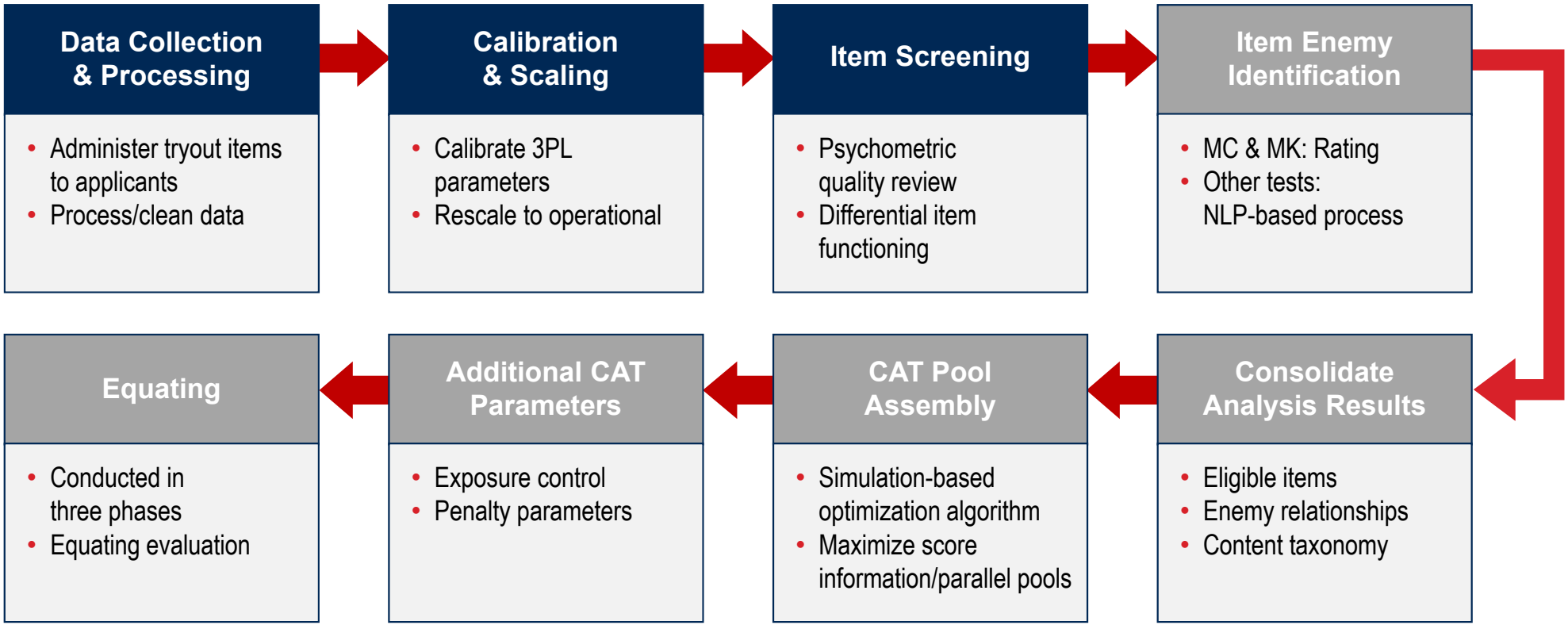
OFFICE OF PEOPLE ANALYTICS

ASVAB Item Development Process: Item Analysis

Matt Reeder
Human Resources Research Organization (HumRRO)

Briefing presented to the DACMPT
August 16, 2023

CAT-ASVAB Pool Development Process Overview



OPA
OFFICE OF PEOPLE ANALYTICS

TRYOUT ITEM DATA PROCESSING

Data Cleaning

- Remove invalid, ineligible, or corrupt records such as
  - Non-Service applicants
  - Invalid person/item identifiers
- Remove records suggesting potentially unmotivated responding such as
  - High % missing responses
  - Anomalous response latencies
- Randomly select one record from applicants who have tested more than once

Pre-Calibration Key Check

- CTT-based analysis
- Evaluate patterns of response option selection and option-total correlations
- Flag items with questionable response patterns; examples include
  - low/negative item-total correlation for key
  - positive option-total correlation for non-keyed response
- Content SMEs review items that are potentially miskeyed, have multiple correct responses, or are otherwise problematic
- Correct any miskeys and rescore as necessary
- Remove items with multiple correct responses or content flaws from IRT calibration

OPA
OFFICE OF PEOPLE ANALYTICS

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

ITEM PARAMETER CALIBRATION, CONT.

- 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
  - 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

OPA
OFFICE OF PEOPLE ANALYTICS

**Reflects current procedures than can evolve as we modify form development goals*

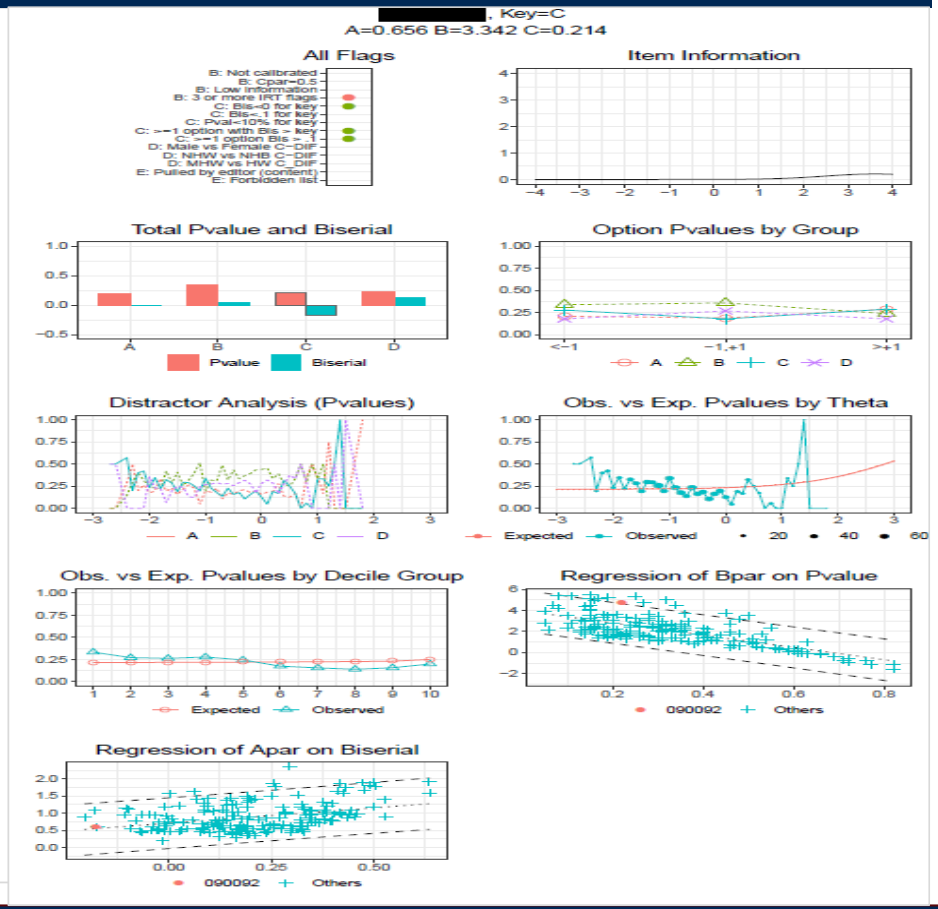
Empirical Item Screening: Psychometric Quality Review

Psychometric Quality Analyses (per item)

- Item information
- Item-model fit
  - Eight fit indices (e.g., Q1, IRT "B" parameter regressed on CTT p-value)
- Distractor analysis
  - Content review as necessary
- Differential item functioning (DIF)
  - Empirical Bayes (EB) enhancement to Mantel-Haenszel (MH) DIF analysis (Zwick, Thayer, & Lewis 1999)
- Screening Rubric
  - Many items automatically eligible for operational status (no item quality flags)
  - Some items automatically ineligible for operational status (e.g., out of bounds parameter estimate)
  - Several require psychometric/content review to determine eligibility

OPA
OFFICE OF PEOPLE ANALYTICS

"One-Pager" Visual Summary (per item)



Empirical Item Screening: Psychometric Quality Review

- Items that are not automatically eligible or ineligible for operational status require additional review
- Inherently subjective task where analysts consider all available empirical evidence of item quality + item content, as necessary
- Two analysts independently rate each item as “keep” or “drop”
- When analysts agree, item eligibility status is final
- When analysts disagree, meet to discuss rationale for ratings and establish consensus
  - Mostly resolved through discussion
  - Occasionally enlist additional rater(s) or SMEs

OPA
OFFICE OF PEOPLE ANALYTICS

Empirical Item screening: Bias Detection

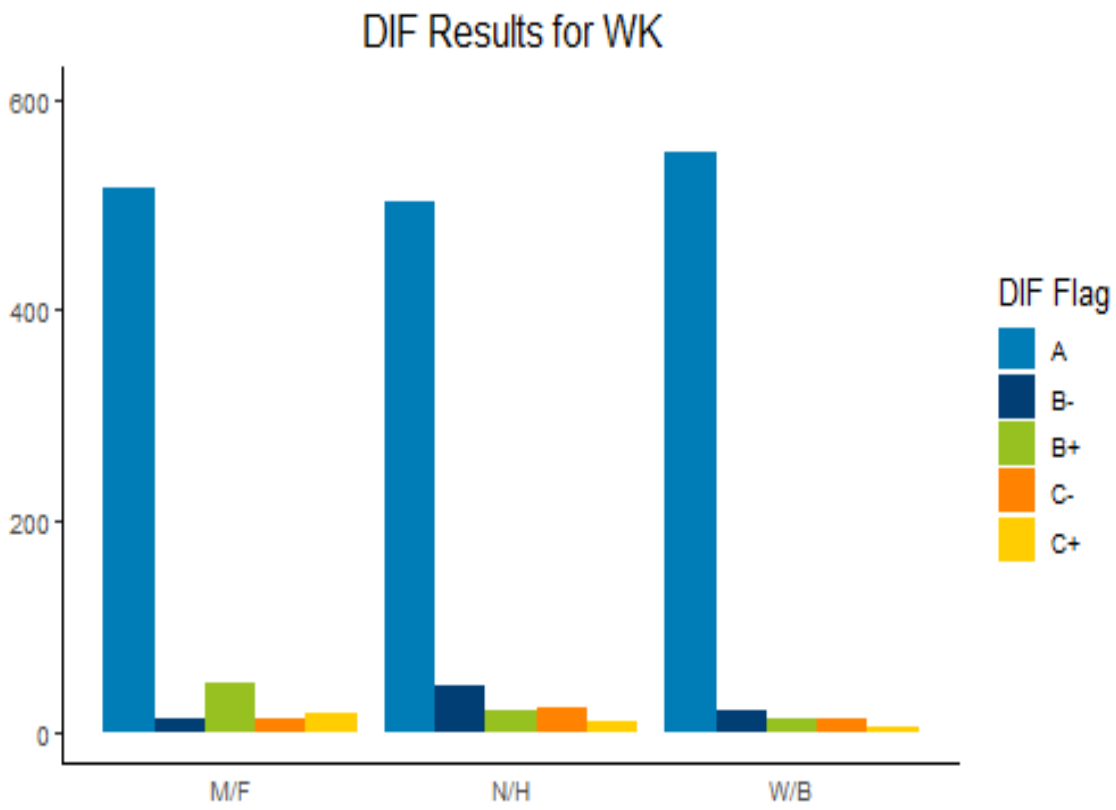
- Item-level sample sizes allow for three item performance DIF analyses
- If we had sufficient data, we would make comparisons across all relevant combinations of ethnicity/race; however, sample sizes are insufficient to make any other comparisons
- Items categorized as "C" or "moderate to severe" DIF according to ETS framework are reviewed for evidence of bias
- Review sessions include members of both focal and reference groups

Pair	Reference Group	Focal Group
1	Non-Hispanic White	Hispanic White
2	Non-Hispanic White	Non-Hispanic Black
3	Male	Female

- Reviewers are trained on basic concepts of DIF and construct irrelevant factors
- Reviewers are provided with several examples of items that include construct irrelevant content
- If reviewers conclude an item includes construct irrelevant factors that might plausibly prevent members of a group of test takers from responding to the item in ways that allow appropriate inferences about their knowledge, skills, or abilities, the item is <u>not eligible</u> to be an operational ASVAB item
- If reviewers conclude an item does <u>not</u> include construct irrelevant factors <u>and</u> the item passes all other psychometric quality screens, the item remains eligible for assignment to an ASVAB form/pool

OPA
OFFICE OF PEOPLE ANALYTICS

Example Empirical Bayes MH Results



- Histogram and table summarize analysis of 600 of 1,000 tryout items (per test), evaluated as part of CAT-ASVAB forms 11–15 assembly
- Each item is part of three

DIF Results for WK					
Comparison	A	B-	C-	B+	C+
M/F	516	11	11	45	17
N/H	504	44	22	20	10
W/B	550	20	13	13	4
Total	1,570	75	46	78	31

OPA
OFFICE OF PEOPLE ANALYTICS

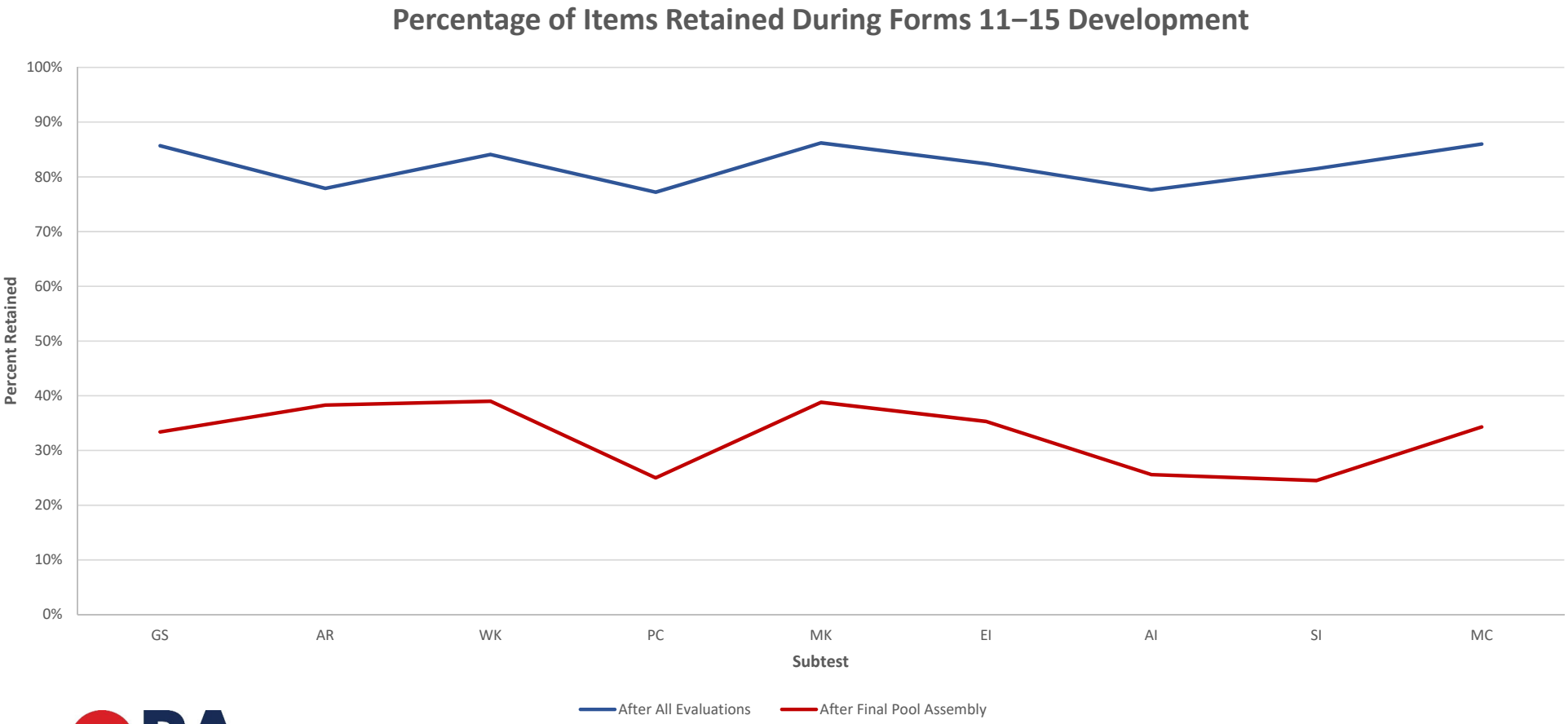
Empirical Item Screening: Bias Detection

ASVAB Test	Moderate to Severe EB (C+-) Indicators*	Items Dropped for Bias
Automotive Information	13	0
Arithmetic Reasoning	9	0
Electronics Information	8	0
General Science	20	0
Mechanical Comprehension	9	1
Math Knowledge	17	0
Paragraph Comprehension	0	0
Shop Information	50	1
Word Knowledge	77	10

** Note: Numbers do not necessarily indicate number of items, as each item is part of three sub-group comparisons of 600 items (per test) and can generate more than one indicator per item*

OPA
OFFICE OF PEOPLE ANALYTICS

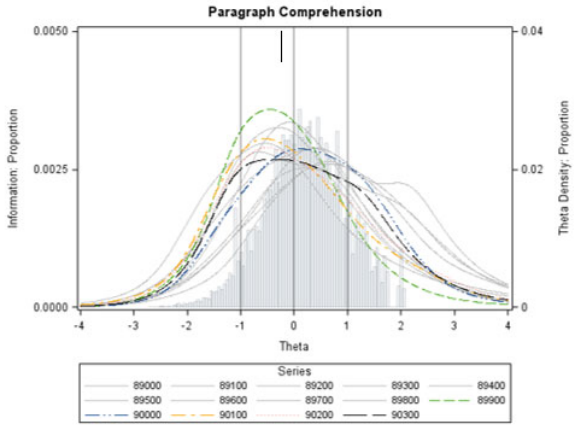
EMPIRICAL ITEM SCREENING: SUMMARY

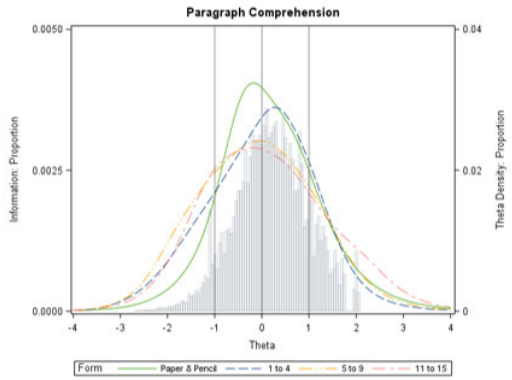


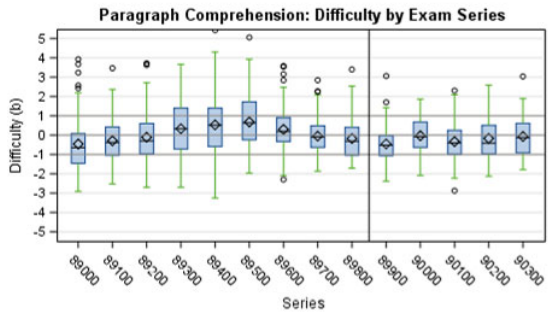
OPA
OFFICE OF PEOPLE ANALYTICS

Feedback to Content Development Team: Current Example from PC

Subtest: Paragraph Comprehension







Median difficulties for the Paragraph Comprehension subtest items are lower than the other subtests. Three of the series have median difficulties of less than 0, with the other two series falling at 0. Thus, the items are relatively well-targeted. Further, the information functions look good in relation to the actual examinee ability distribution, indicating that examinees' paragraph comprehension ability is being fairly well estimated.

OPA
OFFICE OF PEOPLE ANALYTICS

Dimensionality

OPA
OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB DIMENSIONALITY

- Each of the ten CAT-ASVAB subtests is calibrated separately and scored using a unidimensional IRT model
- The CAT-ASVAB algorithm implements content balancing for two subtests (AO & GS) based on the outcome of dimensionality analyses described in Segall, Moreno, & Hetter (1997)*
- IRT models are robust against minor violations of the unidimensionality assumption (Dorans & Kingston, 1985; Drasgow & Parsons, 1983; Reckase, 1979)
- Original developers of CAT-ASVAB considered three approaches to dealing with dimensionality (see next slide)
- Current practices continue to rely on the outcome of the original dimensionality evaluation
  - Constructs/blueprints remain unchanged
  - Item development practices remain unchanged

OPA
OFFICE OF PEOPLE ANALYTICS

*Included as read-ahead material

CAT-ASVAB DIMENSIONALITY

Approach	Calibration	Item Selection	Scoring
Unidimensional Treatment	Combined calibration containing items of each content type	No constraints placed on item content for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
Content Balancing	Combined calibration containing items of each content type	Constraints placed on the number of items drawn from each content area for each examinee	A single IRT ability estimate computed across items of different content using the unidimensional scoring algorithm
Pool Splitting	Separate calibrations of items of each content	Separate adaptively tailored tests for each content area	Separate IRT ability estimates for each content area

OPA
OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB Dimensionality

- Segall, Moreno, & Hetter (1997) present a decision framework for dimensionality that considers several criteria
  - Statistical significance of multidimensionality
  - Interpretability of factor solutions
  - Overlap of item difficulties
    - Academic vs. nonacademic content
  - Factor correlations
- Full information item factor analysis was used to evaluate dimensionality empirically via TESTFACT (Muraki, 1984)

CAT-ASVAB Dimensionality Decision Framework

Case	Factor Statistical Significance	Interpretable Factors	Overlapping Item Difficulties	Factor Correlations	Approach	Finding/Result
1	No	NA	NA	NA	Unidimensional	MC, PC
2	Yes	Yes	Yes	High	Content balance	GS, [AO]*
3	Yes	Yes	Yes	Low	Split pool	--
4	Yes	Yes	No	NA	Unidimensional	AI, AR, EI, SI, WK
5	Yes	No	Yes	NA	Unidimensional	MK
6	Yes	No	No	NA	Unidimensional	--

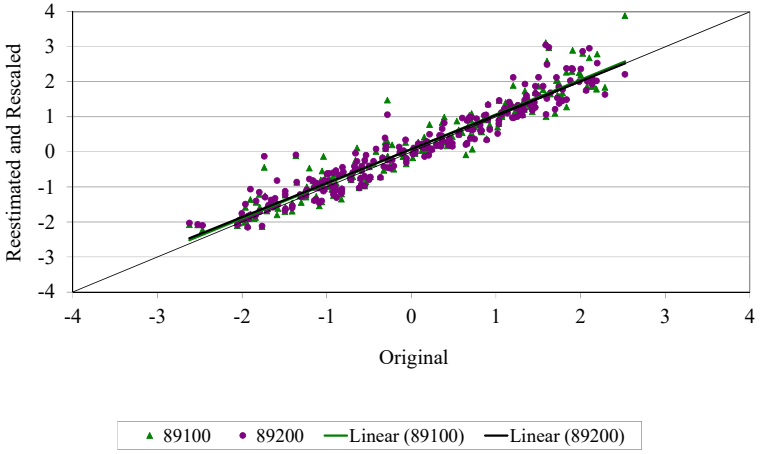
OPA
OFFICE OF PEOPLE ANALYTICS

**AO was not part of analyses documented in Segall et al. (1997), but content balancing is implemented based on similar rationale and findings.*

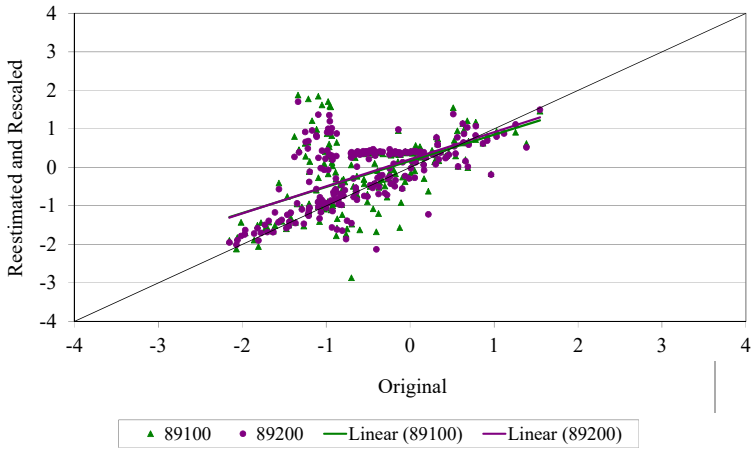
CAT-ASVAB DIMENSIONALITY

- A previous CAT-ASVAB seeding design involved recalibrating operational items along with tryout items
- Recovery of operational difficulty parameter values is indirect evidence of unidimensionality
- Difficulty parameter values recovered as expected for all tests except AO when this seeding design was operational
- Failure to recover operational difficulty parameter values was important in diagnosing multidimensionality in a generation of AO tryout items

Original vs. Reestimated and Rescaled Operational B Parameters, MC



Original vs. Reestimated and Rescaled Operational B Parameters, AO



OPA
OFFICE OF PEOPLE ANALYTICS

Recent Investigations into ASVAB Dimensionality

- DTAC-sponsored research
  - Sparse data dimensionality assessment with the Cyber Test
  - Feasibility of combining AR and MK ASVAB subtests
    - Finding of essential unidimensionality
- Dimensionality analyses rooted in two frameworks
  - IRT-based analyses (e.g., correlations between theta estimates, item misfit)
  - Item factor analysis approaches rooted in bifactor modeling
    - Explained common variance (ECV)
    - Comparison of general (g) factor loadings from one-factor vs. bifactor models
- These and other approaches could be explored in the future for general use in ASVAB item and form development

OPA
OFFICE OF PEOPLE ANALYTICS

CAT-ASVAB DIMENSIONALITY

- Evaluating dimensionality of tryout items is not currently part of the pool development process
- The primary reason is that each examinee is administered 15 of either 200, 400, or 800 tryout items, depending on the test, resulting in a very sparse response data matrix
  - Covariance-based approaches will not work
  - Full information maximum likelihood (FIML) is theoretically appropriate but practically challenging given the degree of missing data (92–98%) at the examinee level
  - IRT model-based approach iFACT: (Segall, 2002) may be an option
    - Has been applied to a somewhat similar data structure in a related context (Gao, 2018)
- Again, we rely on the foundational research on dimensionality + the fact that constructs, blueprints, and development procedures remain constant

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

QUESTIONS FOR THE DAC

- Does the DAC have a recommendation on a process for evaluating dimensionality of ASVAB tryout items under sparse data conditions?
- Does the DAC have feedback on item analysis processes?

Thank you!

For more information, please contact:

Jeff Dahlke
jdahlke@humrro.org

OPA
OFFICE OF PEOPLE ANALYTICS