SLIDES ONLY
NO SCRIPT PROVIDED

OFFICE OF PEOPLE ANALYTICS

Update on Calculator Studies

Andrea Sinclair & Jeff Dahlke
Human Resources Research Organization

Briefing presented to the DACMPT
June 12, 2024

24-P-0583

Briefing Agenda

- Background Information
- Three Calculator Studies:
  - Study 1: Content expert review of Arithmetic Reasoning (AR) and Mathematics Knowledge (MK) items for calculator sensitivity
  - Study 2: Empirical investigation of impact of calculators on ASVAB scores
  - Study 3: Needs assessment to determine what the blueprint would contain in the event a new calculator test is needed
- Next Steps
- Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Background Information

OPA
OFFICE OF PEOPLE ANALYTICS

Overview

- Current ASVAB policy is “no calculators”
- Previous research (Buckland et al., 2021) surveyed subject matter experts (SMEs) across the Services about whether servicemembers are required to apply mathematics knowledge and arithmetic reasoning *without having access* to a calculator or other tool
  - 68% of surveyed military SMEs indicated some form of math, without a calculator, is required in training
  - 56% reported that some form of math, without a calculator, is required on the job
  - Thus, Buckland et al. (2021) recommended the “no calculator” policy continue

Overview (cont.)

- Expressed concerns over current policy with respect to calculators
  - Other national testing programs (e.g., ACT, SAT, GED) allow calculators on the quantitative tests
  - Exclusion of calculators may result in the perception that the ASVAB testing program is not keeping up with trends in assessment
  - High school curricula often allow calculators during instruction and exams
  - Test items requiring manual calculations may result in increased test anxiety as students are not accustomed to performing such calculations without a calculator

Study 1:
Content Expert Review of Arithmetic Reasoning (AR) and Mathematics Knowledge (MK) Items for Calculator Sensitivity

OPA
OFFICE OF PEOPLE ANALYTICS

Study 1: Procedures

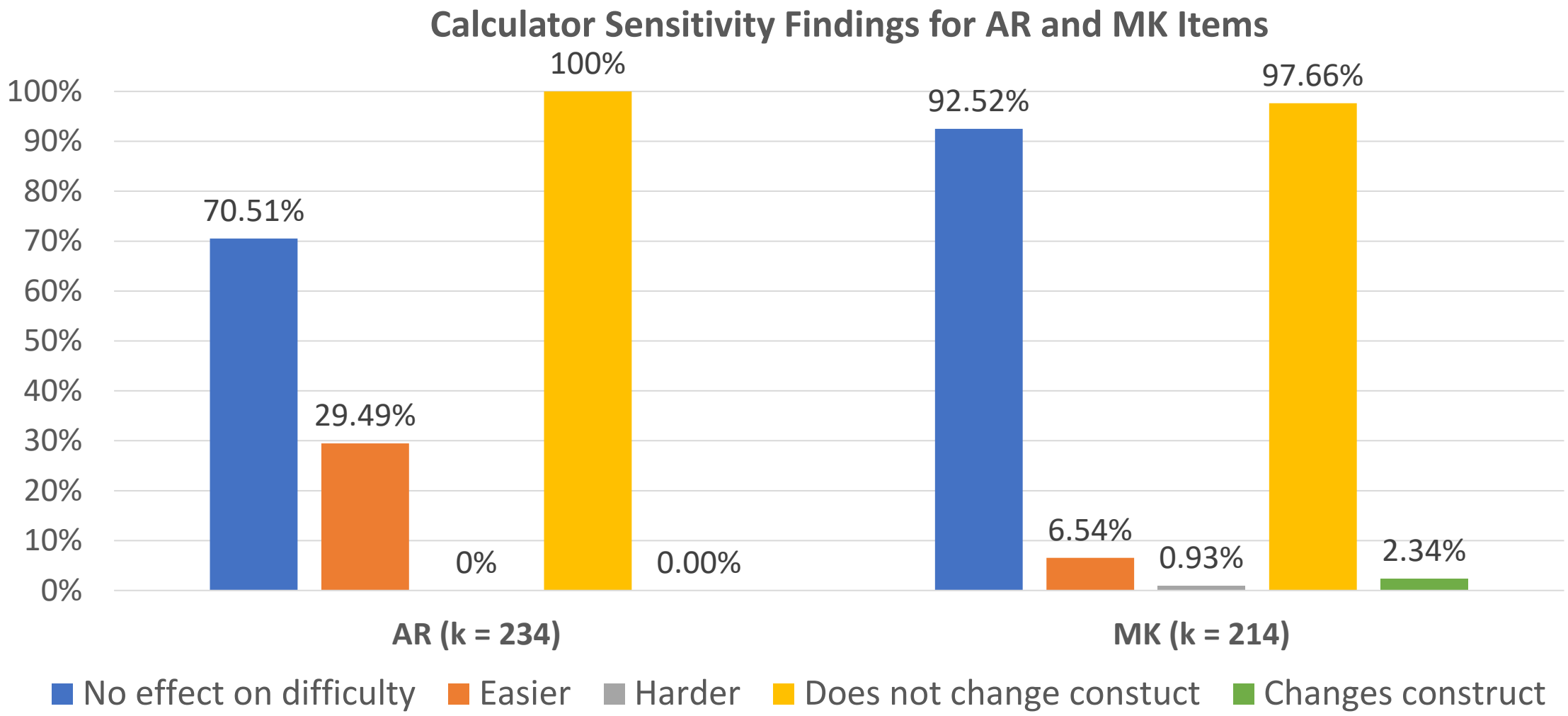
- Two math content experts reviewed 393 CAT-ASVAB items and 55 P&P items for calculator sensitivity (448 in total)
  - 234 AR items
  - 214 MK items
- Calculator Sensitivity operationalized as use of a calculator impacting . . .
  - difficulty of the item (makes item easier or harder)
  - underlying construct measured by the item

OPA
OFFICE OF PEOPLE ANALYTICS

Study 1: Procedures (cont.)

- Experts instructed to read the definition of the construct, review the taxonomy/blueprint, read each item and response options, and then provide three ratings for each item:
  - **To which content area is the item most closely aligned?**
    - Used the content areas associated with each subtest's blueprint
  - **Does the calculator change the difficulty of the item?**
    - Instructed that difficulty is distinct from response time
  - **Does the calculator fundamentally change the skill (construct) being measured?**
    - That is, if the calculator is used, could the item be correctly answered without understanding the underlying mathematical principle(s), thereby changing the nature of what is being assessed from math knowledge to calculator knowledge?
- Experts instructed to consider, "the full range of high school students in Grades 10–12 who typically take the ASVAB" when making their ratings

Study 1: Results



OPA
OFFICE OF PEOPLE ANALYTICS

Study 1: Summary & Conclusions Stemming from Content Expert Review Study*

AR Summary

- Nearly 1/3 of AR items rated as easier with the calculator
  - Content area most impacted was Interest and Percentage
- Calculator did not make any items more difficult
- Calculator did not change the construct for any items

MK Summary

- Less than 1/10 of MK items rated as easier with the calculator
  - Items rated as easier in all content areas except Number Theory
- A couple items rated as more difficult
  - Isolated to items with multiple sets of parentheses/grouping symbols
- A few items rated as changes the construct; isolated to . . .
  - Identify square root (e.g., $\sqrt{36} = ?$)
  - Solve problem with negative numbers (e.g., −30 – 7 = ?)

Conclusions

- Calculator is likely to make test easier, especially for AR subtest
  - Logical finding given what each subtest is designed to measure
- If use of a calculator is allowed, need to consider implications of . . .
  - replacing items that can be correctly answered by recognizing the correct calculator button with calculator neutral items
  - replacing items with calculator neutral items ***or*** conducting an equating study when item difficulty is impacted by use of a calculator

OPA
OFFICE OF PEOPLE ANALYTICS

Aside from main effect findings, many of these results appear inconsistent with empirically-based results for the subset of items included in Study 2.

Study 2:
Empirical Investigation of Impact of Calculators on ASVAB Scores

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Empirical Investigation of Impact of Calculators on ASVAB Scores

- Purpose:
  - Empirically evaluate the impact on examinee test performance and the psychometric properties of the AR and MK subtests when calculators are allowed on the MK and AR subtests of the ASVAB.
- Study design considerations:
  - Maximize generalizability to ASVAB applicant population
  - Minimize security risks to existing ASVAB item pools
  - Minimize disruptions to operational testing of applicants
  - Minimize strain or burden on study participants

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Study Sample

- Individuals similar to those who take the ASVAB under operational testing conditions, with (relatively) recent operational ASVAB scores
- Shippers complete the study during a waiting period on their ship day
- Target sample size = 3,600 (1,800 per condition)
  - Expected to lose about 20% of cases (e.g., low motivation, unmerged data)

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Experimental Conditions

- Two conditions: calculator provided/calculator not provided
  - Same calculator as used in Study 1
- To avoid intermingling or “cross-condition” exposure, all participants on a given day assigned to the same condition
  - Odd days (11th, 19th, 25th of month) = calculator not provided
  - Even days (12th, 20th, 30th of month) = calculator provided

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Test Delivery

- Designed to be as similar as possible to ASVAB operational testing
- Administered in MEPS by Test Administrators/Test Control Officers
- Hosted on HumRRO's online assessment delivery platform
- Follows functionality, look, and feel of CAT-ASVAB
  - P&P items (also included in Study 1), administered as fixed, linear form
- Included post-test survey (contextual information about participants, motivation, calculator usage)
- Pilot testing mid-Dec. 2023–early Jan. 2024
- Operational data collection Jan. 16–Mar. 29, 2024

Study 2: Data Management

- MEPS sent rosters of participants containing user IDs and SSNs to DTAC
- DTAC merged rosters of participants with operational ASVAB records, stripped SSNs, and sent to HumRRO
- HumRRO merged with study data collected on the HumRRO platform (merged by user ID)
- Cases screened for self-reported low motivation and insufficient effort (excluded from analyses)

Study 2: Demographic Characteristics of Analysis Sample*

Demographic	No Calculator		Calculator		Total		FY 2023 Applicants/Accessions	
	n	%	n	%	n	%	n	%
Female	139	9.0	156	10.0	295	10.0	80,986	25.1
Male	1,118	76.0	1,175	76.0	2,293	76.0	237,604	73.5
Data Not Available	210	14.0	207	13.0	417	14.0	4,653	1.4
Hispanic White	262	18.0	329	21.0	591	20.0	80,348	24.9
Non-Hispanic Asian	50	3.0	63	4.0	113	4.0	17,406	5.4
Non-Hispanic Black	286	19.0	282	18.0	568	19.0	87,395	27.0
Non-Hispanic White	567	39.0	568	37.0	1,135	38.0	113,921	35.2
Other†	55	4.0	57	4.0	112	4.0	16,317	5.1
Data Not Available	247	17.0	239	16.0	486	16.0	7,856	2.4

*Shippers from 59 of 65 MEPS participated in the study.

†Participants who provided ethnicity information and identified as American Indian, Alaska Native, Native Hawaiian, or Other Pacific Islander, and/or identified as Hispanic Black or Hispanic Asian

Study 2: Analysis Sample

Service	No Calculator		Calculator		Total		FY 23 Applicants/Accessions	
	n	%	n	%	n	%	n	%
Army	467	32.0	486	32.0	953	32.0	165,358	51.2
Air Force	287	20.0	247	16.0	534	18.0	56,736	17.6
Marine Corps	198	13.0	271	18.0	469	16.0	46,935	14.5
Navy	255	17.0	304	20.0	559	19.0	46,199	14.3
Coast Guard	43	3.0	29	2.0	72	2.0	6,679	2.1
Space Force*	13	1.0	0	0.0	13	0.0		
Invalid/Missing	204	14.0	201	13.0	405	13.0	1,336	0.4
Total	1,467	100.0	1,538	100.0	3,005	100.0	323,243	100.00

*Due to Space Force service code not yet being consistently implemented in data system, Space Force applicants are included with Air Force.

Study 2: Scoring

- Applied the study items' CAT-ASVAB 3PL IRT parameters to participants' item-level scores to compare AR and MK scores from this study with operational scores
  - Matches what is done to compute theta estimates on the CAT-ASVAB

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Planned Analyses*

- *Does calculator usage meaningfully impact the dimensionality of AR and MK subtests (RQ1)?*
  - Factor analyses to evaluate the extent to which factor structures of the subtests differ across study conditions
  - Comparisons with Study 1 findings

*Types of analyses planned for RQ1; analyses not yet conducted as of April 2024.

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Preliminary Results*

- *Do psychometric properties differ between study conditions (RQ2)?*
  - Comparison of mean scores across conditions:

Subtest	Official Scores					Experimental Scores					Estimated Latent Ability Distributions				
	No Calculator Condition		Calculator Condition		*d*	No Calculator Condition		Calculator Condition		*d*	No Calculator Condition		Calculator Condition		*d*
	M	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*		*M*	*SD*	*M*	*SD*	
AR	.49	.85	.43	.86	−.07	.06	.79	.35	.79	.37	.03	.89	.39	.90	.40
MK	.72	.74	.69	.72	−.05	.30	.70	.35	.70	.07	.30	.75	.35	.76	.07

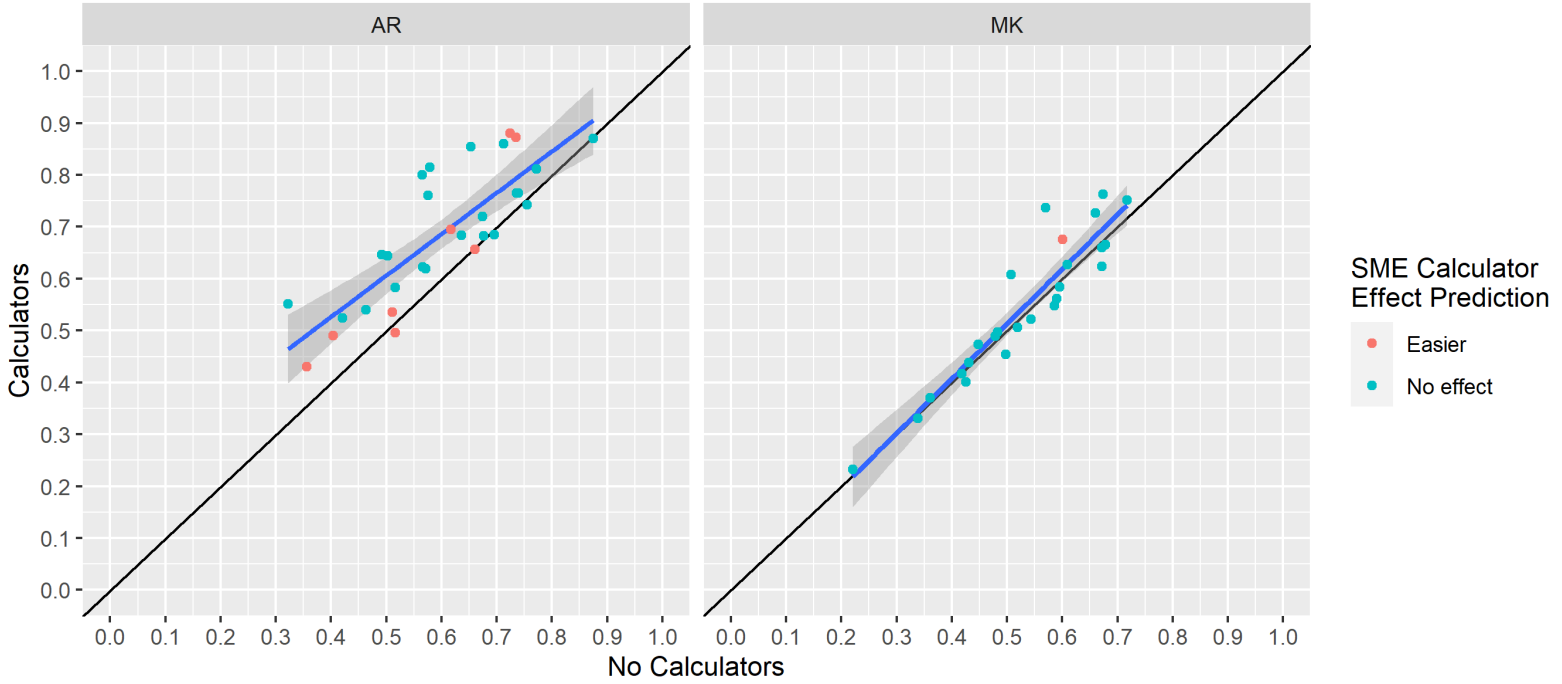
*Results are preliminary as of April 2024.

OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Preliminary Results (cont.)

- *Do psychometric properties differ between study conditions (RQ2)?*
  - Classical test theory *p* value comparisons across conditions:

Subtest	*p* Values				
	No Calculator Condition		Calculator Condition		*d*
	M	*SD*	*M*	*SD*	
AR	.60	.13	.69	.13	.65
MK	.53	.12	.55	.14	.11



OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Preliminary Results (cont.)

- *Do psychometric properties differ between study conditions (RQ2)?*
  - IRT item parameter comparisons between conditions:

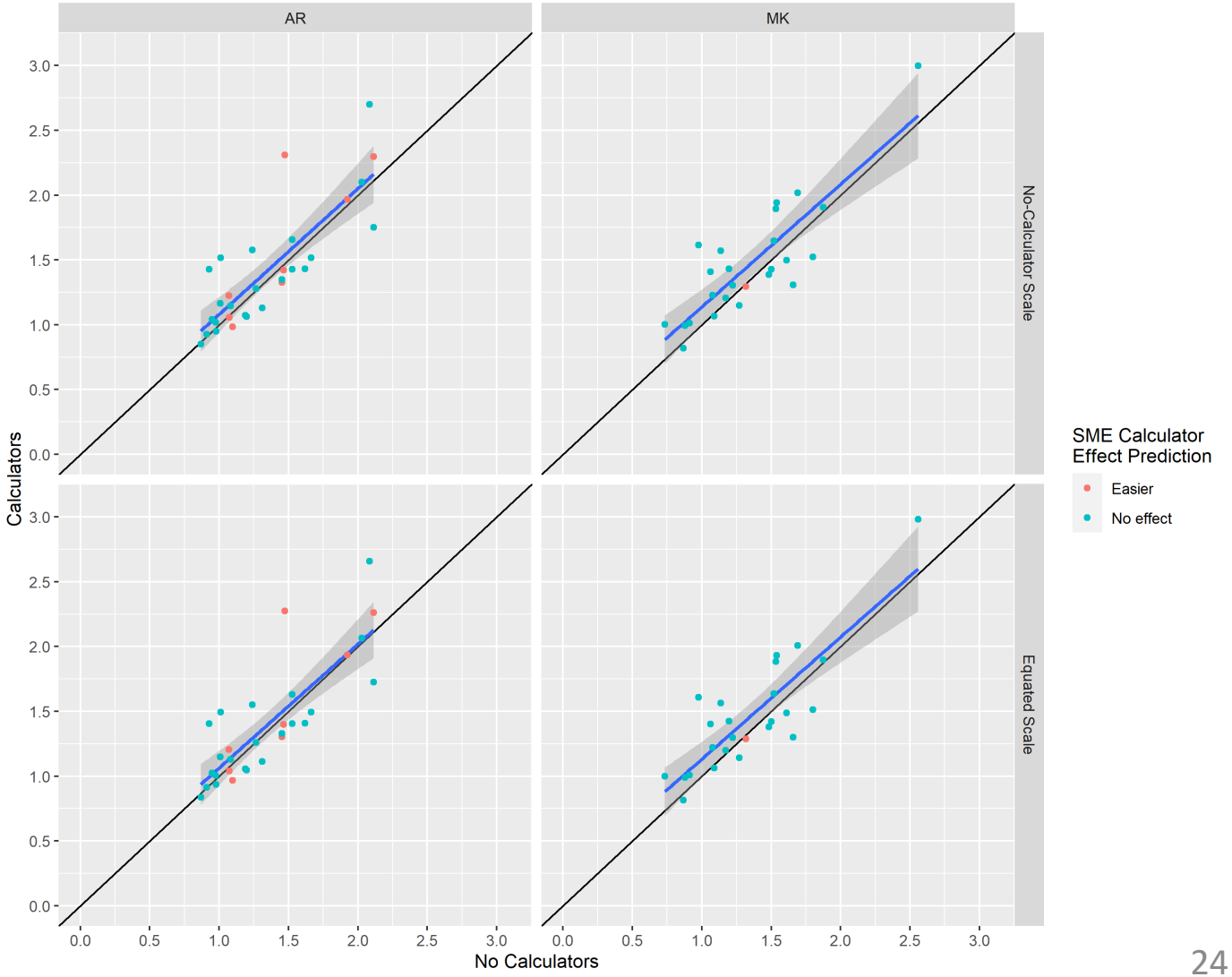
3PL Item Parameter	Subtest	No-Calculator Scale				
		No Calculator Condition		Calculator Condition		*d*
		M	*SD*	*M*	*SD*	
a	AR	1.35	0.39	1.42	0.46	0.16
	MK	1.35	0.40	1.47	0.45	0.28
b	AR	0.02	0.47	-0.31	0.53	-0.67
	MK	0.56	0.32	0.53	0.39	-0.09
c	AR	0.22	0.07	0.23	0.06	0.14
	MK	0.22	0.08	0.23	0.10	0.13

3PL Item Parameter	Subtest	Equated Scale				
		No Calculator Condition		Calculator Condition		*d*
		M	*SD*	*M*	*SD*	
a	AR	1.35	0.39	1.40	0.45	0.11
	MK	1.35	0.40	1.46	0.45	0.26
b	AR	0.02	0.47	0.04	0.54	0.03
	MK	0.56	0.32	0.59	0.39	0.07
c*	AR	0.22	0.07	0.23	0.06	0.14
	MK	0.22	0.08	0.23	0.10	0.13

**c* parameter is not transformed

Study 2: Preliminary Results (cont.)

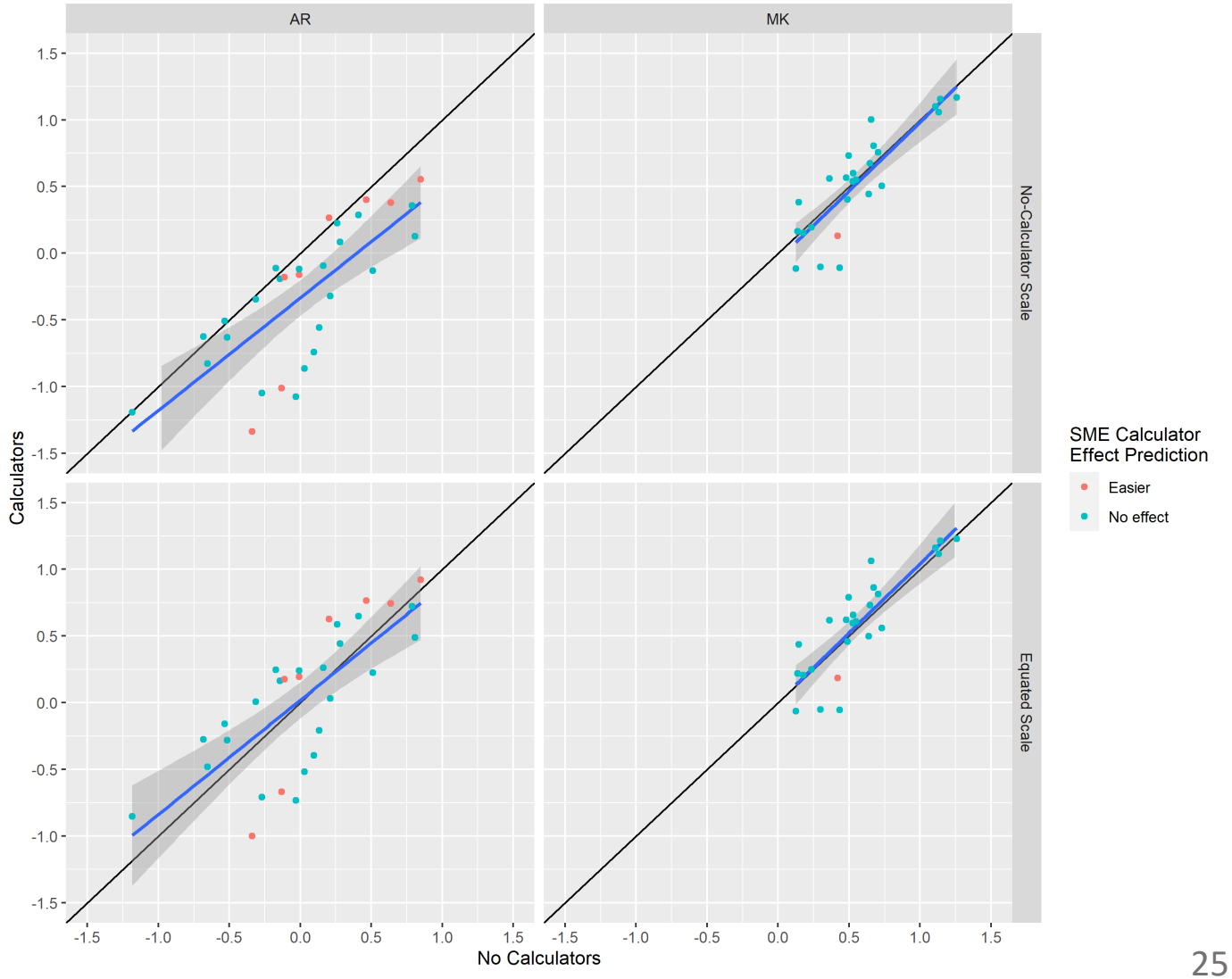
- *Do psychometric properties differ between study conditions (RQ2)?*
  - IRT *a* parameter comparisons between conditions:



OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Preliminary Results (cont.)

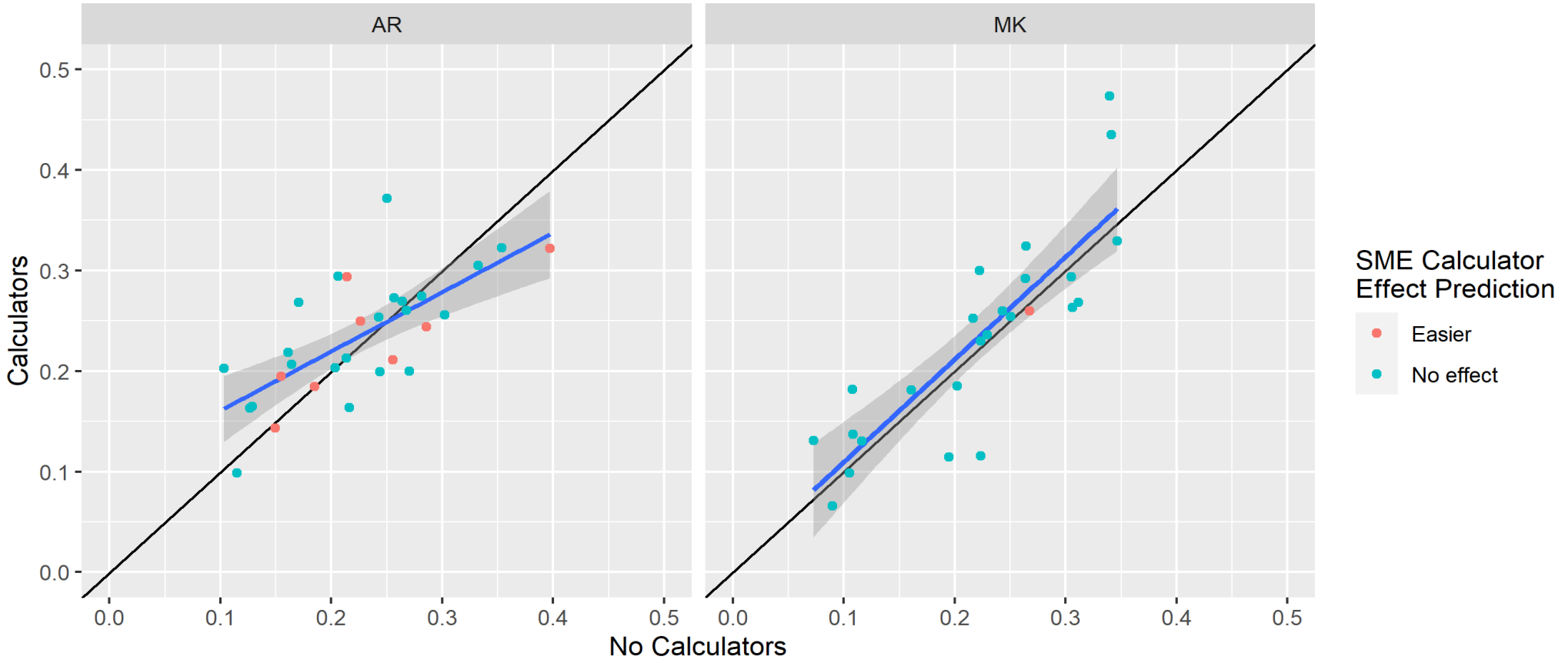
- *Do psychometric properties differ between study conditions (RQ2)?*
  - IRT *b* parameter comparisons between conditions:



OPA
OFFICE OF PEOPLE ANALYTICS

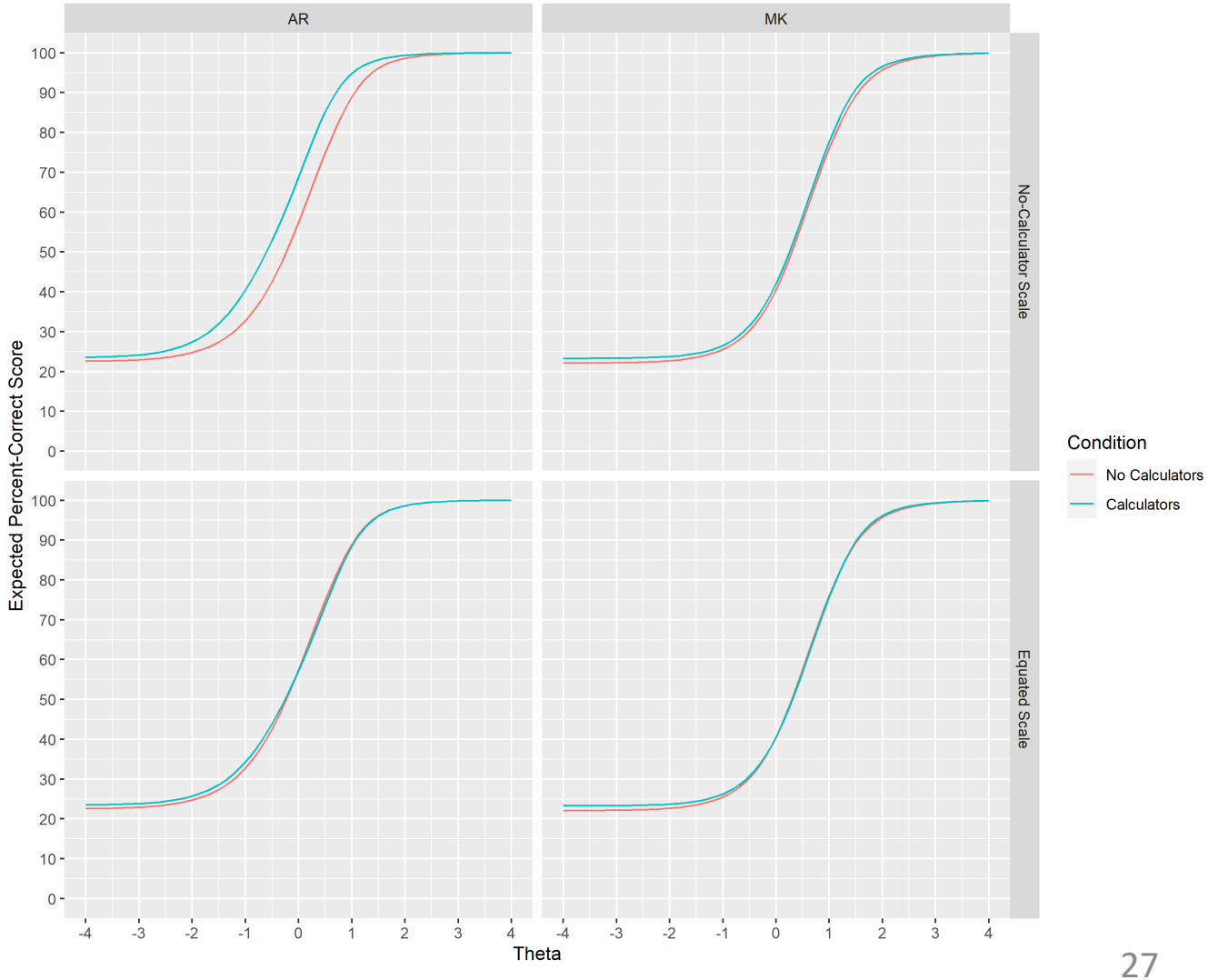
Study 2: Preliminary Results (cont.)

- *Do psychometric properties differ between study conditions (RQ2)?*
  - IRT *c* parameter comparisons between conditions:



Study 2: Preliminary Results (cont.)

- *Do psychometric properties differ between study conditions (RQ2)?*
  - IRT test characteristic curve (TCC) comparisons between conditions:



OPA
OFFICE OF PEOPLE ANALYTICS

Study 2: Preliminary Summary & Conclusions for RQ2

- *Do psychometric properties differ between study conditions (RQ2)?*
  - Calculators make AR items easier by a considerable degree, but have very little impact on the difficulty of MK items
  - The effects of calculators on scores and item difficulty parameters are primarily linear
    - After equating, TCCs for no-calculator and calculator conditions are nearly identical
    - Equating would be an essential component of introducing calculators to operational ASVAB testing (to maintain continuity of scores), so examinees will gain no systematic advantage by using calculators

Study 2: Preliminary Summary & Conclusions for RQ2 (cont.)

- *Do psychometric properties differ between study conditions (RQ2)?*
  - Given the parallelism between conditions' equated TCCs, allowing calculators could put some examinees at a disadvantage if they choose not to make full use of the calculators
    - Choosing not to (consistently) use a calculator could reduce examinees' expected rates of correct responses, and their scores would then be evaluated relative to calculator users
    - Examinees who prefer not to use a calculator would effectively test under no-calculator conditions, but be scored according to calculator-based standards
    - Scores would reflect a function of both math ability and individual differences in calculator use
  - The equivalence of equated TCCs implies administering AR and MK with calculators would (a) have no psychometric impact and (b) introduce concerns about the interpretation of scores due to individual differences in calculator usage

Study 2: Preliminary Results (cont.)

- *Do group differences in performance exist across conditions (RQ3)?*
  - Mean score differences across subgroups:

	No Calculator Scores				Calculator Scores				Effect Size	
	AR		MK		AR		MK		AR	MK
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*d*
Female	−.18	.66	.18	.63	.06	.80	.26	.68	.33	.13
Male	.09	.80	.32	.71	.38	.80	.37	.71	.37	.08
Hispanic White	−.05	.73	.16	.68	.25	.72	.24	.68	.42	.12
Non-Hispanic Asian	.25	.79	.59	.73	.41	.86	.55	.88	.18	−.06
Non-Hispanic Black	−.28	.71	.12	.64	.00	.73	.17	.66	.39	.07
Non-Hispanic White	.28	.79	.45	.69	.57	.80	.51	.69	.37	.09
English Proficiency: Yes	.07	.79	.31	.70	.36	.79	.36	.70	.37	.07
English Proficiency: No	−.38	.67	.12	.67	−.11	.69	.03	.67	.39	−.13

Study 2: Preliminary Results (cont.)

- *Does condition impact amount of time to complete each math subtest (RQ4)?*

	No Calculator Testing Time				Calculator Testing Time				Effect Size	
	AR		MK		AR		MK		AR	MK
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*	*d*
Overall	31.32	9.64	13.29	5.62	28.49	9.06	13.34	5.63	−0.30	0.01
Female	32.86	8.98	13.98	5.99	29.96	9.35	14.56	6.39	−0.32	0.09
Male	31.23	9.68	13.34	5.69	28.29	9.02	13.13	5.50	−0.31	−0.04
Hispanic White	33.27	9.78	13.96	6.55	30.38	8.66	13.71	5.68	−0.32	−0.04
Non-Hispanic Asian	33.56	9.02	13.64	4.83	29.72	9.53	14.92	6.34	−0.41	0.22
Non-Hispanic Black	34.50	9.47	14.54	6.24	32.36	9.71	14.69	6.95	−0.22	0.02
Non-Hispanic White	28.81	9.06	12.72	4.94	25.11	7.66	12.18	4.53	−0.44	−0.11
English Proficiency: Yes	31.25	9.62	13.23	5.58	28.42	9.02	13.29	5.55	−0.30	0.01
English Proficiency: No	36.15	9.05	17.17	6.73	33.78	9.80	16.22	8.75	−0.25	−0.12

Study 3:
Needs Assessment for a Math Test with a Calculator

OPA
OFFICE OF PEOPLE ANALYTICS

Study 3: Purpose

- Conduct a needs and requirements assessment to determine what the taxonomy/blueprint would be in the event a new calculator test is needed (i.e., a test assessing math content that requires a calculator).

Study 3: Procedures

- Developed online needs assessment
  - Identified the types of math that servicemembers use in training and on the job that require a calculator
  - Used existing AR and MK taxonomy as basis for content
  - Identified potential gaps on the existing taxonomies by comparing to alternative math taxonomy identified by Waugh et al. (2015)
  - Content experts (same from Study 1) conducted crosswalk comparison between the AR and MK taxonomies and Waugh et al. taxonomy
    - Of the 51 standards/statements comprising the alternative taxonomy, content experts judged that . . .
      - 9 overlapped entirely with existing AR and MK taxonomies
      - 14 partially overlapped
      - 28 had no overlap with existing AR and MK taxonomies

Study 3: Procedures (cont.)

- Met with MAPWG technical and policy reps to identify training staff *and* occupational managers across Services to receive the online needs assessment
  - Sample includes training courses and occupations covering a variety of content, including some with intensive math requirements (e.g., Air Force Precision Measurement Equipment Laboratory 2P0X1)
  - Response options for each type of math listed:*
    - Are there times in training (on the job) when trainees (servicemembers) must perform this type of math with a calculator? Y/N
    - If yes, do trainees (servicemembers) who enter training (the job) knowing how to do this type of math with a calculator perform better in training (on the job) than those who do not? Y/N
- Administered on HumRRO platform
  - Administration window: anticipated May 2024†

*Training staff respond to phrasing for "training" and "trainees." Occupation managers respond to phrasing for "on the job" and "servicemembers."
†Administration window not finalized as of April 2024.

Study 3: Results

- Needs assessment not completed as of April 2024

Next Steps

OPA
OFFICE OF PEOPLE ANALYTICS

Complete Analyses and Finalize Results

- Results presented today are preliminary
  - Finalizing preliminary results
  - Additional analyses to be conducted (e.g., dimensionality/factor analyses, items flagged for DIF)
  - Need for additional analyses may surface once preliminary analyses are finalized

Consider Implications if Calculator Use Is Allowed on AR and MK

- Logistic considerations
  - Distributing and maintaining calculators (including for overseas testing)
  - Distributing and transporting calculators for ASVAB CEP administrations
  - Determining who will provide and maintain calculators for each Service for Armed Forces Classification Test (AFCT) administrations
  - Creating training/guidance for Test Administrators
    - Including guidance on enforcement of approved calculator

Consider Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Psychometric implications
  - An equating study will be necessary
  - Will affect both the CAT and P&P formats and multiple administration purposes (AFCT, PiCAT, Vtest, ETP, etc.)
    - Will have implications for score scale as forms are recycled for different purposes
  - Between AR and MK, approximately 10,000 items have been developed, calibrated, and scaled under no-calculator conditions
    - All item parameters will need to be rescaled
  - The linear transformation constants used to convert theta estimates to standard scores are based on linking form-specific score distributions to the 1997 Profile of American Youth (PAY97) norms under no-calculator conditions
    - These constants will need to be adjusted to account for calculator effects on score distributions

Consider Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Psychometric implications (cont.)
  - Even if equated, many uncertainties persist
    - Impact(s) on validity: decades of validity evidence is based on ASVAB administered without the use of calculators
    - We have or will have only some knowledge (a snapshot based on 30 AR & 25 MK items) of psychometric impacts on:
      - Difficulty
      - Dimensionality
      - Response time
      - Fairness
      - Norms
      - Composite cut scores

Consider Implications if Calculator Use Is Allowed on AR and MK (cont.)

- Psychometric implications (cont.)
  - Even if equated, many uncertainties persist
    - Score utility
      - Interpretability of scores (score meaning/definition)
      - Potential loss of score utility without a clear score meaning
- Statutory compliance
  - USC, Title 10, Sec 520, mandates how AFQT is to be applied for the purpose of enlistment; specifically, the statute mandates a limitation on enlistment of applicants with an AFQT score between 10 and 30
  - This implies an ability to accurately estimate aptitude—allowing use of calculators on the ASVAB could result in changing the definition of the AFQT scores

Questions for the DAC

OPA
OFFICE OF PEOPLE ANALYTICS

Questions for the DAC

- To what extent should we be concerned about individual differences in calculator use?
- Other input or guidance from the DAC?

OPA
OFFICE OF PEOPLE ANALYTICS

Thank you!

For more information
please contact:

Andrea Sinclair
asinclair@humrro.org
502.966.7015

OPA
OFFICE OF PEOPLE ANALYTICS