MK	15	31/65	25	24
EI	15	10/21	20	9
AI	10	7/18	25**	11**
SI	10	7/17		
MC	15	22/42	25	19
AO	15	18/38	25	15

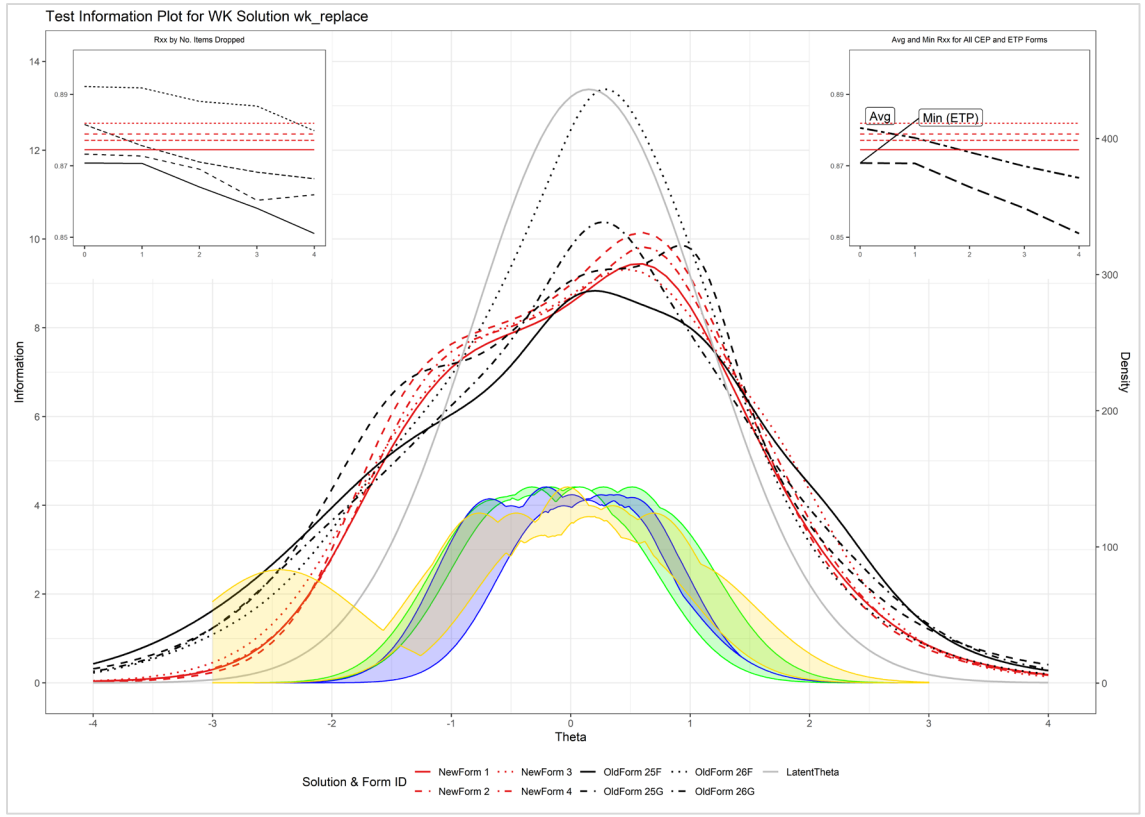
*Without/With tryout items (see slide 5 for design)
**Auto-Shop (AS) = 13 AI + 12 SI items
Limits are set so that at least 99% of examinees can finish in the allotted time

Automated Test Assembly

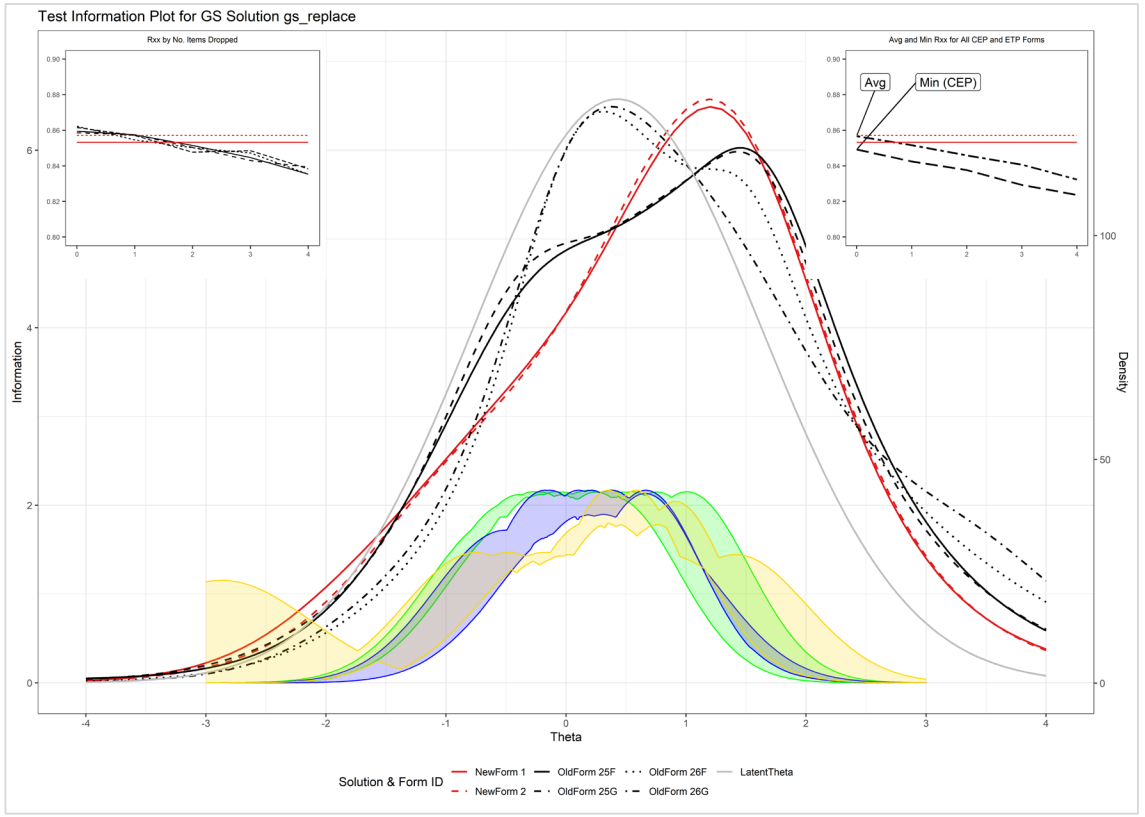
- Use Automated Test Assembly (ATA) optimization model to develop forms parallel to each other and “target” CEP/ETP forms
- Model constraints include
  - Number of items
  - Content blueprint
  - Item key “balance”
  - Item enemies
  - Maximizing the test information functions (TIFs)
  - Minimizing equally weighted sum of the distance between TIFs and test characteristic curves (TCCs) of the forms
- Quantitative evaluation criteria include
  - Similarity to “target” CEP/ETP form TIF/Rxx
  - Alignment with latent distribution
  - Alignment with latent distribution conditional on aptitude area composites
- Final SME review
  - Review assembled form content for evidence of
    - Enemies
    - Obsolete content
    - “Sensitive” content

Automated Test Assembly (cont.)

Example: WK



Example: GS



P&P-ASVAB Technical Challenges & Solutions

Innovative. Responsive. Impactful.

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

P&P-ASVAB Technical Challenges

Auto & Shop (AS)

- When P&P-ASVAB was originally developed, Auto & Shop (AS) was calibrated/scaled as one test
- CAT-ASVAB items are calibrated/scaled as separate Automotive Information (AI) and Shop Information (SI) tests, which are subsequently combined as a composite
- P&P-ASVAB must include AS-scaled item parameters to be compatible with MEPCOM infrastructure
- Rescaling options include:
  - Special data collection to administer new AI & SI items + backup/reserve items + original AS items, followed by calibration and final form assembly
    - Impractical and risky
  - Modified Stocking-Lord Procedure (MSLP)

Paragraph Comprehension (PC)

- When P&P-ASVAB was originally developed, Paragraph Comprehension (PC) items were developed with 5 questions per paragraph stimulus
- CAT-ASVAB items are developed with one question per paragraph stimulus
- Maintaining a fifteen-item PC test would result in increasing the number of paragraphs from three to fifteen
- Twelve additional paragraph stimuli will dramatically increase word count (~125%) and thus increase the **time limit**
- Testing time is extremely valuable, and considerably increasing the time limit will be problematic for CEP and ETP

P&P-ASVAB Technical Solutions: AS

Approach

- Modified Stocking-Lord Procedure (MSLP) for two tests/scales transformed to a common scale
- Iteratively trying out sets of transformation constants (A [scale] and B [location] constants) and searching for the set that best minimizes our objective function
  - AI and SI each has a set of constants that are optimized simultaneously
- Objective function is the sum of squared differences between:
  - Expected number-correct scores based on (a) sets of parameters on the AI and SI scales and (b) a simulated distribution of true-score AI and SI thetas; and
  - Expected number-correct scores based on (a) a set of parameters that have been rescaled using provisional constants and (b) the average of the true-score AI and SI thetas (i.e., true-score AS thetas)

Key Idea: The typical Stocking-Lord Procedure links different sets of item parameters (for a common set of anchor items) with the latent distribution held constant. We can extend this logic to link different latent distributions by rescaling a single set of item parameters.

P&P-ASVAB Technical Solutions: AS (cont.)

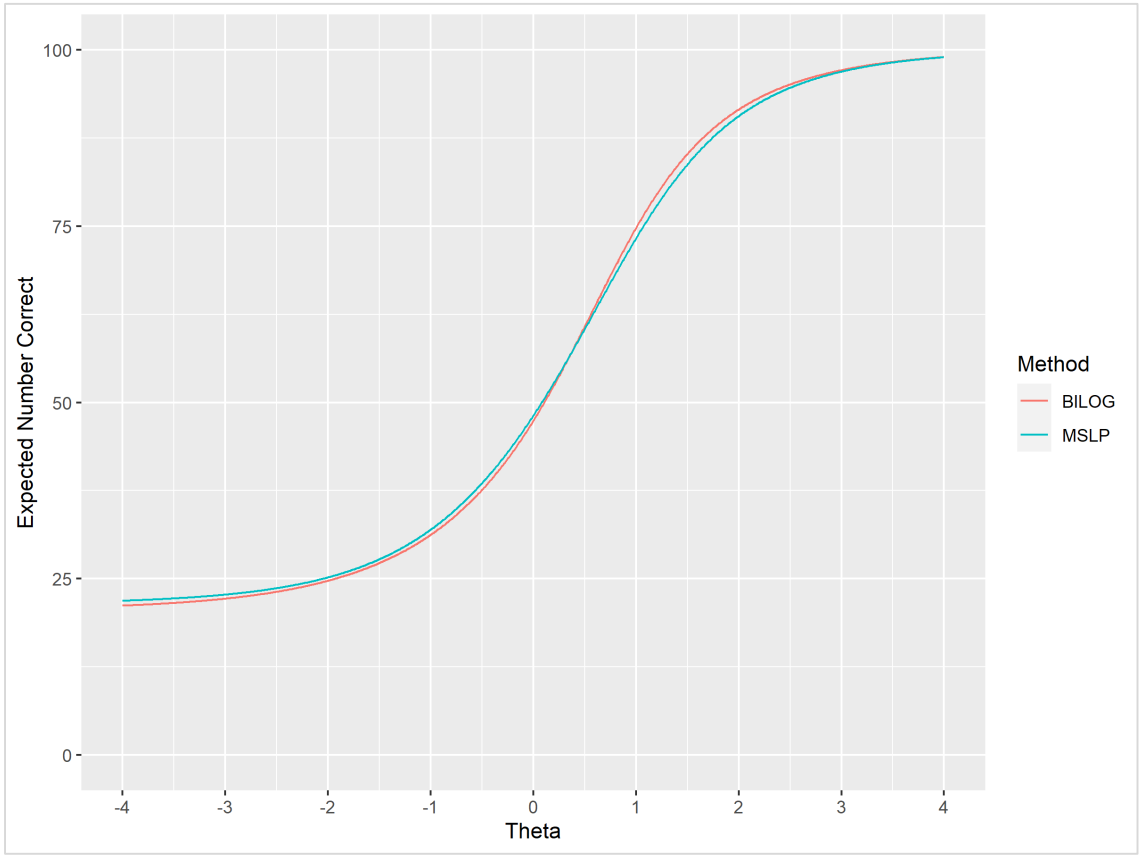
Evaluation

- Compared MSLP to item parameters estimated by calibrating AI and SI items together in BILOG-MG*
  - **Simulation 1:** Large-*N* single-form proof of concept
    - Purpose: Determine if the scaling procedure works as expected under ideal conditions with ample data
    - Simulated 50 AI items, 50 SI items, and 10k "simulees" (fixed form; no seeding or randomized administration)
    - Calibrated AI and SI separately with BILOG-MG, then applied MSLP
    - Calibrated AI and SI together with BILOG-MG
  - **Simulation 2:** 100 forms assembled from items calibrated using a joint AI+SI seeding design
    - Purpose: Determine if the scaling procedure works as expected with 25-item forms
    - Simulated 200 AI items, 200 SI items, and 16k simulees (15 random items per subtest per simulee → ~1.2k simulees per item)
    - Calibrated AI and SI separately with BILOG-MG, then randomly assigned 25 items (13 AI + 12 SI) to forms and applied MSLP
    - Calibrated AI and SI together with BILOG-MG, then matched AS-scaled parameters with the assembled forms
- Consistent results in both simulations
  - Very close correspondence between test characteristic curves (TCCs) for MSLP- and BILOG-scaled parameters
  - Slightly lower test information functions (TIFs) for MSLP, but BILOG-based TIFs are likely inflated due to violating assumption of unidimensionality

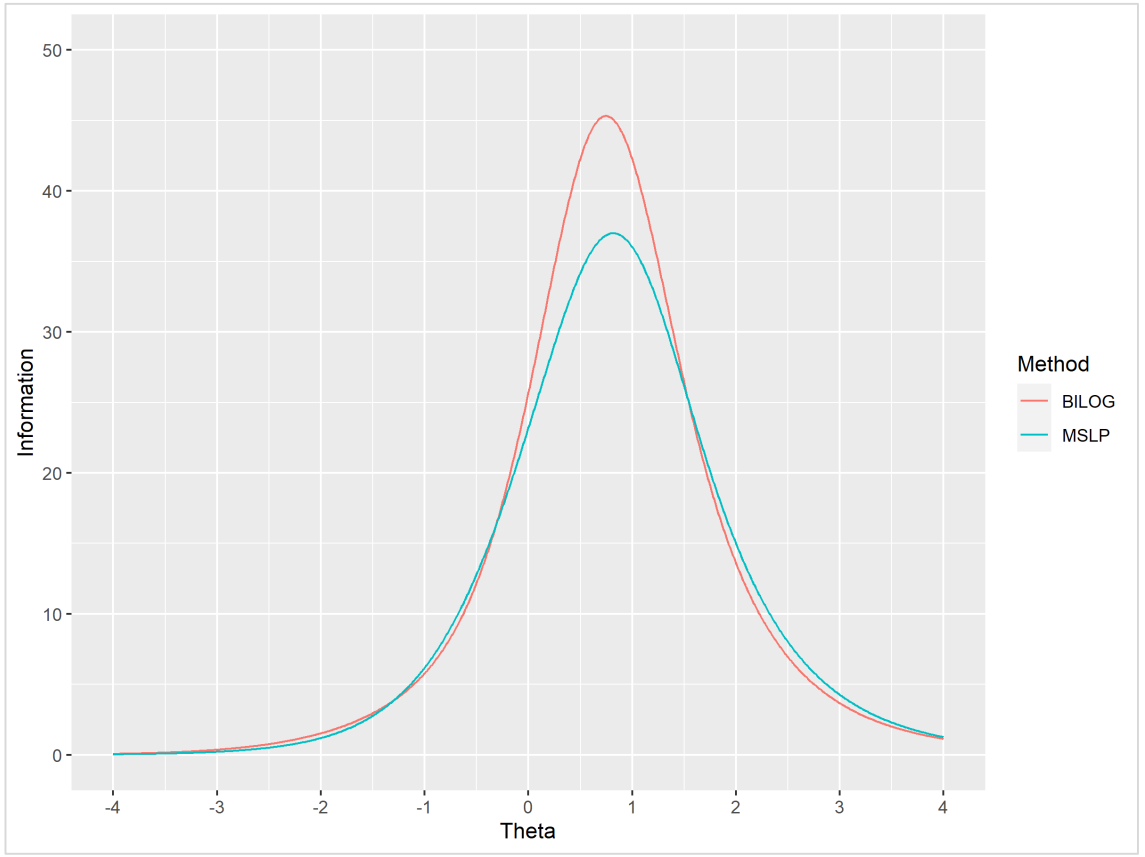
*BILOG-MG calibration includes DTAC's established parameter-rescaling process

P&P-ASVAB Technical Solutions: AS (Simulation 1)

TCC

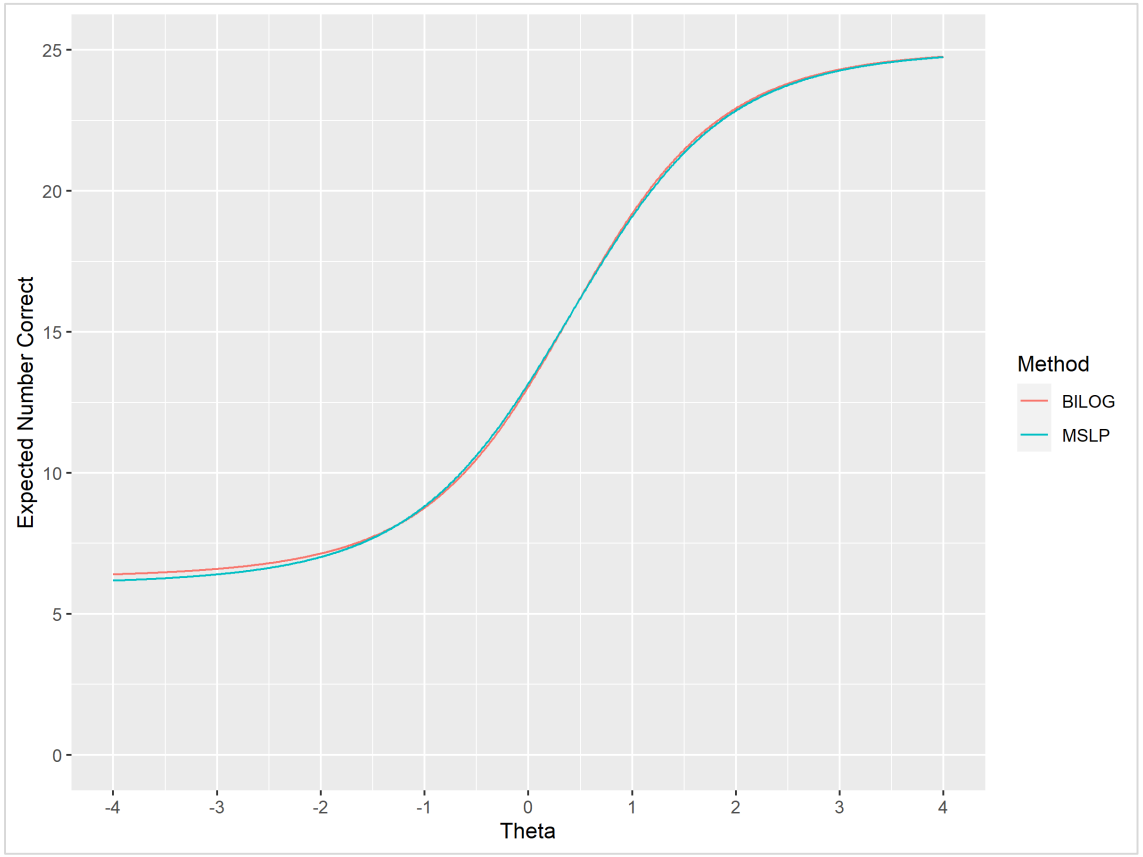


TIF

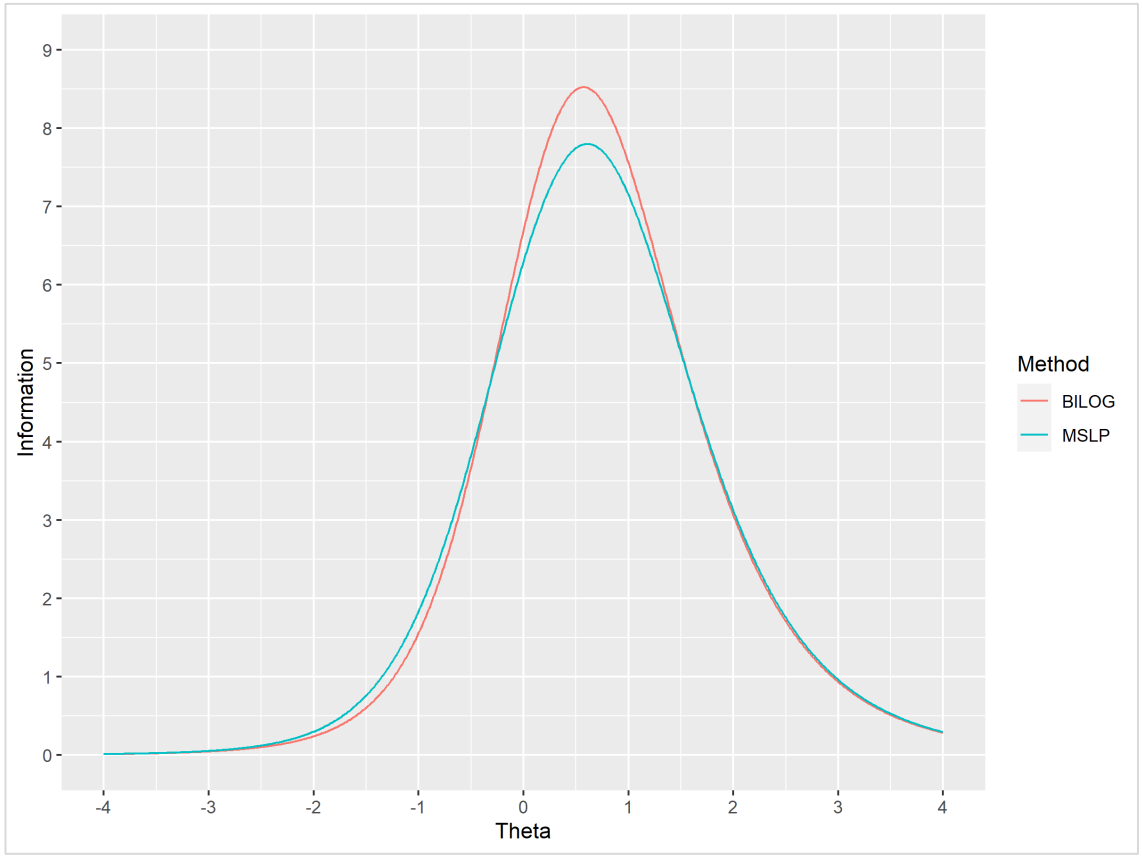


P&P-ASVAB Technical Solutions: AS (Simulation 2)

Average TCC Across Forms



Average TIF Across Forms



P&P-ASVAB Technical Solutions: PC

Solution

- Modify automated test assembly (ATA) optimization algorithm
- Existing constraints
  - Content blueprint
  - Item key “balance”
  - Item enemies
  - Maximize the test information functions (TIFs)
  - Minimize equally weighted sum of the distance between TIFs and test characteristic curves (TCCs) of the forms
- New Constraint
  - Minimize projected response time
- Variable
  - Number of items (9–15)

Summary Findings

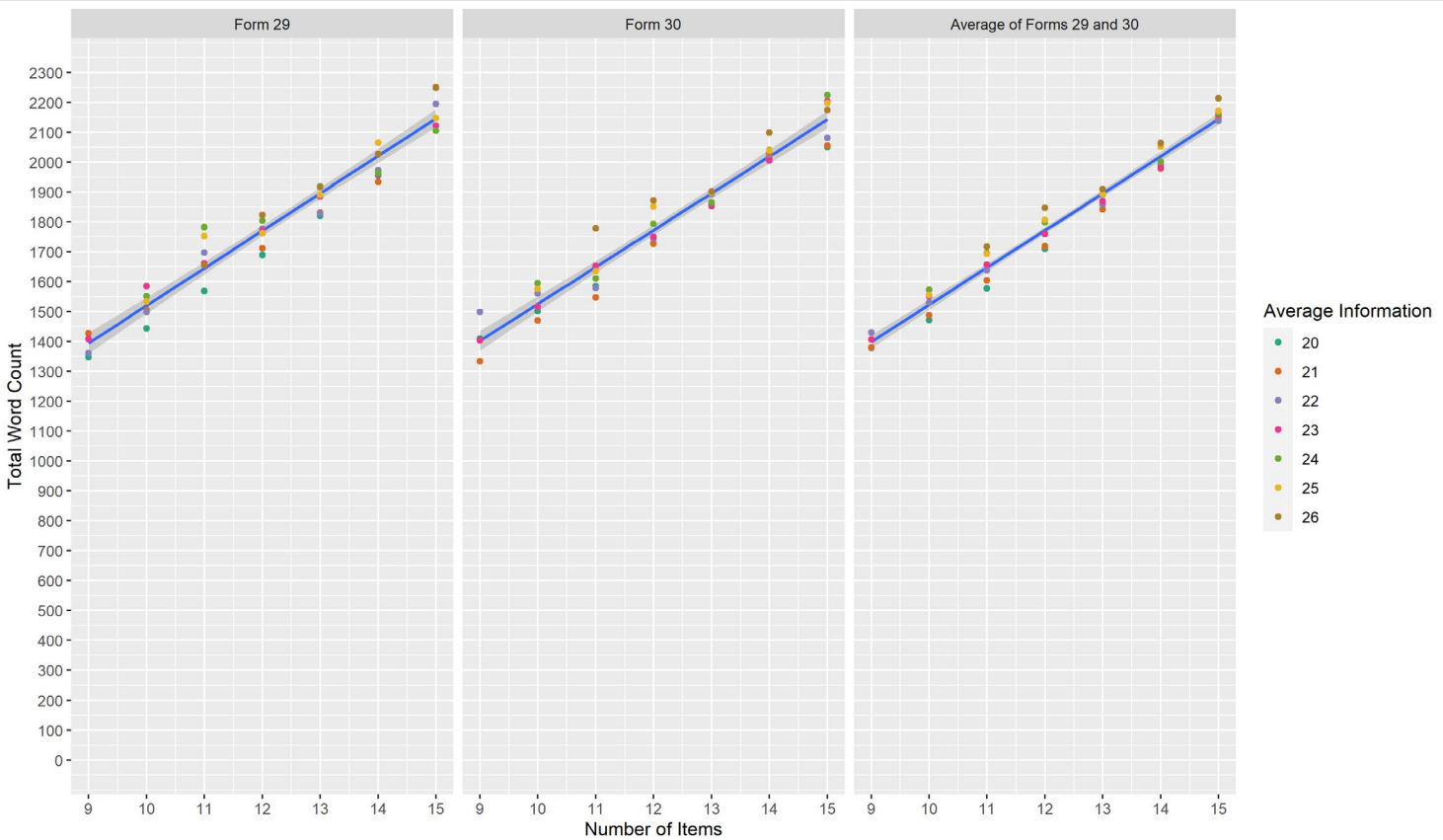
- Ten-item solution is optimal for
  - Minimizing response time while
  - Minimizing loss of test-level reliability and
  - Maintaining composite reliability
- The current P&P-ASVAB PC time limit is 13 minutes
  - Unlike CAT-ASVAB where time limit is imposed on the individual, this is a test session time limit that applies to all test takers in a proctored environment
- DTAC will provide a recommended solution at an upcoming MAPWG when the details are finalized

P&P-ASVAB Technical Solutions: PC (cont.)

PC

- There is an obvious/strong relationship between number of items and word count
- More items = more words = longer testing time
- 15-item solution projected time limits > 20 minutes, nearly doubling current time limit of 13 minutes, which is prohibitive

Number of PC Items vs. Total Word Count

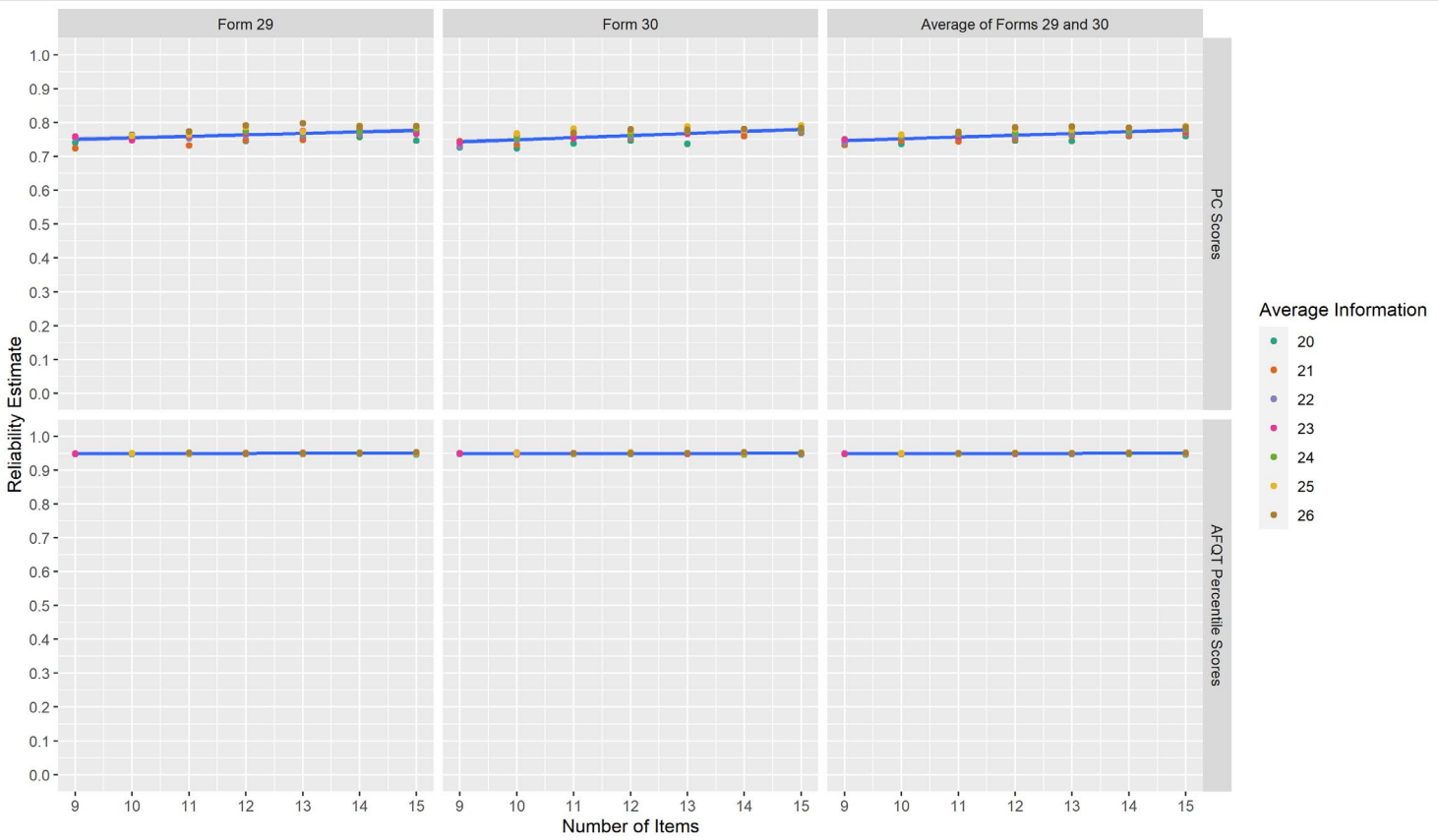


P&P-ASVAB Technical Solutions: PC (cont.)

PC

- There is a much weaker relationship between number of items and
  - reliability of PC scores
  - reliability of AFQT percentile scores
- 10-item solution represents optimal compromise between Rxx and projected testing time
- DTAC will prepare recommendation based on comprehensive research

Number of PC Items vs. Test-Retest Rxx



P&P-ASVAB Form Development Status

Current Status

- Career Exploration Program and Enlistment Testing Program forms
  - All solutions except PC are complete, QC'd, and ready for delivery to DTAC
  - Final AS solutions mostly unaffected by scaling decision, but experimenting with some "what if" scenarios regarding MSLP order of operations
  - ETP MK solution is ready for delivery, but another "what if" analysis underway as we wait for PC and AS solutions to be fully resolved and implemented

Next Steps

- Finalize PC under "unified" approach
  - Develop 6 parallel PC forms (2 CEP + 4 ETP) using latest optimization algorithm
- Finalize AS scaling decision
  - Heavily favoring MSLP approach over additional data collection
- Summarize research findings and present recommendation to MAPWG
- Finalize all P&P form deliverables
  - CAT-ASVAB Pools 16–20 eligible items become known
- Return full focus of project team to CAT-ASVAB pool development

DAC Guidance & Feedback

P&P-ASVAB Form Development

- Process was created for this last wave of development
- No intention to repeat the process in the future

- Comment or concerns over solutions to technical challenges faced with AS and PC?

- Questions or concerns over other aspects of this process?

- Other recommendations, observations, or advice?

HumRRO Project Team

- Maura Burke
- Jeff Dahlke
- Ted Diaz
- Olga Golovkina
- Ki Ho Kim
- Matthew Reeder
- Stephen Robertson
- Matthew Trippe
- Liz Waterbury