OFFICE OF PEOPLE ANALYTICS

Exploring the Efficacy of Using Natural Language Processing and Machine Learning to Enhance CAT-ASVAB Form Development

Ted Diaz and Olga Golovkina
Human Resources Research Organization

Briefing presented to the DACMPT
June 12, 2024

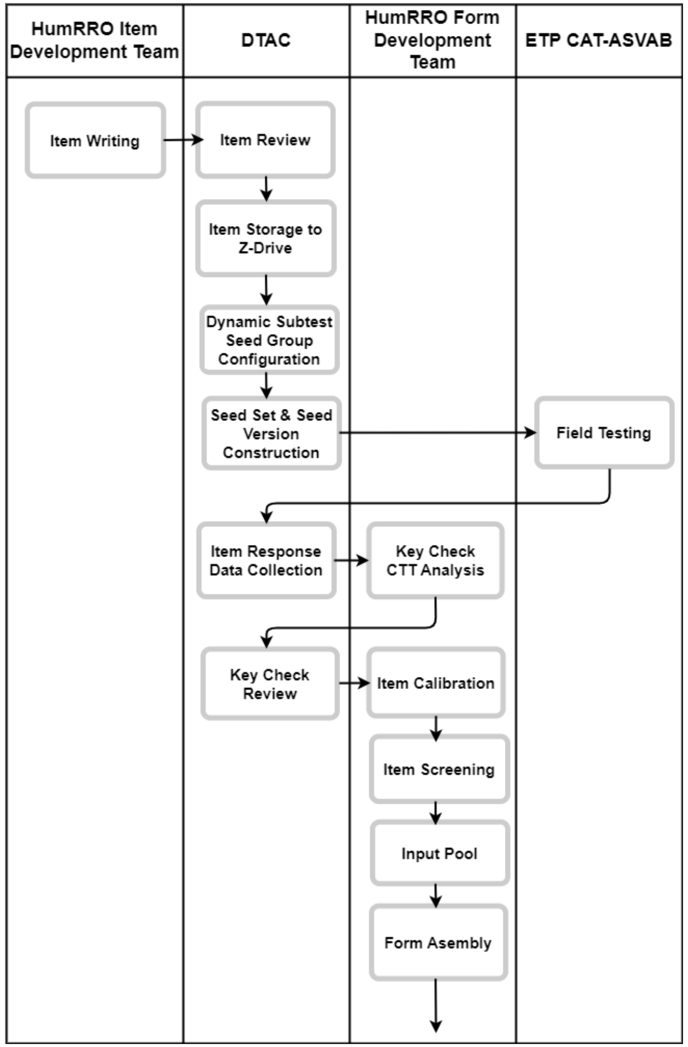
Briefing Agenda

- Background Information
- Recommended Process Improvements
- Supporting Analyses
- Summary
- Questions for the DAC

Background Information

CAT-ASVAB form development is a laborious process

- CAT-ASVAB form development involves several stages over several years and affects multiple stakeholders
- However, individual item and CAT form performance, in terms of resource and psychometric efficiency, becomes apparent only at the final form assembly stage, when little can be done about candidate items considered for assignment to operational forms



HumRRO Item Development Team

DTAC

HumRRO Form Development Team

ETP CAT-ASVAB

Item Writing

Item Review

Item Storage to Z-Drive

Dynamic Subtest Seed Group Configuration

Seed Set & Seed Version Construction

Field Testing

Item Response Data Collection

Key Check CTT Analysis

Key Check Review

Item Calibration

Item Screening

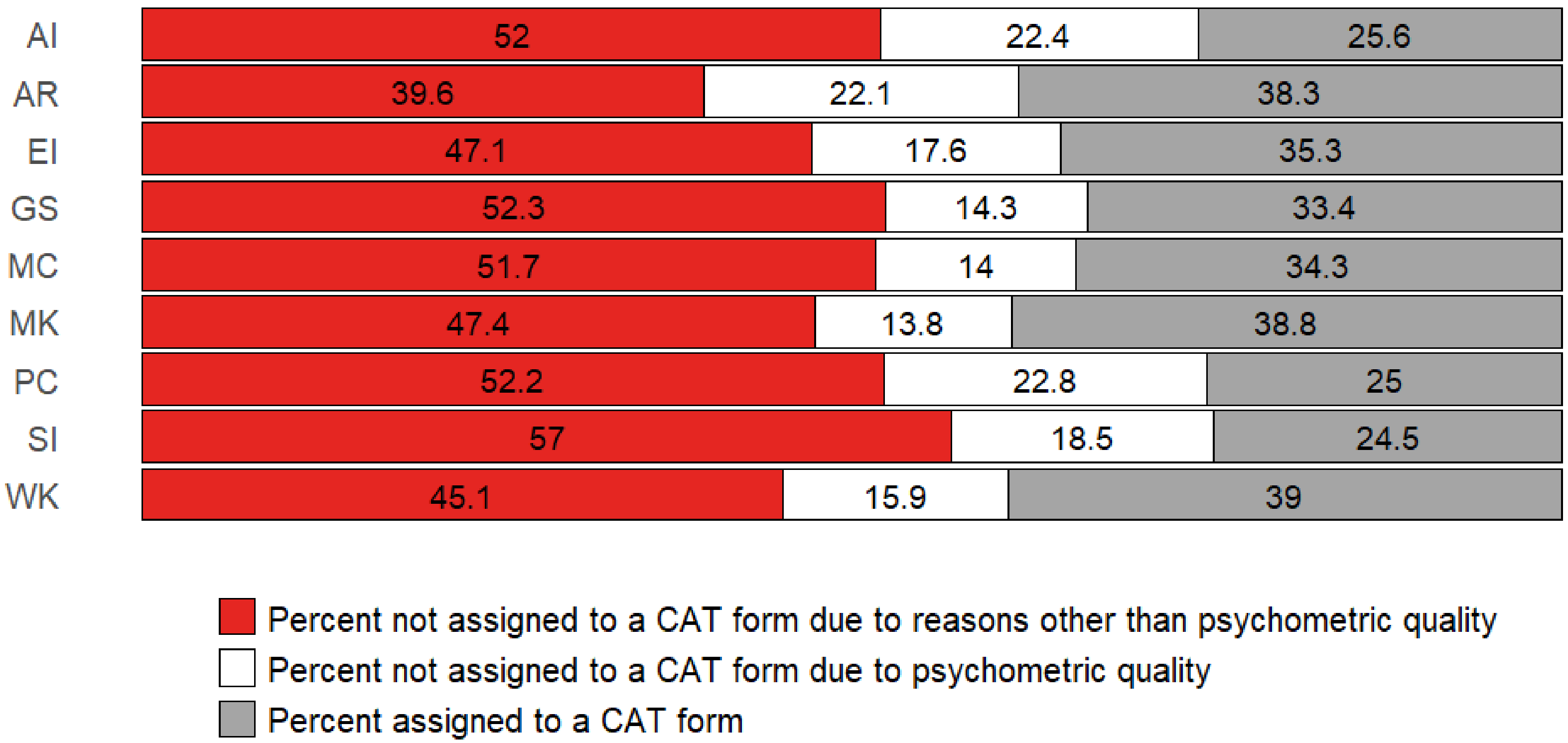
Input Pool

Form Asembly

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.

Most developed items do not make it into the final CAT-ASVAB forms

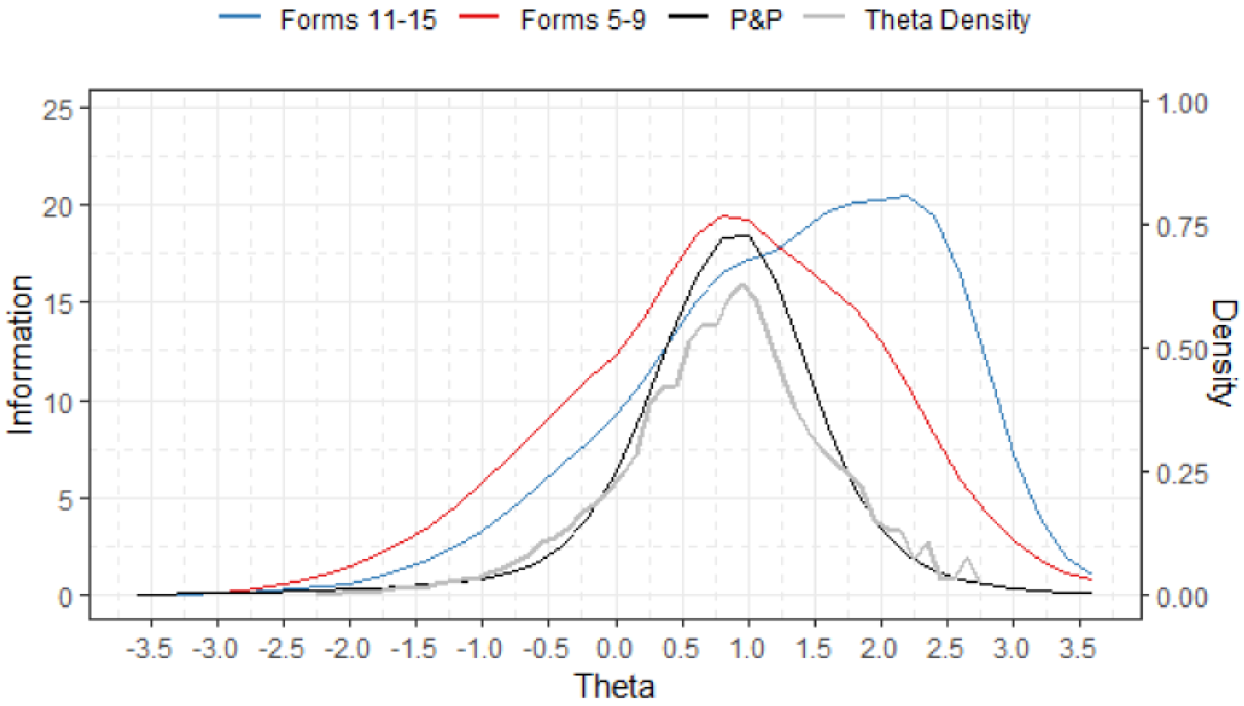
During the development of CAT-ASVAB Forms 11-15, only about 25-39% of seed items were assigned to a CAT form



OPA
OFFICE OF PEOPLE ANALYTICS

Mathematics Knowledge (MK)

- Score information for CAT-ASVAB Forms 11-15 is highest in the high ability range, where we expect few examinees, and lower in the low to moderate ability range, where we expect most examinees
- There is a surplus of difficult items
- Drifting score information function (SIF) can also raise concerns about parallelism of future forms



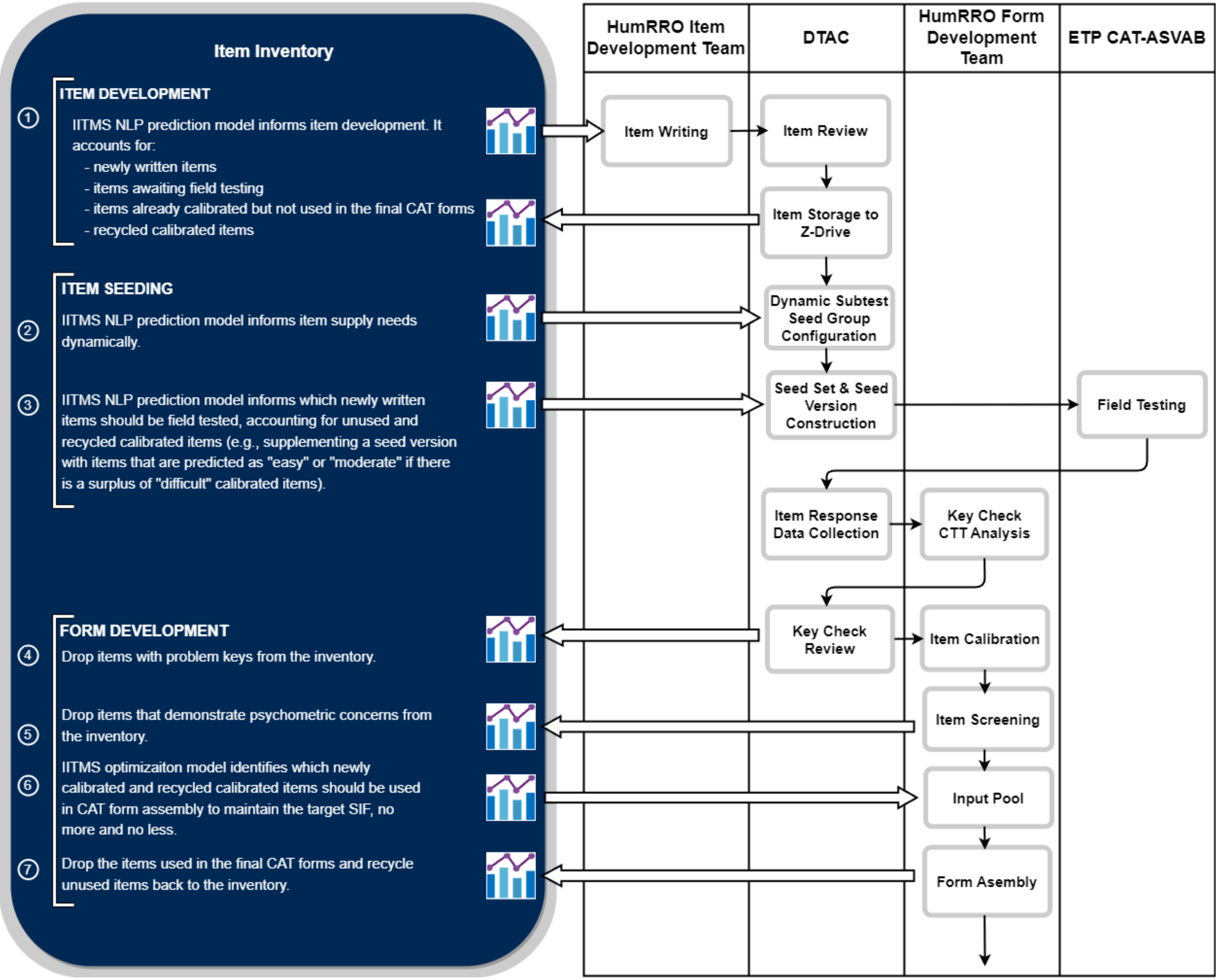
Recommended
Process Improvements

ASVAB Item Inventory Total Management System (IITMS)

- A total system approach to enhance CAT form development that integrates three components:
  - Natural language processing (NLP) prediction model
    - Predicts item quality and difficulty from item content
    - Guides item development and seeding to ensure steady supply of items with targeted characteristics for CAT form assembly
  - CAT optimization model
    - Optimally selects items from the inventory to make CAT form construction efficient
    - Moderates CAT algorithm “greediness” for highly discriminating items while maintaining psychometric targets
  - System simulation
    - Applies the NLP prediction and CAT optimization models within a total system simulation of the form development process to inform decisions at key steps—item development, field testing, and CAT form assembly

ASVAB Item Inventory Total Management System (IITMS)

- IITMS supports more frequent CAT form replacement, allowing:
  - Item writers to develop items with desired quality and difficulty that are aligned with psychometric targets
  - DTAC to identify item supply needs dynamically (e.g., moderate or easy items) and field test items accordingly
  - The form development team to efficiently assemble CAT forms, using only items needed to maintain test information targets, saving desirable items for future form development cycles



Item Inventory

ITEM DEVELOPMENT

① IITMS NLP prediction model informs item development. It accounts for:
- newly written items
- items awaiting field testing
- items already calibrated but not used in the final CAT forms
- recycled calibrated items

ITEM SEEDING

② IITMS NLP prediction model informs item supply needs dynamically.

③ IITMS NLP prediction model informs which newly written items should be field tested, accounting for unused and recycled calibrated items (e.g., supplementing a seed version with items that are predicted as "easy" or "moderate" if there is a surplus of "difficult" calibrated items).

FORM DEVELOPMENT

④ Drop items with problem keys from the inventory.

⑤ Drop items that demonstrate psychometric concerns from the inventory.

⑥ IITMS optimizaiton model identifies which newly calibrated and recycled calibrated items should be used in CAT form assembly to maintain the target SIF, no more and no less.

⑦ Drop the items used in the final CAT forms and recycle unused items back to the inventory.

HumRRO Item Development Team	DTAC	HumRRO Form Development Team	ETP CAT-ASVAB
Item Writing	Item Review Item Storage to Z-Drive Dynamic Subtest Seed Group Configuration Seed Set & Seed Version Construction Item Response Data Collection Key Check Review	Key Check CTT Analysis Item Calibration Item Screening Input Pool Form Asembly	Field Testing

NLP Prediction Model

- Features:
  - Item metadata (e.g., content taxonomy category)
  - Syntactic features (e.g., average number of words before main verb, average prepositional phrase count, etc.)
  - Complexity and readability features (e.g., monosyllable count, Flesch Reading Ease, etc.)
  - Cognitive features (e.g., familiarity rating, concreteness rating, etc.)
  - Mathematics tokens (e.g., powers and polynomials by type, fraction by type, etc.) for MK items
  - Additional features (e.g., difficulty of the target word, prevalence of the target word, etc.) for WK items
- True labels are based on thresholds set for A and B Item Response Theory (IRT) parameter estimates:
  - We label an item as "high quality" if A is greater than the threshold and "not high quality" otherwise
  - We label an item as "difficult" if B is greater than the threshold and "not difficult" otherwise
  - For items that are "not difficult," we label an item as "easy" if B is below an additional threshold and "moderate" otherwise
- Models and architectures:
  - Model predicting high-quality items: two-layer neural network with 16 and 8 nodes in the first and second layers, respectively
  - Model predicting difficult items: two-layer neural network with 16 and 8 nodes in the first and second layers, respectively
  - Model predicting easy items: two-layer neural network with 16 and 12 nodes in the first and second layers, respectively
  - The neural networks predict labels based on thresholds on probabilities of item quality and difficulty labels

CAT Optimization Model

- ASVAB CAT automated test assembly (ATA) algorithm maximizes information across theta for each form while minimizing difference in information between forms (i.e., parallelism) in a form development cycle
  - Since there are no formal information constraints, it is a greedy algorithm that selects the most informative items available in the input pool
  - Information can drift from one development cycle to another based on the item quality within the input pool, with excess or deficit for specific ability ranges compared to previous forms (see slide 6)
  - Items that are not selected are considered in future form development cycles, but they are typically lower in quality
- The proposed CAT enhancement, Optimal CAT (Optim-CAT) ATA, adds information targets and optimal sampling of items to the current CAT ATA
  - It is analogous to mixed-integer-linear-programming-based ATA methods for fixed forms
  - It samples items optimally from item clusters of similar quality and difficulty using cluster-specific sampling proportions
  - It selects just enough items of given quality and difficulty to achieve information targets
- Benefits of Optim-CAT ATA:
  - Produces consistent information from one form development cycle to another
  - Works well with the item-reuse policy that combines fresh items with items not used in previous cycles
  - Circumvents ATA greediness when using fresh items, making it no longer necessary to globally sample from fresh items (a naive fix) to save some quality items for future form development cycles
  - Effectively replaces the convention of using 200 tryout items per form, allowing CAT ATA to use all available items

CAT Optimization Model

- Bayesian optimization implementation of Optim-CAT
  - S1: Construct a provisional form and evaluate information as follows:
    - Assign items to C item clusters with similar A and B parameters (alternatively, similar item information)
    - Sample items from each cluster using cluster-specific sampling proportions $P_1, \dots, P_C$ (set by the algorithm in S3)
    - Input sampled items to ATA algorithm and compute information from the constructed CAT forms
  - S2: Expensive black box objective function
    - Evaluate the root-mean-square deviation (RMSD) of new-form information relative to target information
    - Gaussian process with $P_1, \dots, P_C$ as parameters uses a *surrogate model* to probabilistically approximate RMSD
  - S3: Iterate between S1 and S2, using Bayesian optimization to identify “promising” values of $P_1, \dots, P_C$
    - Bayesian optimization algorithm identifies $P_1, \dots, P_C$ that will likely minimize RMSD
    - Improve the Gaussian process approximation of RMSD using parameter values $P_1, \dots, P_C$ and “true” RMSD across iterations
    - Stop after the given maximum number of iterations
  - S4: Optimally sampled items correspond to minimum RMSD across iterations

System Simulation

- Statistical model of the sequential data generation process underlying CAT form development
  - To mimic item development, one can sample from curated items (i.e., items that have certain desired characteristics after NLP prediction modeling) based on target characteristics
  - To mimic CAT form assembly, one can use the Optim-CAT ATA algorithm, while ignoring item enemies and using items not used in previous form development cycles
  - One can then evaluate the psychometric properties of constructed forms and the remaining items in the inventory
- One can apply the statistical model above to evaluate what-if scenarios

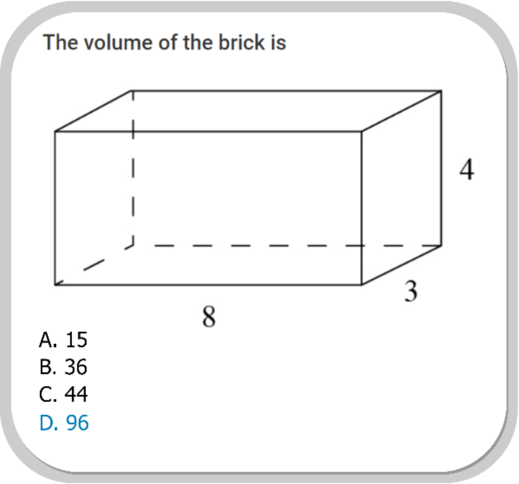
Supporting Analyses

OPA

OFFICE OF PEOPLE ANALYTICS

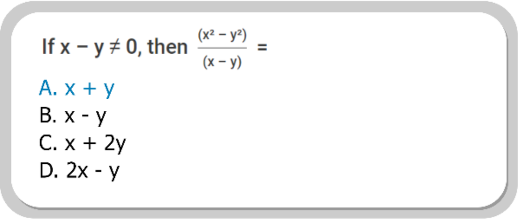
Mathematics Knowledge (MK)

- Analyses:
  - Item quality and difficulty modeling using NLP
  - CAT form development simulation
    - Without CAT optimization and with CAT optimization
    - Input pool size efficiency analysis
- 2,749 MK items used



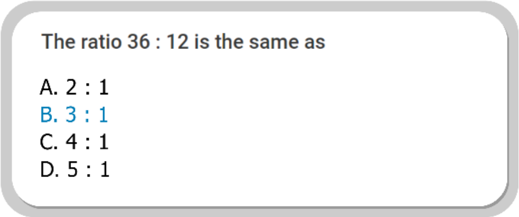
The volume of the brick is

A. 15
B. 36
C. 44
D. 96



If $x - y \neq 0$, then $\frac{(x^2 - y^2)}{(x - y)} =$

A. x + y
B. x - y
C. x + 2y
D. 2x - y



The ratio 36 : 12 is the same as

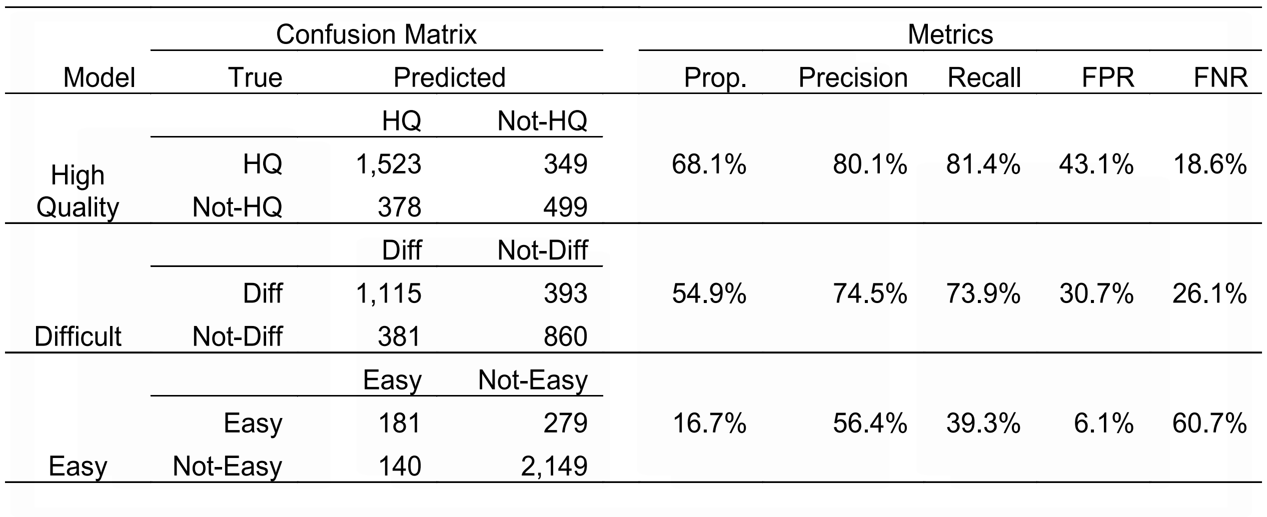
A. 2 : 1
B. 3 : 1
C. 4 : 1
D. 5 : 1

Sample MK Items; Source: https://www.officialasvab.com/mathematics-knowledge-mk/

NOTE: In addition to the analyses above, we performed NLP prediction modeling, CAT form development simulation without CAT optimization, and input pool size efficiency analysis for the remaining Armed Forces Qualification Test (AFQT) subtests—Arithmetic Reasoning (AR), Paragraph Comprehension (PC), and Word Knowledge (WK).

MK: NLP models predict item quality and difficulty with accuracy

- We assigned the following item quality and difficulty labels to all items:
  - True “high-quality” and predicted “high-quality” items (i.e., “True HQ” and “Pred HQ”)
  - True “not-high-quality” and predicted “not-high-quality” items (i.e., “True Not-HQ” and “Pred Not-HQ”)
  - True “difficult” and predicted “difficult” items (i.e., “True Diff” and “Pred Diff”)
  - True “not-difficult” and predicted “not-difficult” items (i.e., “True Not-Diff” and “Pred Not-Diff”)
  - True “easy” and predicted “easy” items (i.e., “True Easy” and “Pred Easy”)
  - True “not-easy” and predicted “not-easy” items (i.e., “True Not-Easy” and “Pred Not-Easy”)

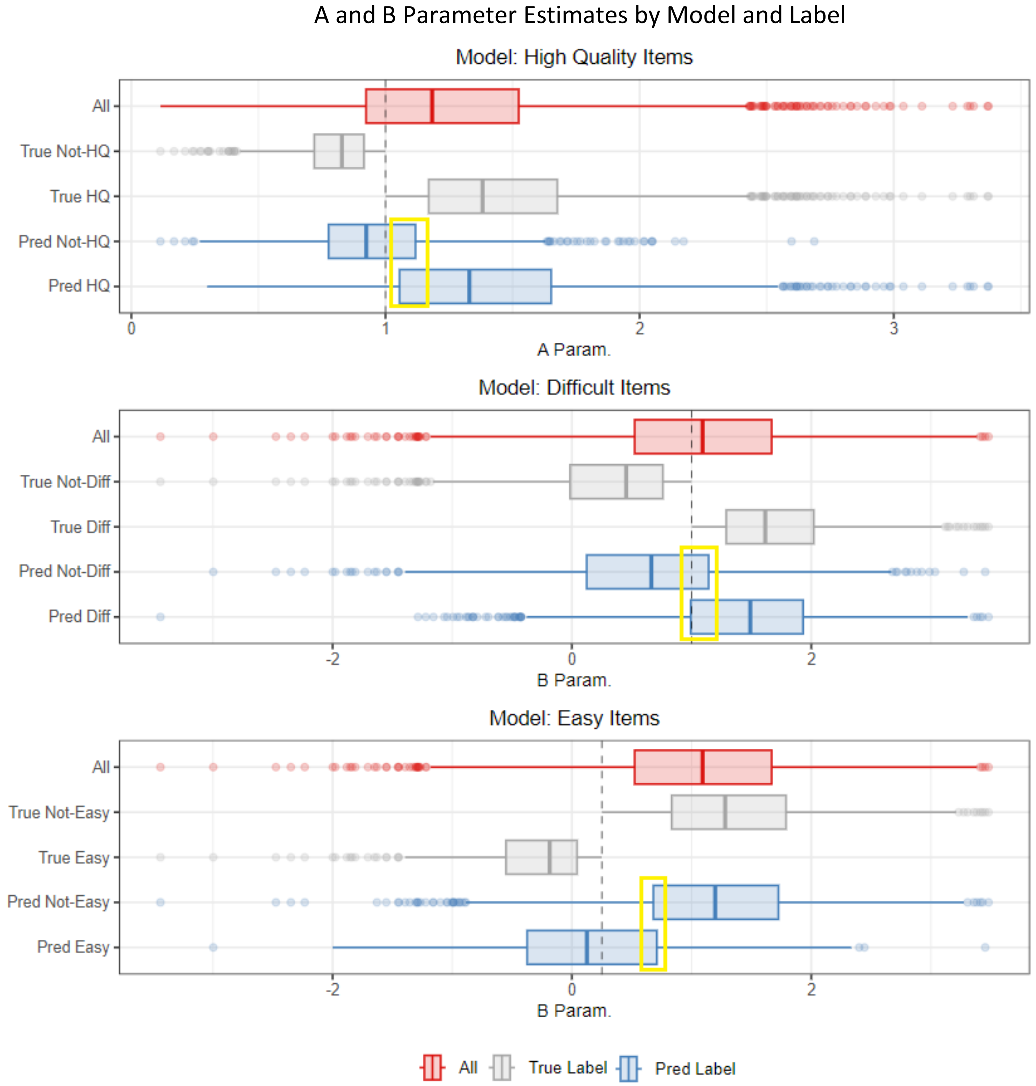


	Confusion Matrix			Metrics				
Model	True	Predicted		Prop.	Precision	Recall	FPR	FNR
		HQ	Not-HQ					
High Quality	HQ	1,523	349	68.1%	80.1%	81.4%	43.1%	18.6%
	Not-HQ	378	499					
		Diff	Not-Diff					
	Diff	1,115	393	54.9%	74.5%	73.9%	30.7%	26.1%
Difficult	Not-Diff	381	860					
		Easy	Not-Easy					
	Easy	181	279	16.7%	56.4%	39.3%	6.1%	60.7%
Easy	Not-Easy	140	2,149					

OFFICE OF PEOPLE ANALYTICS

MK: NLP models predict item quality and difficulty with accuracy

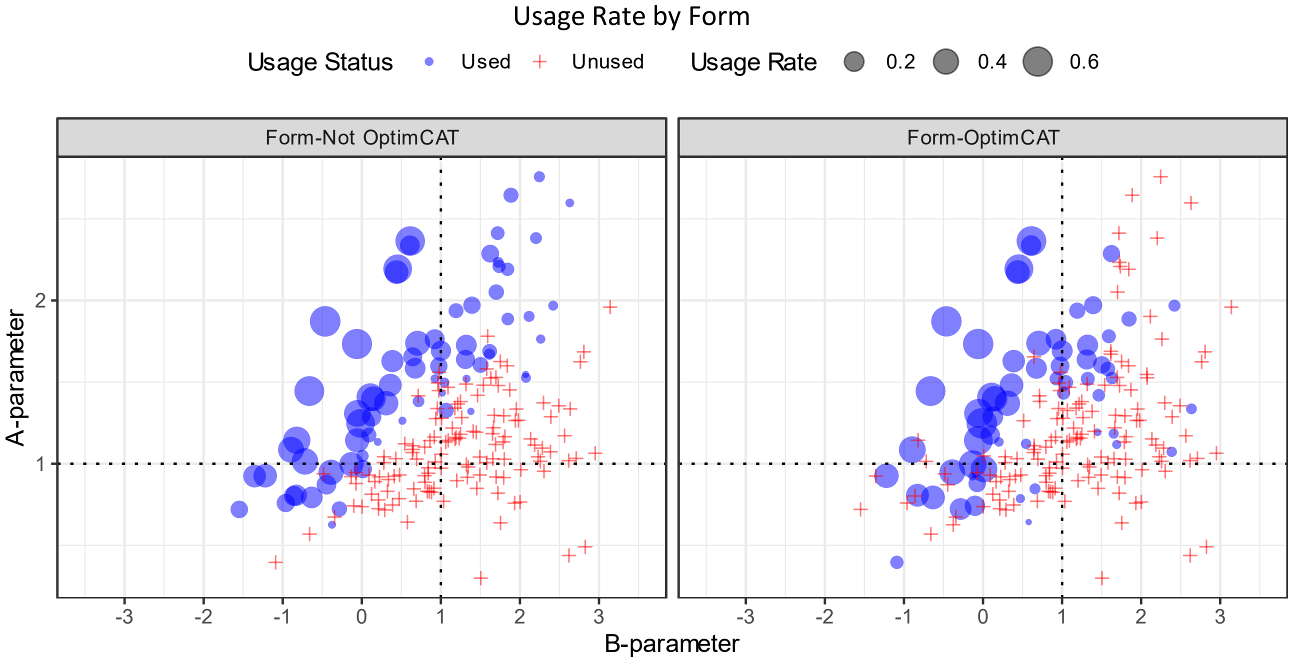
- We compared the A and B parameter estimates of all items, items grouped by true label, and items grouped by predicted label to assess model performance
- There is good separation between predicted group memberships across the three models, with small overlap between the middle 50% (i.e., the interquartile range) of the blue distributions
- The NLP model can provide reliable feedback that can be used to guide item development and item seeding as well as to develop CAT forms



OFFICE OF PEOPLE ANALYTICS

MK: Form assembly with CAT optimization can moderate the "greediness" of the ATA algorithm

- The ATA algorithm is "greedy" and selects the most informative items (i.e., items with highest A parameter) across examinee ability range
- Over time, this effect is compounded and the items remaining in the inventory are not sufficient to construct additional forms that meet targets
- Form assembly with CAT optimization can mitigate this effect, saving quality items for future cycles and, over time, increasing the number of forms that meet or maintain the target SIF



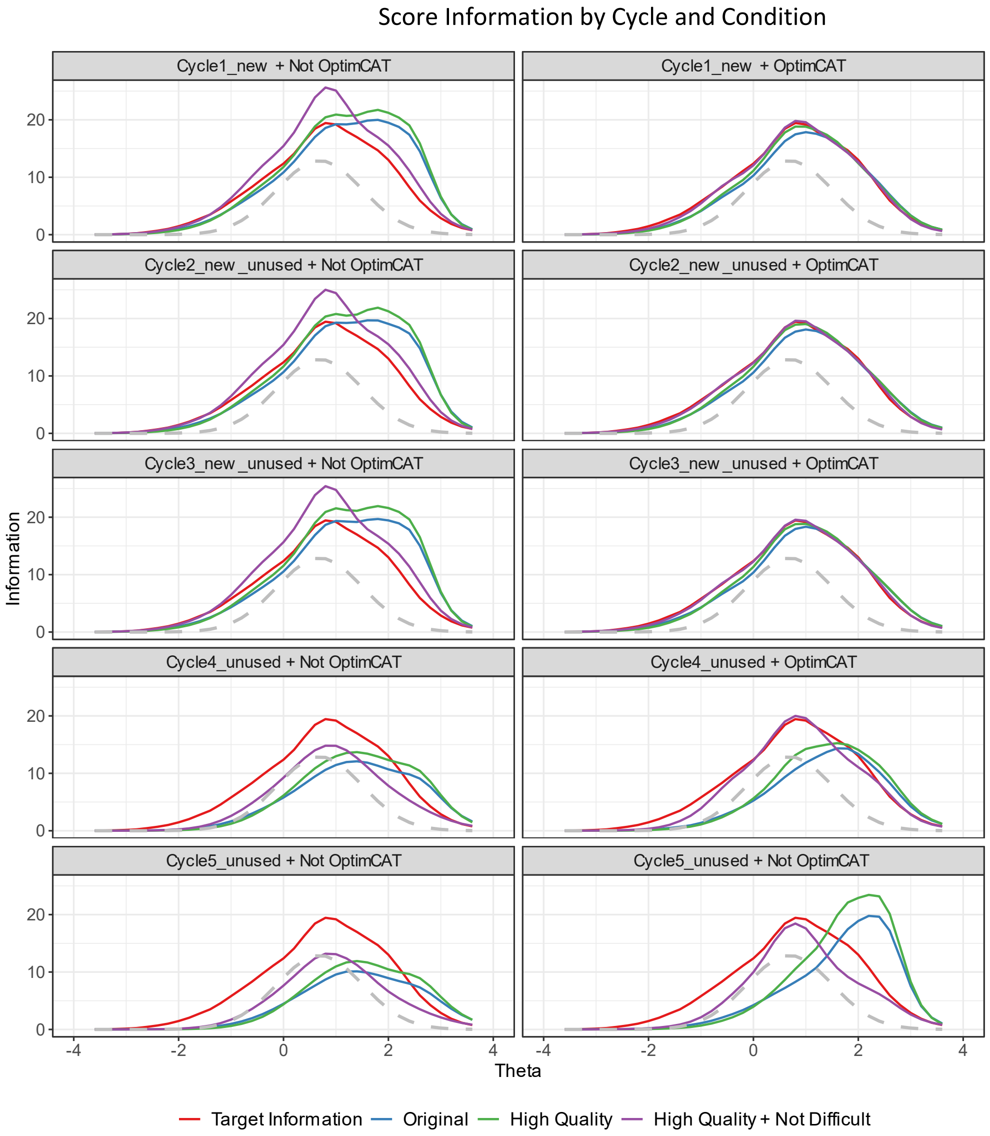
OFFICE OF PEOPLE ANALYTICS

MK: ASVAB IITMS Analysis

- We assembled five forms across five cycles under different conditions:
  - without paying attention to item quality and difficulty labels (i.e., “original”) and without CAT optimization (i.e., the current approach)
  - without paying attention to item quality and difficulty labels (i.e., “original”) and with CAT optimization
  - using only items predicted to be “high quality” (i.e., “high quality”) and without CAT optimization
  - using only items predicted to be “high quality” (i.e., “high quality”) and with CAT optimization
  - using only items predicted to be “high quality” and “not-difficult” (i.e., “high quality + not difficult”) and without CAT optimization
  - using only items predicted to be “high quality” and “not-difficult” (i.e., “high quality + not difficult”) and with CAT optimization
- We implemented the item-reuse policy across the five cycles:
  - In Cycles 2-5, we reuse items from preceding cycles
  - In Cycles 1-3, we add a supply of fresh items
  - In Cycles 4 and 5, we do not add fresh items and construct forms from reused items only

MK: ASVAB IITMS can improve form quality

- SIF results demonstrate that:
  - Curating the input pool based on quality and difficulty can lead to forms that are aligned with the latent distribution (i.e., the purple lines vs. the blue lines)
  - Assembling forms with CAT optimization can lead to more forms that are aligned with the target SIF (i.e., Cycles 1-3 with Optim-CAT)
  - Curating the input pool and assembling forms with CAT optimization can balance the competing demands of meeting target criteria and leaving the inventory sufficiently supplied to build additional forms (i.e., the purple line in Cycle 4 with Optim-CAT)
  - This combined approach, across several cycles, can align form score information with the latent examinee ability distribution and improve score information in the lower examinee ability range, when compared with the current approach

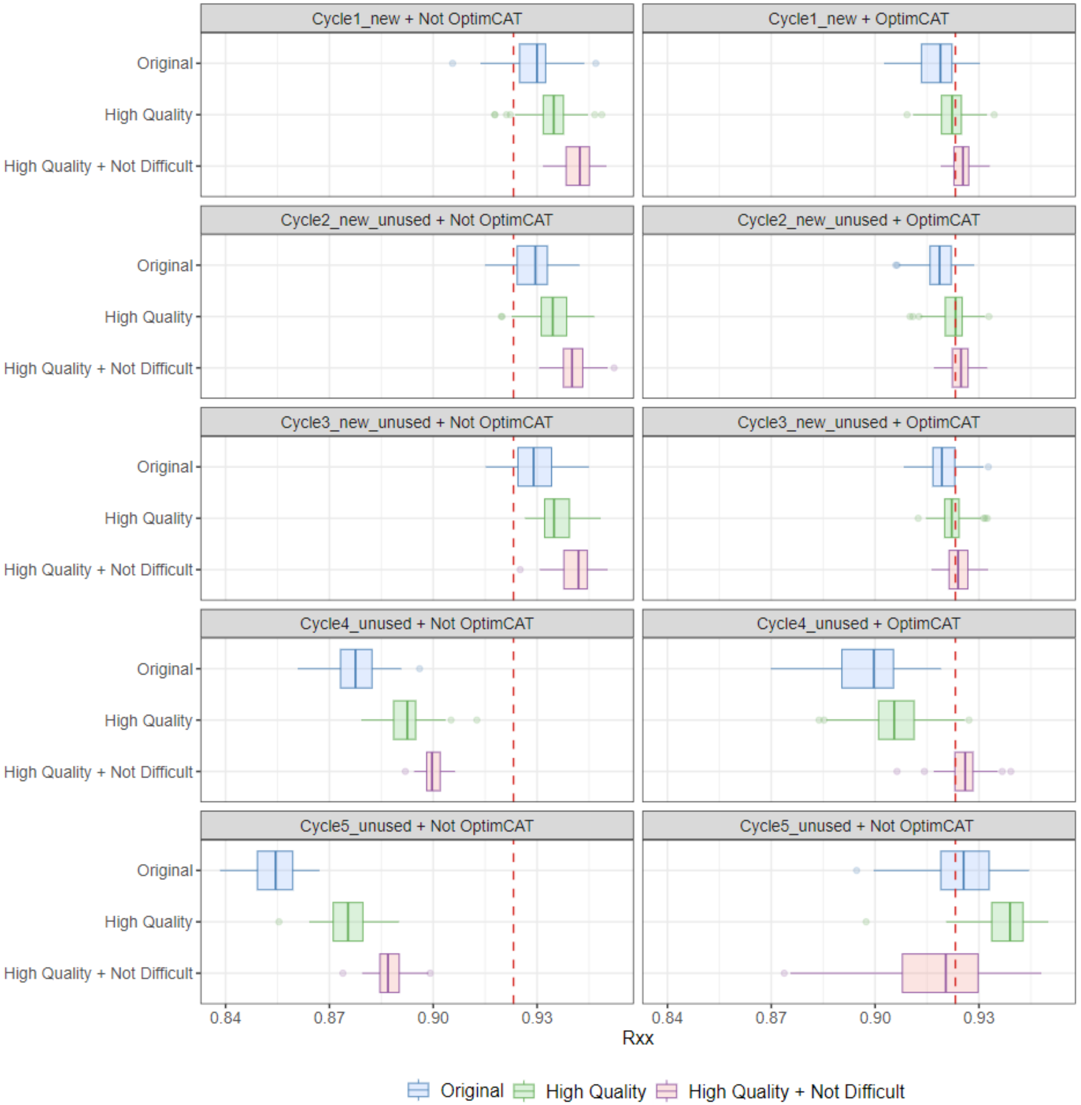


NOTE: The gray dashed line represents the latent examinee ability distribution.

MK: ASVAB IITMS can improve form quality

- Furthermore, reliability results demonstrate that:
  - Curating the input pool based on quality and difficulty can increase overall reliability across cycles (i.e., the purple vs. the blue or green box plots in Cycles 1-4)
  - Assembling forms with CAT optimization can stabilize overall reliability across cycles (i.e., the left column vs. the right column)
  - Curating the input pool and assembling forms with CAT optimization can lead to overall reliability across cycles that is more consistent with the target, when compared with the current approach (i.e., the purple box plots on the left vs. the purple box plots on the right)

Marginal Reliability by Cycle and Condition

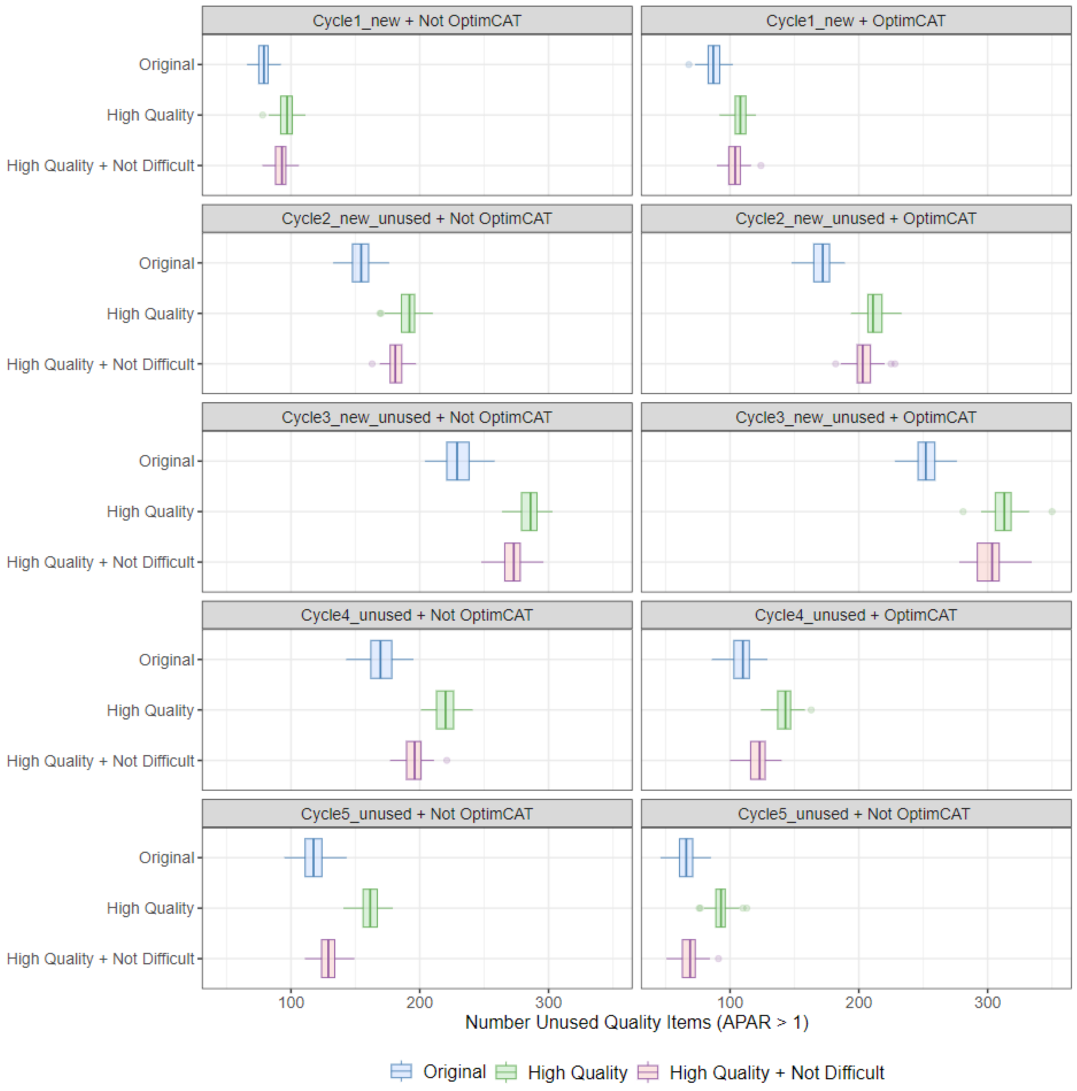


NOTE: The red dashed line represents reliability corresponding to the target score information (i.e., the red line in the plot on the previous slide).

MK: ASVAB IITMS can save quality items for future cycles

- Additionally, examination of unused items demonstrates that:
  - Assembling forms with CAT optimization can save more quality items (i.e., the boxplots on the left and the right for a single color, within Cycles 1-3)
  - Assembling forms with CAT optimization can efficiently use the curated input pool, leading to more leftover quality items for future form development cycles (i.e., the gap between the blue and the green boxplots and the gap between the blue and the purple boxplots, between the left and the right, across Cycles 1-3)

Unused Quality Items by Cycle and Condition

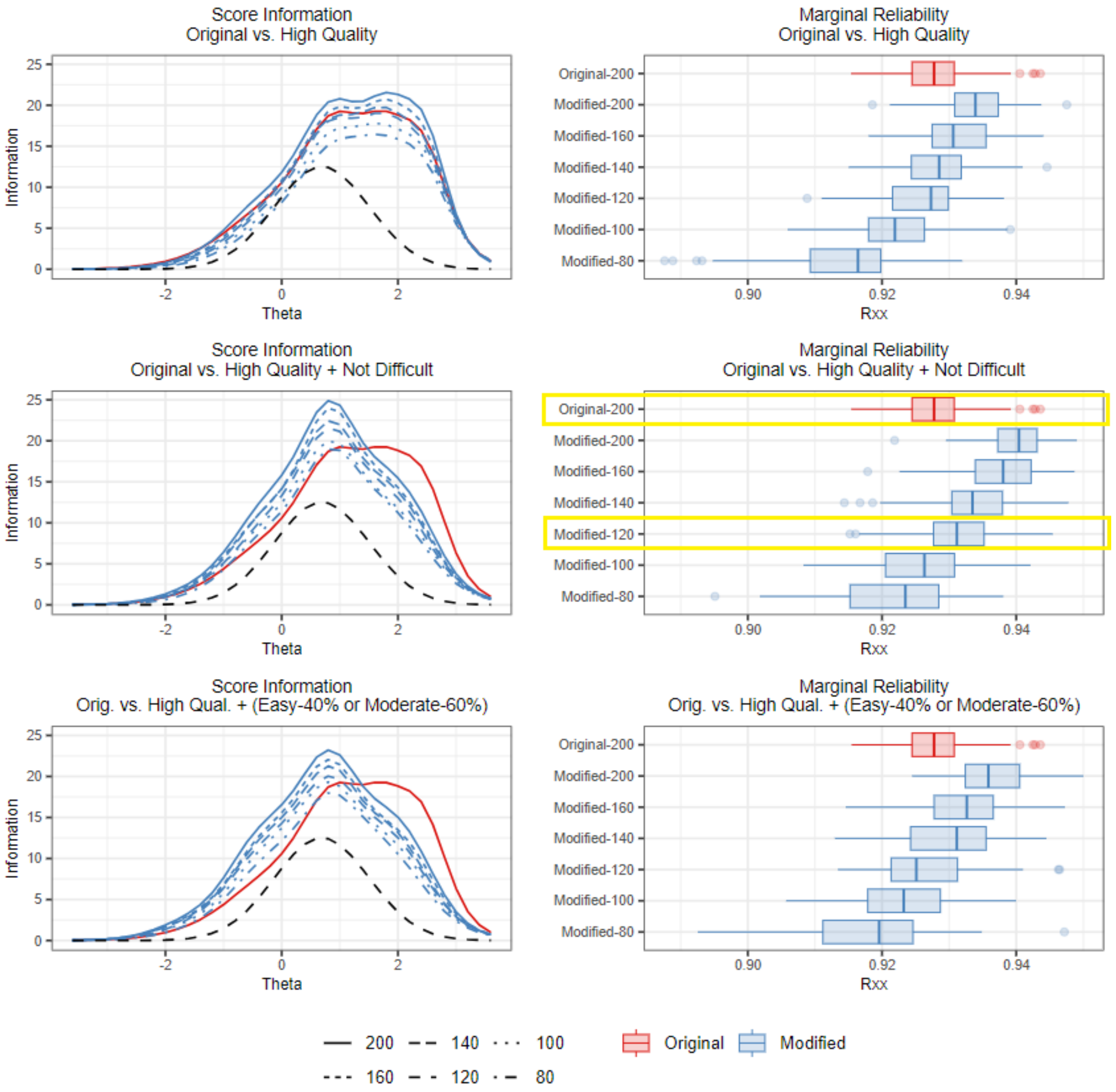


Operational Implications of ASVAB IITMS

- In sequential form development cycles, NLP model prediction and Optim-CAT:
  - Can align form score information with the latent examinee ability distribution and improve score information in the lower examinee ability range, when compared with the current approach
  - Can lead to higher and more consistent overall reliability across cycles, when compared with the current approach
  - Can save more quality items for future form development cycles
  - Can lead to an effective input pool size of 150 per form when producing four forms (i.e., 600 items for four forms versus 800 items)
- These simulation analysis results provide insight into how IITMS can inform future item development and seeding decisions
  - IITMS can create look-ahead forms after form development to forecast needs
  - IITMS can provide feedback to DTAC on which items to field test from items that have already been written
  - If there are not enough items, IITMS can provide feedback to item developers on which items to write
- We performed additional analyses to address input pool size of Optim-CAT ATA

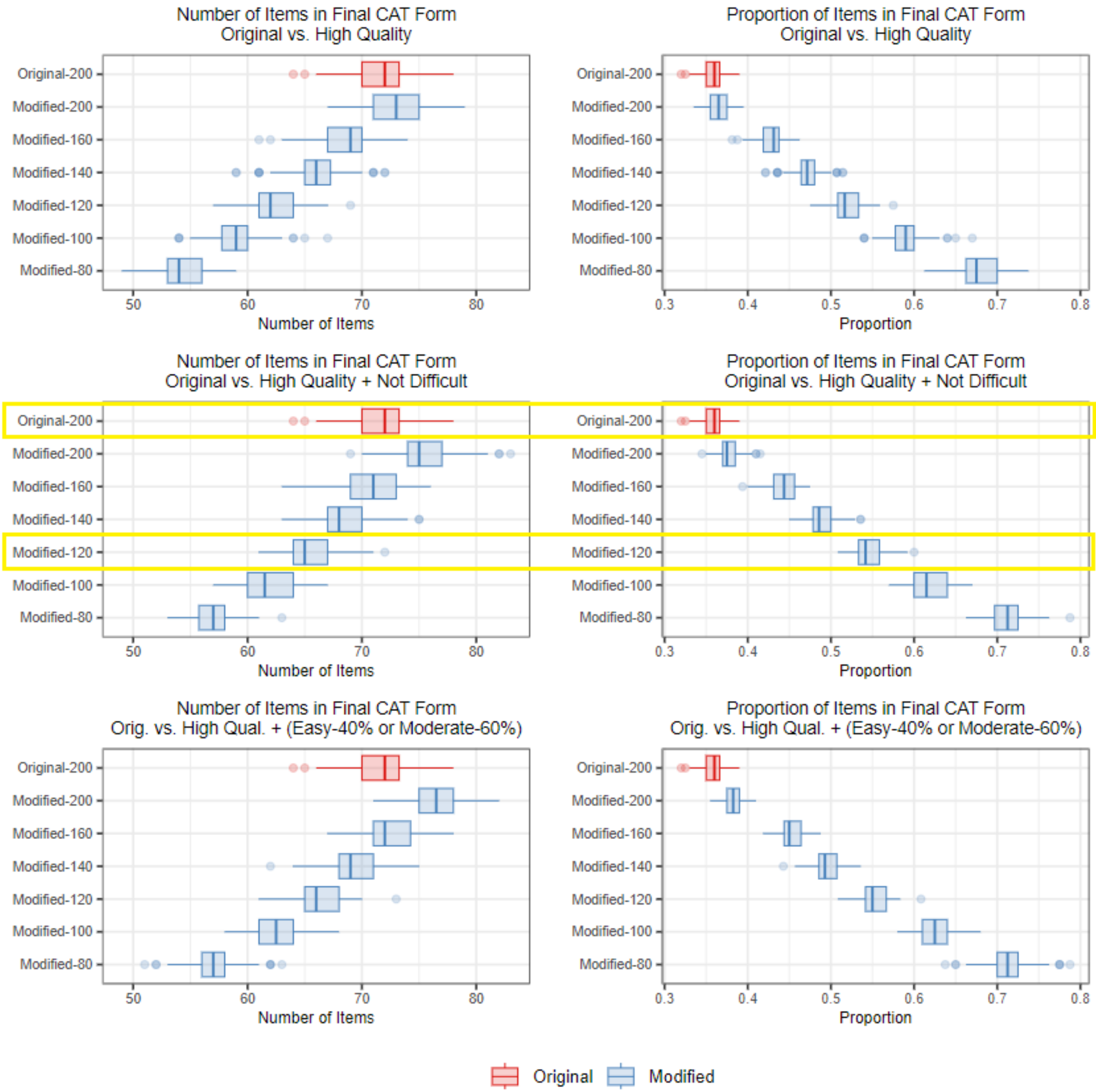
MK: A smaller, curated input pool can match or improve CAT form quality

- We formed input pools without Optim-CAT under different conditions:
  - The original input pool contained 200 items (i.e., the current approach)
  - We created modified input pools using only items predicted as "high quality" (i.e., "high quality")
  - We created modified input pools using only items predicted as "high quality" and "not-difficult" (i.e., "high quality + not difficult")
  - We created modified input pools using only items predicted as "high quality" and "not-difficult," such that 40% of the input pool contained "easy" items and 60% contained "moderate" items (i.e., "high quality + easy-40% or moderate-60%")
- SIF and reliability results demonstrate that:
  - A smaller input pool can align score information with the latent examinee ability distribution, improve score information in the lower examinee ability range, and increase reliability, when compared with the original input pool
  - The smaller input pool size can range from 160 to as low as 120 items
  - When compared with the current approach, the 120-item input pool (i.e., 40% savings) under condition "high quality + not difficult" demonstrates improved score information in the lower ability range and improved reliability



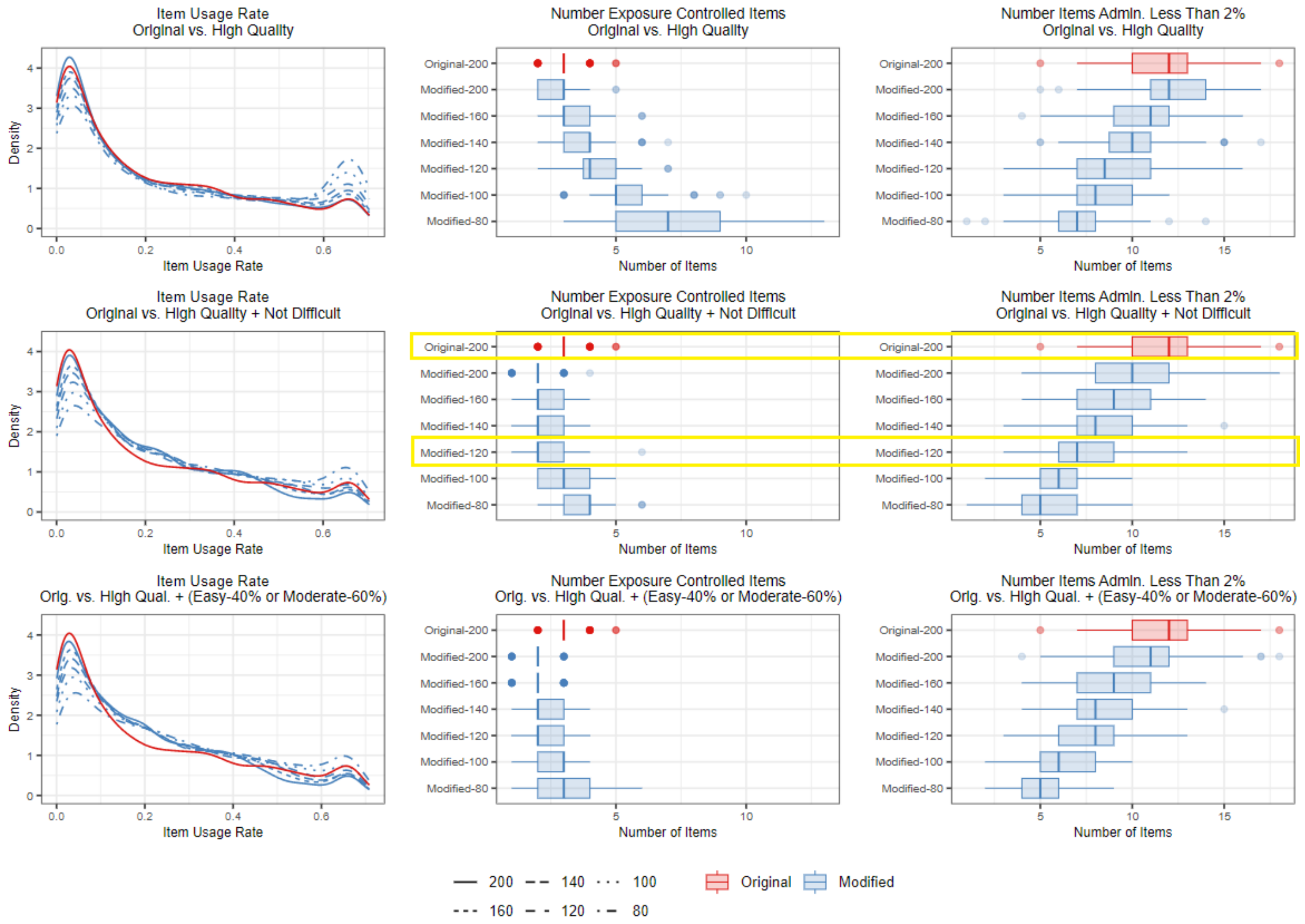
MK: A smaller, curated input pool can create a form with fewer items

- CAT form composition results demonstrate that:
  - Filtering items by item quality and difficulty almost always produces shorter CAT forms
  - The 160- to 120-item input pools under condition “high quality + not difficult” lead to CAT forms with 60 or more items
  - At the same time, the relatively smaller numbers of items in the CAT forms represent higher proportions of items in their respective input pools (i.e., higher survival rate)
    - Recall that, during development of Forms 11-15, only about 39% of seed items were assigned to a CAT form
    - However, the 120-item input pool under condition “high quality + not difficult” produced a CAT form that kept about 55% of items
    - 55% is a substantial improvement in survival rate compared with 39% (i.e., an improvement of over 40%)



MK: A smaller, curated input pool can reduce the number of underutilized items

- Item usage results demonstrate that:
  - When comparing the full distribution of item usage rates across input pools, there are noticeable differences in the lowest and highest usage rate ranges but similarities in the middle of the distribution
  - For 160- to 120-item input pools under condition “high quality + not difficult,” the number of exposure-controlled items is comparable to that of the original input pool
  - For 160- to 120-item input pools under condition “high quality + not difficult,” there are fewer items administered less than 2% of the time for the best modified input pools (i.e., about 5 to 10 items) compared to the original input pool (i.e., 10 or more items)
    - Items administered less than 2% of the time are inefficiently used or underutilized items
  - When compared with the current approach, the 120-item input pool under condition “high quality + not difficult” demonstrates more efficient item utilization



Summary

OPA
OFFICE OF PEOPLE ANALYTICS

Summary

- ASVAB IITMS would allow the stakeholders of CAT-ASVAB form development to proactively monitor item characteristics in the item inventory to support future form development
    - The NLP predictive model would ensure a steady supply of quality items that are aligned with psychometric targets (i.e., SIF and the latent distribution)
    - The optimization model would ensure items are used efficiently from resource and psychometric perspectives (i.e., forms consistently meet SIF targets)
        - This is an important improvement in upstream steps (i.e., developing and seeding items efficiently) and downstream steps (i.e., ensuring parallel forms)
    - A smaller, curated input pool can create a form with fewer items, smaller proportion of unused items, and fewer items administered infrequently (i.e., less than 2%)
    - For the MK subtest, NLP modeling and CAT simulation without Optim-CAT led to savings of 40%, from 200 to 120, in the number of items needed in the input pool and savings of 40%, from 16,000 to 9,600, in the number of examinees required during field testing, while retaining the quality of the test

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- Does the DAC have feedback on the recommended process improvements?

Thank you!

For more information please contact:

Ted Diaz and Olga Golovkina
tdiaz@humrro.org
ogolovkina@humrro.org

OPA
OFFICE OF PEOPLE ANALYTICS