AFRL-RH-WP-TR-2023-0059

Cyber Test Development

Christopher R. Huber
Jeffrey A. Dahlke
Brenda D. Ellis
Matthew Trippe
Sean Baldwin

Human Resources Research Organization

Interim Report
September 2023

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

The Air Force Research Laboratory Public Affairs Office cleared this report for public release and it is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2023-0059 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH THE ASSIGNED DISTRIBUTION.

CARRETTA.THOMAS.R.1228984929
Digitally signed by CARRETTA.THOMAS.R.1228984929
Date: 2023.10.16 11:52:00 -04'00'

THOMAS R. CARRETTA, PhD
Work Unit Manager
Performance Optimization Branch
Air and Space Biosciences Division

WILLIAMS.LOGAN.ANDREW.1273597634
Digitally signed by WILLIAMS.LOGAN.ANDREW.1273597634
Date: 2023.10.16 17:24:39 -04'00'

LOGAN A. WILLIAMS, PhD
Human Performance Product Area Lead
Product Development
Performance Optimization Branch
Air and Space Biosciences Division

This report is published in the interest of scientific and technical information. And its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*
12-10-2023

2. REPORT TYPE
Interim

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Cyber Test Development

5a. CONTRACT NUMBER
FA8650-21-F-4104

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Christopher R. Huber, Jeffrey A. Dahlke, Brenda D. Ellis,
Matthew Trippe, and Sean Baldwin

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H12Q

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700
Alexandria, VA 22314-1578

8. PERFORMING ORGANIZATION REPORT NUMBER
2023 No. 125

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Air and Space Biosciences Division
Performance Optimization Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHBCP

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2023-0059

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
AFRL-2023-5123, cleared 12 October 2023

14. ABSTRACT

Since 2008, several phases of research have been conducted to develop and evaluate the Cyber Test (formerly known as the Information and Communications Technology Literacy Test, or ICTL), which is used as a pre-enlistment assessment across the Services and has been shown to predict success in entry-level training for cyber-related military occupations. Due to the rapidly evolving nature of cyber knowledge, it is necessary to periodically reevaluate and refresh the Cyber Test item pool to maintain the test's functioning. To this end, we calibrated, equated, and screened 215 experimental items to supplement the existing Cyber Test item pool. We also revisited Cyber Test items previously approved for operational use and culled items that were at risk of becoming obsolete in the near future. We then assembled the updated pool of items into three parallel forms (or pools) from which the computerized adaptive test (CAT) algorithm will draw items. In addition, we updated the test blueprint with the assistance of cyber subject matter experts from across the Services and developed 205 new experimental items for pilot testing.

15. SUBJECT TERMS
Knowledge test, information test, pre-employment test, psychological tests, assessment, measurement, test scoring, cyber test, information communications technology literacy, cyber knowledge

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	37	**19b. TELEPHONE NUMBER** *(Include Area Code)*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

There are individuals not listed as authors who made important contributions to the work described in this report. We thank Dr. Brittany Crawford and Dr. Mike Zweifel for their assistance with item reviews. We are grateful to Suzanne Clark, Russell Fenton, Justin Hunsaker, Kelly Larsen, and Larry McLean for providing their valuable expertise to develop, review, and refine test content. We extend our gratitude to the military points of contact and subject matter experts who helped us update the Cyber Test blueprint. In addition, we would like to thank Dr. Jennifer Klafehn for her helpful feedback on this report. Finally, we are deeply grateful to Dr. Tom Carretta, our technical point of contact at the Air Force Research Laboratory, for his guidance and support throughout this project.

1.0 EXECUTIVE SUMMARY

Since 2008, multiple phases of research have been conducted to develop and evaluate the Cyber Test (formerly known as the Information and Communications Technology Literacy Test, or ICTL), which is used as a pre-enlistment assessment across the Services and predicts success in entry-level training for cyber-related military occupations (Trippe et al., 2014). Until recently, the Cyber Test employed a static administration format; that is, each examinee who completed a given test form saw the same set of items in the same order. In 2018, Human Resources Research Organization (HumRRO) conducted research supporting the transition to a computer adaptive test (CAT) format, wherein each examinee recieves a unique set of items that are algorithmically selected to optimize estimation of their individual ability level (Koch et al., 2018). The products of this effort included a solution to divide the Cyber Test item pool into three parallel 87-item forms (i.e., item pools)[1] from which the CAT algorithm could draw items, as well as 215 new items that were ready for pilot testing.

Given the rapidly evolving nature of cyber knowledge, the current effort sought to review, refresh, and expand the Cyber Test item pool to maintain the test's functioning and build larger, more robust CAT forms. To this end, we pilot tested, calibrated, equated, and screened the 215 items developed in 2018 for inclusion in new CAT forms. We also reviewed all items that were deemed eligible for operational use in the 2018 study and discarded items with an elevated risk of becoming obsolete in the near future. Next, we used automated test assembly (ATA) to create three new CAT forms with 112 or 113 items each. These forms were comparable with respect to coverage of target constructs and demonstrated strong evidence for psychometric parallelism. In addition, we conducted extensive screening to detect pairs of item enemies, or items that should not be co-administed due to the potential for cueing (i.e., one item's content provides or strongly suggests the correct response to another item), and we divided all enemy pairs between different CAT forms. We also designed an automated enemy detection program to supplement human review of the growing item pool and minimize false negatives in the enemy review process.

In addition to the development of new operational forms, we took steps to further expand the Cyber Test item pool in preparation for future test development efforts. First, we conducted a large review exercise with cyber subject matter experts (SMEs) from across the Services to update the test blueprint. We then recruited SME item writers to develop and refine new items based on the test blueprint. This development and review process produced 205 new experimental items, which we delivered to the Defense Testing Assessment Center (DTAC) for pilot testing.

[1] In keeping with the terminology used by DMDC (Defense Manpower Data Center, 2008), we refer to the CAT item pools as "forms".

2.0 INTRODUCTION AND BACKGROUND

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude test battery used for selection and classification of enlisted trainees across all branches of the U.S. Armed Services. Numerous studies have shown that scores on the ASVAB are valid predictors of training and on-the-job performance (e.g., Campbell & Knapp, 2001; Ree & Earles, 1992; Welsh et al., 1990). At the request of the Office of the Assistant Secretary of Defense, the Defense Manpower Data Center (DMDC) began a review of the ASVAB in 2005 because of concerns that the content had become dated due to changes in the work performed by and the attributes required of military personnel (e.g., more diverse missions, more complex organizations and systems, enhanced technology). An expert panel was convened to review the ASVAB program and to make recommendations for improvements and enhancements to the program. The review panel presented its findings in March 2006 (Drasgow et al., 2006), which included 22 recommendations. One of the panel's recommendations was that research should be conducted to develop and evaluate a test of ICTL[2], as many military jobs now require working with information and communications technology.

In response to the ASVAB review, the Air Force Personnel Center (AFPC) initiated a project in October 2007 to develop and evaluate a test of ICTL. The resulting test, now known as the Cyber Test, covers four broad content areas within the cyber domain: Networking and Telecommunications (NT), Computer Operations (CO), Security and Compliance (SC), and Software Programming and Web Design (SPWD). The Cyber Test was designed to predict success in entry-level training for cyber-related military occupations, and subsequent research confirmed that Cyber Test scores are predictive of cyber training success (Russell & Sellman, 2009, 2010; Trippe & Russell, 2011; Trippe et al., 2014). In addition to screening out candidates who are not a good fit for cyber occupations, the Cyber Test is used to "screen in" candidates who fall marginally short of entry standard requirements, thereby expanding the pool of candidates with a high potential for success in cyber occupations.

Since the Cyber Test's inception, HumRRO has carried out several test development efforts to refresh the test to keep abreast of the changing cyber domain, as well as to enhance its psychometric properties. Most recently, Koch et al. (2018) provided support for transitioning the Cyber Test from a static format to a CAT platform. This effort resulted in three 87-item CAT forms that recently rotated into operational use on the CAT ASVAB platform. In addition, Koch and colleagues developed 215 new items, which have since been pilot tested in a sample of applicants for enlisted service, to support the ongoing renewal and expansion of the Cyber Test item pool. This report documents the psychometric evaluation of these experimental items, the development of three updated CAT forms (using the experimental items and previously approved items), and the development of 205 new experimental items for pilot testing.

[2]An independent committee sponsored by the National Academy of Engineering and the National Research Council made a similar recommendation in 2006 (Garmire & Pearson, 2006).

3.0 ITEM ADMINISTRATION, CALIBRATION, AND EQUATING

3.1 Item Administration

The 215 experimental items (developed under a previous effort; see Koch et al., 2018) were provided to DTAC for administration on the ASVAB platform. Experimental items were "seeded" within existing Cyber Test forms in a similar manner to experimental ASVAB items. Specifically, 10 randomly selected experimental items from the item pool were administered to 152,693 Armed Service applicants between January 2021 and May 2022. This kind of randomization effectively controls for many potential extraneous factors (e.g., order effects) encountered in traditional pilot testing.

To create our analysis sample for calibration and equating analyses, we applied several data cleaning filters to the raw experimental dataset. First, we eliminated invalid records by identifying those with testing times of less than a minute, missing item identification values, or invalid social security numbers. We then eliminated exact duplicate records and limited the data of repeat testers to their first testing instance. We further identified and removed corrupt or otherwise invalid records, including records with (a) non-Service values, (b) invalid response values, (c) invalid response time values, (d) invalid test time values, or (e) missing response data. Characteristics of the sample used for item analysis, calibration, and equating analyses are summarized in Table 1.

Table 1: Demographic Characteristics of the Calibration Sample

Characteristic	*n*	% of Sample
Service/Component		
Air Force Guard	6,597	4.49
Air Force Regular	38,142	25.97
Air Force Reserve	4,130	2.81
Army Guard	2,540	1.73
Army Regular	4,528	3.08
Army Reserve	567	0.39
Coast Guard Regular	53	0.04
Coast Guard Reserve	12	0.01
Marine Regular	47,159	32.11
Marine Reserve	7,239	4.93
Navy Regular	33,882	23.07
Navy Reserve	2,015	1.37
Gender		
Female	20,605	14.03
Male	72,610	49.44
Unknown/Missing	53,649	36.53

Table 1: (Continued)

Characteristic	n	% of Sample
Race		
American Indian or Alaskan Native	345	0.23
Asian or Pacific Islander	2,248	1.53
Black or African American	20,926	14.25
White or Caucasian	67,259	45.8
Other	1,141	0.78
Unknown	480	0.33
Missing/Did not specify	54,465	37.09
Ethnicity		
Hispanic or Latino	4,312	2.94
Not Hispanic or Latino	28,942	19.71
Missing	113,610	77.36
Total	146,864	100.00

Note. The proportions of missing data were unusually large for gender, race, and ethnicity due to database issues at DTAC at the time our data extracts took place.

3.2 Item Calibration and Equating

We analyzed all Cyber Test items using an Item Response Theory (IRT) measurement model known as the three-parameter logistic model (3PL; Lord, 1980; Lord & Novick, 1968). In essence, IRT assumes that examinees' responses to test items are the result of underlying levels of ability possessed by those individuals. IRT is facilitated by fitting, or calibrating, statistical models to examinee responses. Application of these statistical models results in the simultaneous scaling of item difficulty and examinee (population) ability.

IRT algorithms estimate item parameters, which capture a nonlinear relationship between ability and the likelihood of answering each item correctly. In the 3PL model, the probability that an examinee with an ability estimate, theta (θ), responds correctly to item i is

$$P_i(\theta) = \frac{1 - c_i}{1 + e^{-1.702 \times a_i \times (\theta - b_i)}}$$

where a_i is the item discrimination, b_i is the item difficulty and c_i is the pseudo-guessing parameter.

Items that fit the IRT model will exhibit a pattern of lower probabilities of correct responses from low-ability examinees and higher probabilities of correct responses from high-ability examinees. This is reflected in an item characteristic curve (ICC), as depicted in Figure 1. Items vary in difficulty such that the position of the inflection point on the ICC is higher or lower (i.e., to the right or to the left) along the ability (theta) scale. For example, the inflection point of the

curve for the sample item in Figure 1 is centered at zero, which is the mean of the ability scale. An efficient test will consist of items with ICCs similar to Figure 1 but with varying difficulties ("*b*" parameters) that discriminate along the entire ability scale (i.e., feature inflection points at higher and lower values of theta). ICCs also differ in their lower asymptotes (i.e., "*c*" parameters), which reflect how easy it is to answer an item correctly by guessing, and the gradients of their slopes at the inflection point (i.e., "*a*" parameters), which indicate an item's effectiveness at differentiating between higher- and lower-ability examinees.

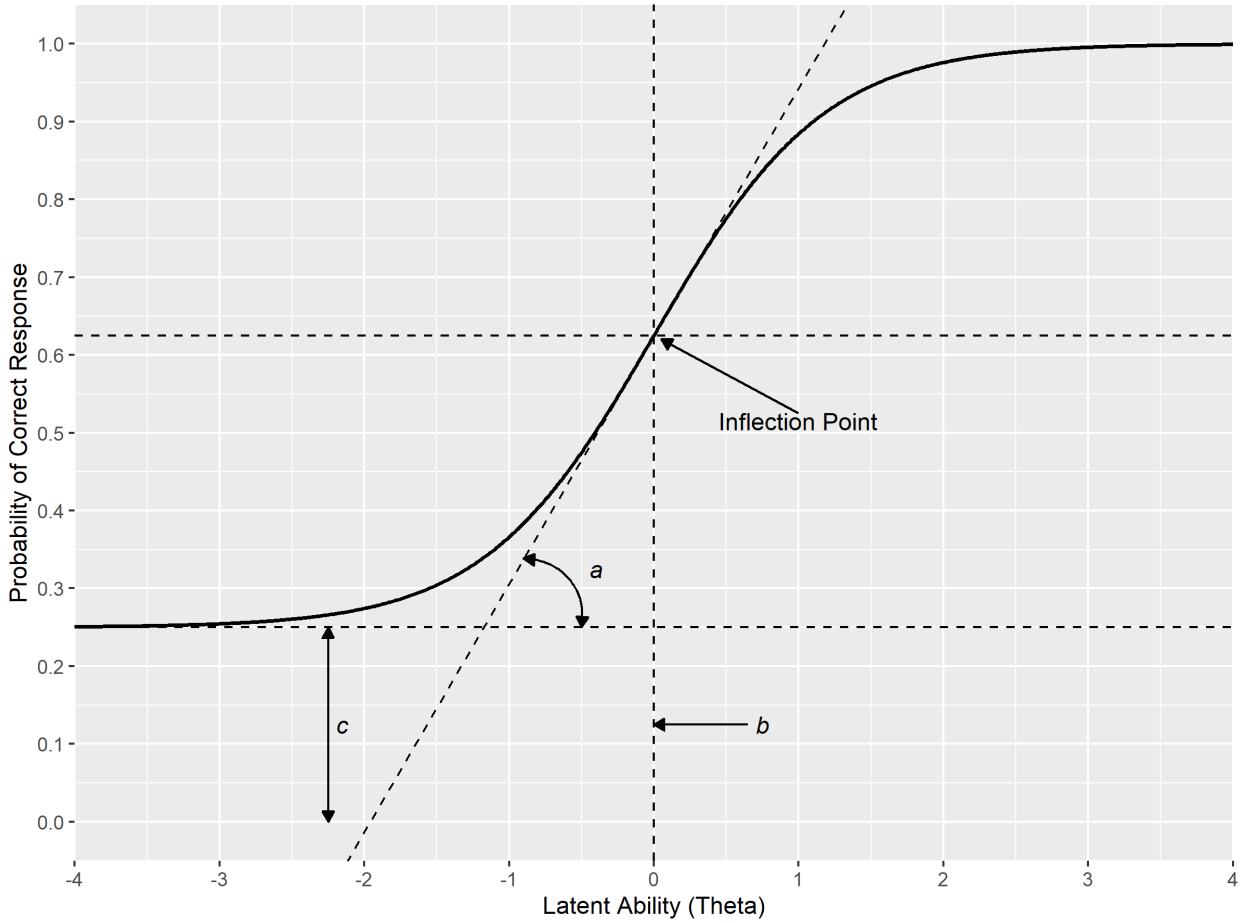

Figure 1: Example Item Characteristic Curve in the Three-Parameter Logistic Model

For the current effort, we conducted a 3PL calibration using the software program MULTILOG (Thissen, 2003). Each individual Service applicant in the calibration sample was randomly administered one of two operational Cyber Test forms (i.e., Form 3 or 4) with 10 randomly seeded experimental items, of which there were 215 in total. Each of the experimental items was administered to an average of 6,831 individuals (range: 6,570 to 7,079) in the randomized design.

We used a "maximum likelihood for fixed theta" approach for calibration, whereby parameter values are derived from a fixed or "known" ability value and an array of item responses. In this approach, we first calibrated item parameters for the 58 operational items using the traditional Marginal Maximum Likelihood (MML) framework, in which the algorithm simultaneously estimates item parameters and examinee ability parameters in an iterative fashion. We then scored each of the applicants in the calibration sample using these operational parameter values alone. The theta estimates were then standardized to a distribution with a mean of zero and standard deviation of one to counteract the compression that often results from maximum a posteriori (MAP) scoring in IRT (Embretson & Reise, 2000). It should be noted that this approach places theta estimates from two operational forms on a common scale without the use of anchor items. As such, it relies on the assumption that the two forms are psychometrically parallel and were administered to randomly equivalent groups.

After obtaining theta estimates using the operational test items, we calibrated parameter estimates for the 215 experimental items using the fixed theta framework. The fixed theta framework has a few advantages related to the stability of the calibration. First, individual item parameter values are derived independently such that a poorly estimated item cannot influence any other item. The fixed theta calibration also strongly ties the parameter estimates to the original operational construct, which minimizes the influence of potential construct drift that can result from an off-topic or otherwise poorly functioning experimental item.

The resulting item parameter estimates were on a somewhat arbitrary scale that needed to be linked back to the original operational scale established in 2011, which we accomplished via an equating process. Item parameter estimates for the operational items were previously equated in 2018 (Koch et al., 2018), and we used these 2018 values to establish the "base scale" for equating in this project. The equating process involved using items administered both in 2018 and in this effort as anchor items and applying the Stocking-Lord procedure (Stocking & Lord, 1983) to link the newly calibrated item parameters to the base scale. Specifically, we used item parameters from the current effort and the 2018 study to calculate a test characteristic curve (TCCs) for each set of parameters. We then calculated a transformational multiplier (M1) and additive constant (M2) to transform the current TCC to match the original TCC as closely as possible.

One significant threat to the validity of equating analyses is parameter drift, or legitimate changes in the statistical functioning of items over time. Specifically, changes in the functioning of anchor items undermines their use for establishing a common scale across time. As such, we evaluated anchor items for evidence of potential parameter drift. First, we placed the old and new anchor parameters on a common scale. Next, we calculated values of the squared differences at 31 quadrature points (the same used in the Stocking-Lord procedure) and computed the mean of the 31 squared differences for each item. As a relative quantitative indicator of parameter drift, items were flagged if their mean squared difference (or mean d-square) exceeded the 90^{th} percentile. We progressively removed items with the largest mean d-square values one at a time and observed how this affected mean d-square values of other anchor items and the overall solution. We also recorded items with persistent or recurring large d-square values across several iterations of the progressive equating solutions. In addition to the quantitative drift analysis, we reviewed the content of all anchor items for potential obsolescence or other evidence of construct-irrelevant influences on item performance (e.g., item meaning had changed over time

or a formerly esoteric concept had become mainstream). While we did not detect content problems in any of the 58 anchor items, we did remove two anchor items with exceptionally large mean d-square values based on quantitative evidence alone. We then implemented the Stocking-Lord (1983) procedure using a total of 56 operational items as anchors. Finally, we used the resulting constants (M1 = 1.0723, M2 = -0.1617) to transform all experimental item parameters to the original operational scale.

4.0 TECHNICAL AND SENSITIVITY REVIEW

4.1 Post Hoc Sensitivity Review

We examined all 215 experimental items for evidence of differential item functioning (DIF), which occurs when members of different groups (e.g., males and females) with the same latent ability score have different probabilities of answering an item correctly. Items that showed substantial evidence for DIF were flagged for *post hoc* sensitivity review to avoid potential bias against a particular subgroup.

We conducted DIF analyses in five subgroup samples: males, females, non-Hispanic Blacks, non-Hispanic Whites, and Hispanic Whites. These groups were chosen to be consistent with designations used by the ASVAB testing program (Defense Manpower Data Center, 2014). Table 2 summarizes the subgroup comparison analyses in which item performance of a focal group is compared to performance of a reference group. We computed an Empirical Bayes Mantel-Haenszel statistic (Zwick et al., 1999) for each item and subgroup comparison. Table 3 summarizes the classification framework for Mantel-Haenszel results described in Zwick et al. (1999).

Table 2: Subgroup Comparisons in Differential Item Functioning Analyses

Label	Reference Group	Focal Group
F/M	Males	Females
B/W	Non-Hispanic White	Non-Hispanic Black
H/W	Non-Hispanic White	Hispanic White

Black/White (B/W), female/male (F/M)

Table 3: Classification of Differential Item Functioning Results

Notation	Description	Mantel-Haenszel Value
A	Negligible DIF	$\lvert EB_{MH} \rvert < 1.0$
B	Slight to moderate DIF	$1.0 \leq \lvert EB_{MH} \rvert < 1.5$
C	Moderate to severe DIF	$\lvert EB_{MH} \rvert \geq 1.5$
+	Direction favors Focal group	$EB_{MH} > 0.0$
-	Direction favors Reference group	$EB_{MH} < 0.0$

As shown in Table 4, we did not find evidence of DIF based on race or ethnicity, but one item showed category C (moderate to severe) DIF between males and females. However, statistical evidence of DIF is not sufficient to conclude that an item is biased and should be removed. Differences in relative difficulty may, in fact, represent construct-relevant variance that is not necessarily bias. That is, there may be true between-group differences in the specific aspect of the construct domain assessed by a given item. An item or test cannot be said to be truly biased unless the source of the differential functioning is determined to be construct-irrelevant, which requires logical analysis of the item or test content (Camilli & Shepard, 1994). As such, we reviewed the item identified as demonstrating category C DIF between males and females. We did not identify any construct-irrelevant factors that could lead to unfairness in the item and therefore did not remove any items due to results of the DIF analyses.

Table 4: Number of Items Displaying Differential Item Functioning

Comparison	*n* A Items	*n* B+ Items	*n* C+ Items	*n* B- Items	*n* C- Items
F/M	204	0	0	10	1
B/W	215	0	0	0	0
H/W	215	0	0	0	0
Total	634	0	0	10	1

As noted earlier, we were missing gender, race, and/or ethnicity data for a significant proportion of our analysis sample due to database issues at DTAC. As such, our DIF statistics were estimated with less precision than they would be otherwise. However, it is unlikely that the outcomes of our sensitivity review would be substantially different given more complete data. Even with this limitation, our subgroup sample sizes were still very large (see Table 1), providing more than adequate precision regardless. Furthermore, it is unlikely that any biased items escaped notice, since (a) all items underwent a sensitivity and bias review when they were first developed and (b) items selected for the eligibility pool received several additional rounds of content review; as such, the post hoc sensitivity review was only one of many opportunities to uncover inappropriate content.

4.2 Post Hoc Item Quality Review

In addition to items that we calibrated during this research, the eligibility pool for our CAT form assembly process included older items from earlier phases of content development. Although all historic items we considered had already passed quality reviews during previous phases of Cyber Test development, we needed to verify they were still relevant and had not become obsolete due to the rapid rate at which computing and networking technologies advance. Therefore, we conducted quality and obsolescence reviews for all items that could potentially be included in the eligibility pool for CAT form assembly. The focus of these reviews differed slightly depending on whether an item was newly calibrated or carried forward from the previous round of form development. These reviews are described in the sections that follow.

4.2.1. Review of Newly Calibrated Items

Two researchers reviewed all 58 operational items analyzed in this research for obsolescence as part of our calibration process. We did not find evidence that any of these items assessed knowledge about obsolete technologies.

After completing our calibration analyses, we conducted another review of both operational and experimental items to identify items with psychometric problems and/or obsolete content. We preemptively excluded 43 items from this review because they had negative *a* (item discrimination) parameters and were therefore ineligible for future operational use. In addition to any other obsolete content, we instructed reviewers to look for items with the following risk factors for future obsolescence:

- References to trending, fad, or niche technology
- References to *specific* versions of software, hardware, or a standard (e.g., Windows 11, USB-C)
- Temporal references (e.g., "new," "newer," "old," "older")
- References to application-specific content (e.g., Excel vs. spreadsheet software)
- Content related to scale (e.g., MB, megahertz [MHz])

We also reminded reviewers that items were less likely to become obsolete if they referenced established technologies, fundamental concepts (e.g., storage, memory, speed), and content that is likely to remain stable over years rather than months.

For the psychometric component of the review, we flagged items that met any of the following criteria:

- Negative item-total correlation
- Positive item-total correlation less than 0.1
- Proportion of correct responses (p-value) less than 0.1
- Distractor had an option-total correlation greater than the item-total correlation for the keyed response
- Distractor had an option-total correlation greater than 0.1
- B-level or C-level DIF for any subgroup contrast

Two researchers who had not previously seen the item content independently evaluated all 230 items within the scope of this review, with a third researcher serving as adjudicator to resolve any disagreements. In one case, we consulted with cyber SMEs—i.e., two of the item writers described in the New Item Development section below—to resolve a technical question about an item that was ultimately discarded. As a result of this evaluation, the reviewers dropped 66 items due to psychometric or content-related issues, leaving 164 items in the preliminary eligibility

pool for form assembly. In addition to the reviews described above, we continued to evaluate item quality during subsequent rounds of form assembly and item enemy review. Three additional items were ultimately discarded based on these reviews, leaving 161 newly calibrated items in the final eligibility pool.

4.2.2. Review of Items from Past Calibrations

In addition to the 58 operational items discussed in the previous section, 208 other historic items were considered eligible for use in 2018 when the previous set of CAT pools was developed. Since we did not recalibrate these items, it was not necessary to conduct a new psychometric review; however, two researchers reviewed all 208 items for obsolescence. Since these items were older than the experimental items and did not have recent parameter estimates, our approach for identifying obsolescence was more stringent and included a verification that the technologies and standards referenced in the items were still in use. The reviewers dropped 31 items due to concerns about obsolescence or problems related to content, leaving 177 of these items eligible for assignment to the new CAT forms.

4.2.3. Final Eligibility Determinations

After accounting for the decisions from all reviews, the eligibility pool for CAT form assembly contained a total of 338 items. See Table 5 for a breakdown of item eligibility decisions.

Table 5: Summary of Item Eligibility Decisions

Item Source Label	Development Year	Total Items Considered	Dropped Due to Negative *a* Parameter	Dropped During Item Reviews	Eligible for CAT Form Assignment
Exp P&P	2011	45	0	4	41
190 Seed	2014	106	1	15	90
251 Seed	2017	115	0	22	93
215 Seed	2018	215	42	59	114
Total	-	481	43	100	338

5.0 PRELIMINARY FORM ASSEMBLY

After obtaining equated parameters for all items and specifying the pool of eligible items, the next step was to develop parallel item pools, or forms, for use with the CAT ASVAB assessment platform. We consider these forms "preliminary" because they do not represent the final operational forms. Rather, the final forms are established by DTAC based on operational testing simulations conducted using the preliminary forms produced by HumRRO. If the CAT algorithm never selects a given item for administration to simulated test-takers, that item is dropped, resulting in operational forms that are a smaller subset of the preliminary forms discussed here.

Consistent with the previous form development effort (Koch et al., 2018), we developed three parallel CAT forms from the pool of eligible items. Although this test will be administered by a CAT algorithm and a given examinee will not be exposed to every item on a form, it is still important to ensure that the forms are relatively parallel so the test's measurement properties (e.g., information provided at a given theta level) will not differ across CAT forms. While true parallelism is an impossible-to-achieve abstraction if the forms contain non-identical items (Lord, 1980), it is possible to achieve effective or practical parallelism/equivalence by balancing key psychometric and content characteristics across the forms. For instance, forms should be balanced with respect to (a) item content, (b) difficulty, (c) discrimination, and (d) keyed responses. In addition, item "enemies" that include a high degree of overlapping content, such that viewing one item effectively provides the correct response to the other, should be accounted for.

It is very difficult, however, to balance all of these objectives simultaneously. For example, two forms manually constructed to have equal content and key distributions will likely differ dramatically in their difficulty and discrimination. Therefore, we used ATA (van der Linden, 2005) to determine the optimal distribution of items between forms to balance competing test specifications. Although ATA can refer to a variety of different algorithms for test assembly, a common approach is to use binary/integer programming to reframe the problem as a mathematical optimization process. In this approach, an objective function is identified, which is the quantity that is to be minimized or maximized, and each of the test specifications is recast as a mathematical inequality that constrains the set of possible solutions.

To solve our specific problem, we used the basic ideas presented in van der Linden (2005) and Diao and van der Linden (2011) but developed our own implementation in SAS (Statistical Analysis Software) using PROC OPTMODEL. The objective function we minimized was the distance between the forms' test information functions (TIFs) and the target TIF. We defined the target TIF as the information function of the entire eligibility pool divided by the number of assembled forms (i.e., three). This target was set recognizing that all items would be assigned and that the goal of CAT form assignments should be to maximize the statistical similarity of all assembled forms. We also defined a TCC target as the TCC of the input file divided by the number of forms to be assembled. We minimized the objective function within several content and statistical constraints, including the number of desired items per form, items per content area, frequency of item keys, minimal differences between the form TCCs and the target TCC (<=.40 raw score points), and the exclusion of any known enemy co-assignments.

The above approach used content and statistical targets defined by the eligibility pool itself, recognizing that optimally dividing the input file into three forms is not dependent upon other external information (e.g., a stand-alone test blueprint or defined statistical targets). Accordingly, the linear program solver had sufficient computing power to make draft assignments in two steps (first for one form; next for the remaining two forms). These draft assignments were used as a warm start to our final solve, in which we tightened the allowable boundary about the target TCC and target TIF and reconsidered item-to-form assignments for all three forms simultaneously, allowing across-form swaps to improve the correspondence of each form to the statistical targets.

Two researchers reviewed the resulting draft forms for enemy co-assignments. Due to the size of the existing eligibility pool and its continual growth, existing enemy coding in the eligibility pool

is never complete. The content review of draft assignments led to the identification of previously unknown enemies and a revised enemy code for each affected item. This revised enemy information was accounted for in a secondary ATA program. This second program was nearly identical to the original ATA draft programming but additionally minimized item swaps from the original item-to-form draft assignments. With each new round of enemy coding, some draft item assignments broke content constraints. The solver sought to minimally change the draft assignments while addressing newly identified enemy co-assignments. This was an iterative process between content review and form revision.

As the eligibility pool grows across each iteration of test development, the number of potential enemy pairs to consider increases exponentially. This, in turn, increases the likelihood of overlooking true enemy pairs and reduces the overall feasibility of the iterative review process described above. To address this growing problem, we developed an automated enemy detection program in R and Python to supplement our iterative reviews of individual draft forms. This program compared item text for every possible combination of two items in the eligibility pool and generated two similarity metrics: one based on overlap in semantic meanings between the two items, and the other based on the proportion of overlap in nontrivial words. Item pairs with relatively high similarity scores on either metric were then flagged for human review. After testing and refining this method during the first round of draft form reviews, we applied it to the entire eligibility pool to preemptively avoid enemy co-assignments in subsequent rounds of form assembly.

Our third round of revisions resulted in the form assignments described in greater detail below. These assignments met the assembly team's standards and included no known enemy co-assignments.

5.1 Solution

We divided the item eligibility pool into three parallel CAT forms with 112 or 113 items each. Figure 2 presents the TCCs for the final three-form solution, and Figure 3 presents the forms' TIFs. Due to the extremely high degree of overlap in the TCCs, separation between the forms cannot be readily seen in Figure 2. The high degree of similarity between the TIFs also provides strong evidence of psychometric parallelism.

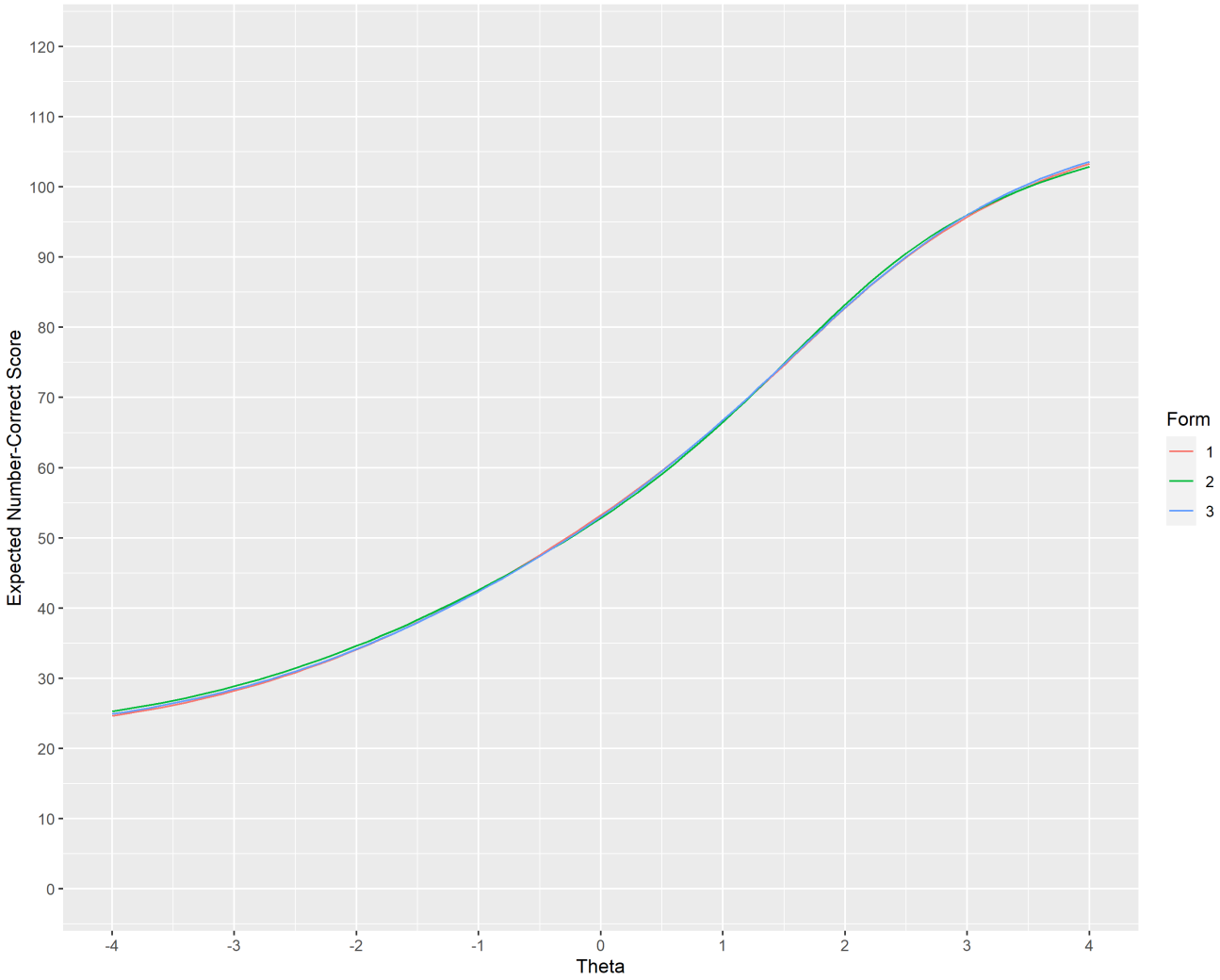

Figure 2: Test Characteristic Curves (TCCs) for the Final Three-Form Solution

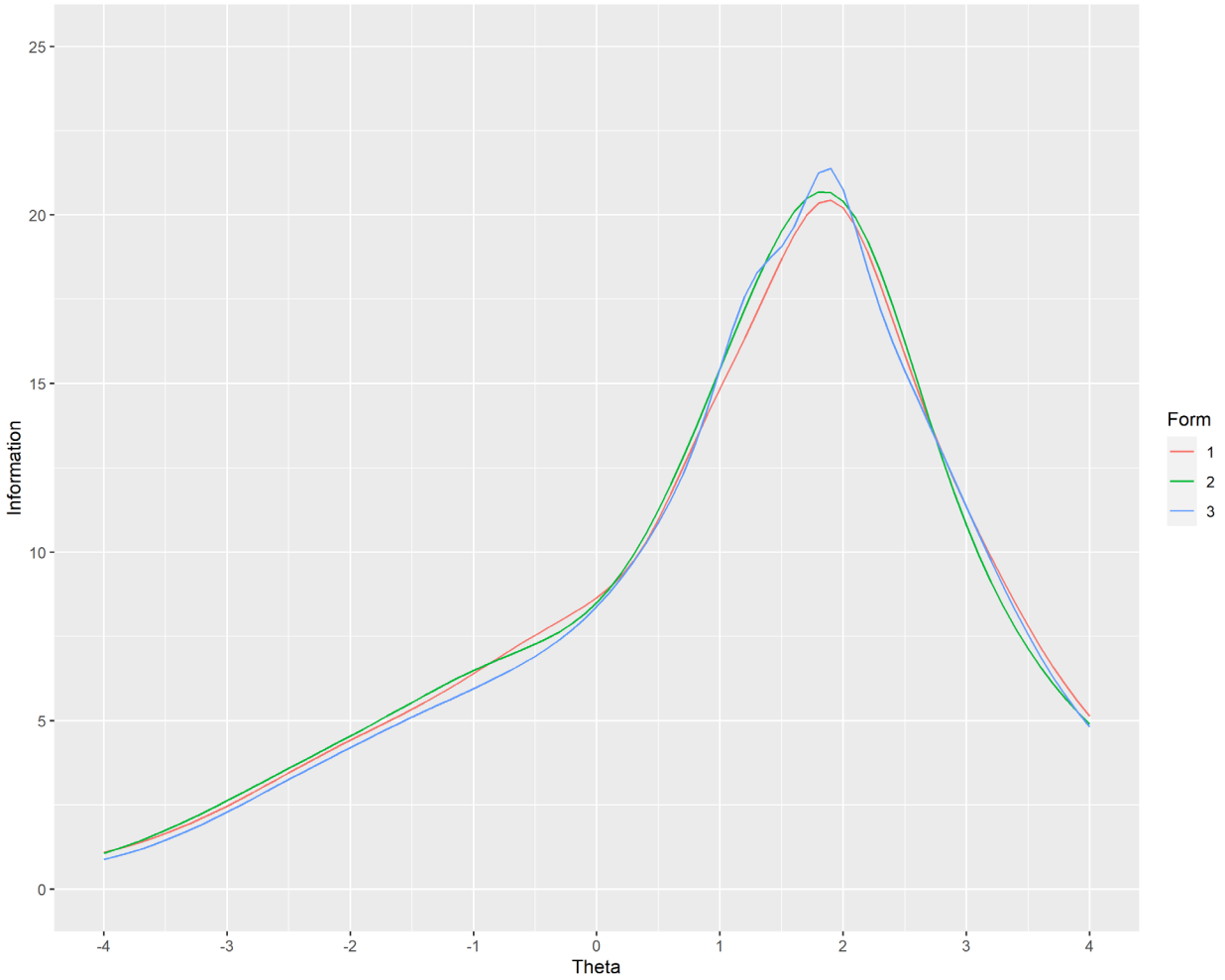

Figure 3: Test Information Functions (TIFs) for the Final Three-Form Solution

Tables 6 and 7 summarize the content and keyed response distributions for the final ATA solution. The three forms were close to one another in terms of their content distributions and tended to be quite close to the distribution specified in the test blueprint (discussed in greater detail later). There were, however, small deviations from the blueprint due to the properties of the overall eligibility pool. We placed less emphasis on equalizing the keyed response distributions and obtaining a perfect 25% in each category to focus more heavily on including as many items as possible and optimizing more important aspects of parallelism across forms. All of the keyed response percentages were between 22% and 29%, which indicates a good balance of key distributions across forms.

Table 6: Content Distributions for New CAT Forms

Content Area	Blueprint %	Form A		Form B		Form C	
		n	%	*n*	%	*n*	%
CO	30	38	33.63	40	35.71	38	33.63
NT	30	31	27.43	32	28.57	32	28.32
SC	25	30	26.55	28	25.00	30	26.55
SPWD	15	14	12.39	12	10.71	13	11.50

Note. CO = Computer Operations; NT = Networking & Telecommunications; SC = Security & Compliance; SPWD = Software Programming & Web Design. To avoid confusion with the names of existing operational forms (i.e., Forms 1-3), we used the labels "Form A", "Form B", and "Form C" here. However, Forms A-C are labeled as Forms 1-3 in other deliverables from the current project.

Table 7: Keyed Response Distributions for New CAT Forms

Form	% A	% B	% C	% D
1	23.01	26.55	22.12	28.32
2	24.11	26.79	22.32	26.79
3	24.78	24.78	22.12	28.32

Note. Due to rounding, the row-wise sums may not equal 100%.

5.2 Comparison Between New and Operational CAT Forms

The Air Force and DTAC recently implemented the set of three CAT forms HumRRO developed in 2018 (Koch et al., 2018). To ensure our new forms met or exceeded the operational standard, we examined the psychometric parallelism between the 2018 forms and our three-form ATA solution. TCCs and TIFs represent summations of values across all items on a form, which means that both metrics are affected by the number of items on a form. Since the new CAT forms include more items than the operational CAT forms, we shifted our focus from summations of item-level values to averages of item-level values to account for differences in item counts while still making useful comparisons between the two sets of forms.

Figure 4 depicts the average ICC for each CAT form. The average ICCs for both sets of CAT forms are very similar, although the average ICCs for the new set of forms are slightly lower than the average ICCs for the operational forms at both tails of the ability distribution. Since the forms will be administered adaptively rather than as fixed forms, exact parallelism of ICCs/TCCs is not required for the forms to produce comparable ability estimates. Furthermore, achieving such parallelism would come at the cost of omitting acceptable items from the eligibility pool or allowing items to overlap between forms; otherwise, average properties of the overall eligibility pool will carry over when that pool is subdivided into parallel forms.

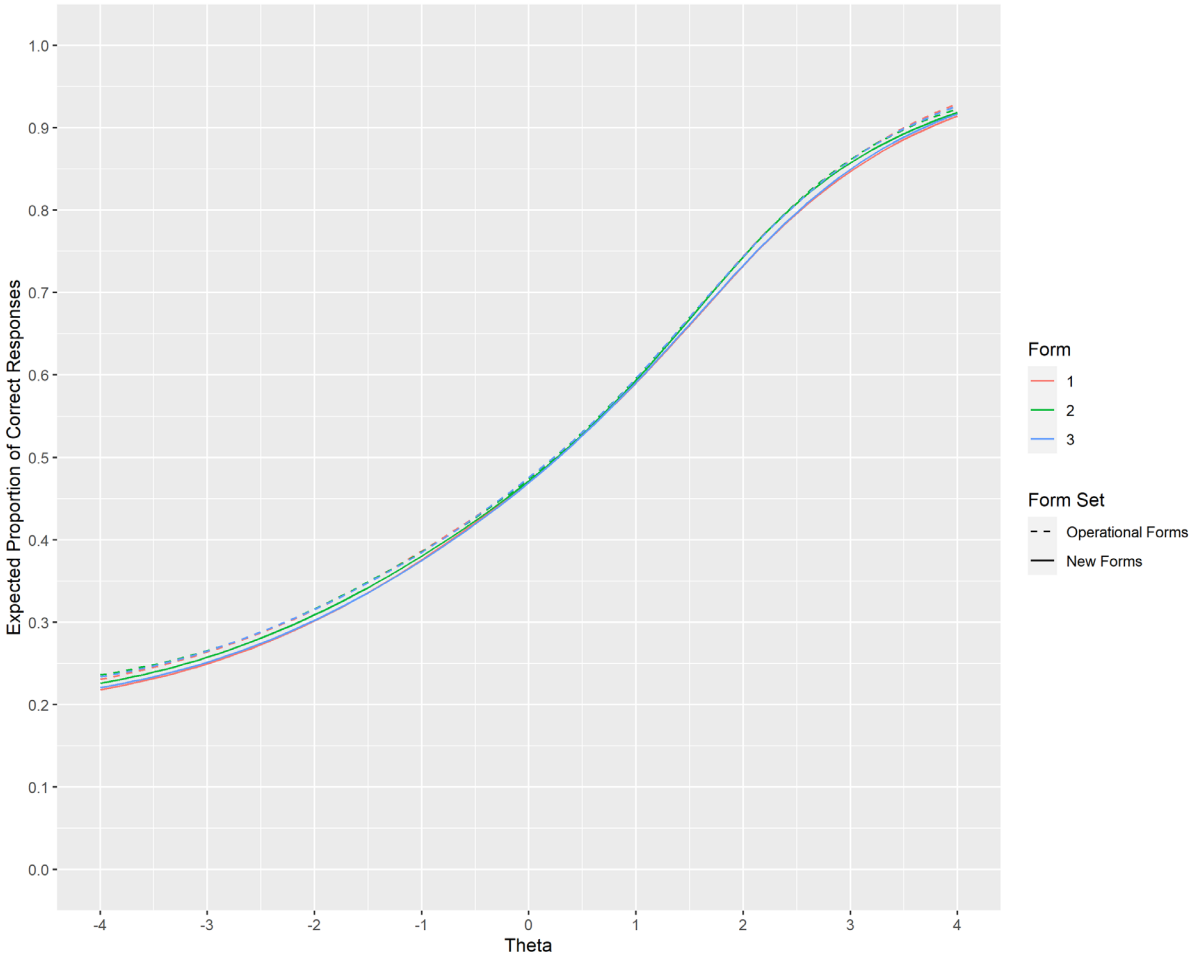

Figure 4: Average Item Characteristic Curves (ICCs) for the Final Three-Form Solution Compared to the Operational Cyber Test CAT Forms

When comparing CAT forms, it is more important to evaluate parallelism in terms of information. Figure 5 depicts the average item information function (IIF) for each CAT form. The average IIFs were similar between the new forms and the operational forms, which suggests that both sets of CAT forms should be able to produce scores with similar levels of precision. On the other hand, maintaining parallelism in terms of the average information per item means that, by increasing the number of items, the new forms increase the depth of the CAT pools across the ability distribution. This could improve estimation insofar as a small quantity of discriminating items at any given theta is a limiting factor for the operational CAT. For example, while both sets of forms have relatively little information around -2 compared to 2 on the theta scale, that same relative deficit may be more impactful in a smaller CAT pool, where the CAT algorithm is more likely to run out of optimally discriminating items to administer to low-ability examinees.

Differences in difficulty and peak precision between the newly assembled and current operational exams are expected due to our current approach of assigning all eligible items into one of three unique CAT forms. Statistical differences between eligibility pools will necessarily evidence themselves in the assembled forms as long as every item is assigned to a single form.

Given the small magnitude of the differences, as well as a CAT's ability to dynamically select optimal items from pools with different average characteristics, this is not a significant cause for concern. Regardless, in keeping with past phases of this work, we recommend that future efforts continue to monitor for "drift" of the average IIF peak toward the high end of the distribution. As we will discuss later, we are engaged in ongoing efforts to prevent this drift by promoting the development of less difficult items.

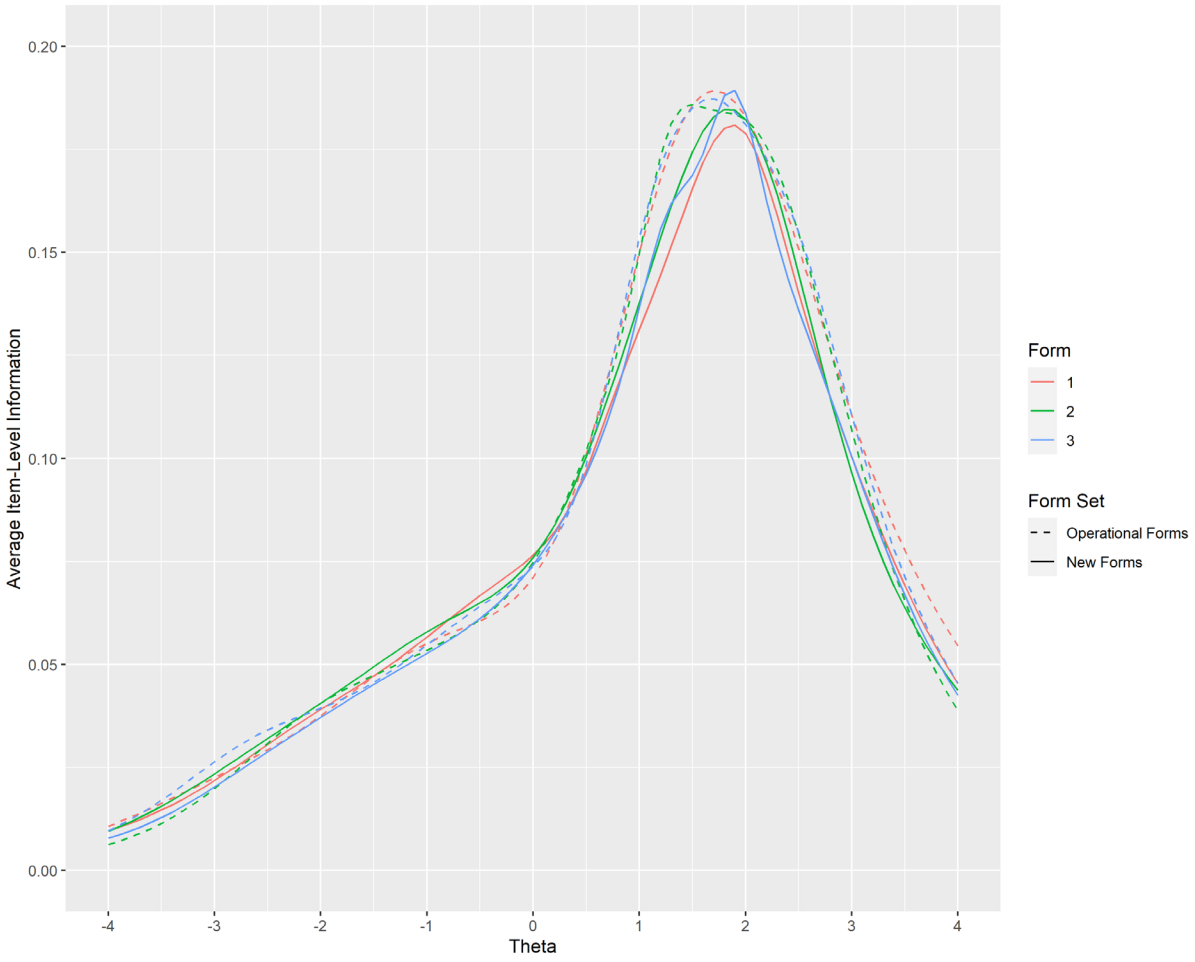

Figure 5: Average Item Information Functions (IIFs) for the Final Three-Form Solution Compared to the Operational Cyber Test CAT Forms

5.3 Form Assembly Summary

We used ATA to develop a three-form solution with forms that were as close as possible to parallel while also balancing other test specifications, such as content and key distributions. This process was successful, with all available items being used, no items being used on multiple forms, TIF and TCC plots that appeared largely parallel, and content and key distributions that were quite close to test specifications. Though outside the scope of the current effort, the next step will be to evaluate how well these functionally parallel forms perform in a CAT simulation.

6.0 NEW ITEM DEVELOPMENT

Due to factors such as item obsolescence and item exposure, there is an ongoing need to refresh the Cyber Test item pool. The final task in this project involved updating the test blueprint and developing a new set of items suitable for pilot testing.

6.1 Blueprint Validation

The test blueprint upon which the Cyber Test is based was originally developed in 2008 (Russell & Sellman, 2009) and updated in 2012 (Trippe et al., 2014), 2015 (Koch et al., 2017), and again in 2018 (Koch et al., 2018). The blueprint is organized hierarchically with four broad content areas at the highest level—namely, NT, CO, wC, and SPWD. Subsumed within each broad content area are several sub-content areas that are more specific and focused. At the lowest, most specific level of the blueprint hierarchy are knowledge, skills, and abilities(KSAs) statements that serve as the basis for item development.

Due to the transitory nature of some content within the scope of the blueprint, it is useful to periodically review the existing blueprint and KSA statements for both obsolescence and job relevance. With assistance from the Air Force, HumRRO convened a panel of military SMEs to conduct a blueprint review and validation workshop. The purpose of this review was to determine the relevance of the most recent blueprint to contemporary entry-level training for cyber-related occupations. We held a Joint-Service teleconference to introduce the Cyber Test research program and solicit input on the blueprint from SMEs. Sixty-two noncommissioned officers (NCOs) in cyber-related occupations from the Air Force (30), Navy (10), Army (4), Marine Corps (12), and Coast Guard (6) were invited to participate in the teleconference and provide input on the blueprint via a follow-up survey. We received 33 responses to the blueprint validation survey. The responding SMEs were affiliated with the Air Force (17), Navy (6), Army (4), Marine Corps (4), and Coast Guard (2). SMEs reported an average of 11.91 years of experience in cyber-related occupations, with a range of 0 to 21 years of experience.

We provided SMEs with the broad content areas, sub-content areas, and 41 KSA statements from the most recent Cyber Test blueprint. Additionally, we asked SMEs to rate 73 KSA statements from the National Initiative for Cybersecurity Education (NICE)'s Cybersecurity Workforce Framework, which describes a variety of cybersecurity jobs and provides a list of KSAs required to perform these jobs (see http://csrc.nist.gov/nice/framework/). For each KSA statement, we asked SMEs to provide a rating regarding (a) how important the KSA is for successful performance in entry-level training for enlisted cyber occupations (not at all important, a little important, somewhat important, very important, extremely important), (b) whether the KSA should be acquired prior to enlistment (yes, no), and (c) how stable the KSA will be over time (likely to change in 2 years or less, likely to change in 2 to 5 years, likely to change in 5 to 10 years, likely to change in 10 years or more, not likely to change at all).

Table 8 displays the 10 KSA statements with the highest importance ratings. Overall, the most important KSA statements tended to capture fundamental concepts that serve as the building blocks for more specific KSAs at lower levels of the blueprint hierarchy.

Table 8: KSAs Receiving Highest Importance Ratings

Category	KSA Statement	*M*	*SD*
CO	Ability to search on-line and other resources to obtain information that will help solve a problem (e.g., using Boolean logic to customize searches).	4.30	1.02
NT	Knowledge of common network terminology.	3.94	1.00
CO	Knowledge of basic computer concepts (bit, byte, CPU).	3.94	1.09
CO	Ability to connect PC hardware components (e.g., monitor, printer).	3.67	1.29
SC	Knowledge of computer-based technology that has potential for exploitation by adversaries.[a]	3.63	1.07
NT	Knowledge of common network tools (e.g., ping, traceroute, nslookup) and interpretation of the results.	3.55	1.18
NT	Knowledge of network addressing concepts.	3.55	1.06
SC	Knowledge of network security features (e.g., firewalls).	3.55	1.06
SC	Knowledge of cybersecurity and privacy principles.[a]	3.52	1.06
SC	Knowledge of basic system administration, network, and operating system hardening techniques.	3.50	1.14

Note. NT = Networking and Telecommunications; CO = Computer Operations; SC = Security and Compliance. Importance was rated on a scale from 1 to 5, where 1 = Not at all important, 2 = A little important, 3 = Somewhat important, 4 = Very important, and 5 = Extremely important.

[a]Not included in final blueprint due to low Needed at Entry ratings.

In addition to importance ratings, we asked SMEs to provide “needed at entry" ratings for each KSA statement. Specifically, we instructed SMEs to indicate whether each KSA should be acquired prior to enlistment (needed at entry) or following enlistment (not needed at entry). Table 9 shows the 10 KSA statements that received the highest “needed at entry” ratings. Like the importance ratings, this list primarily reflected basic, foundational KSAs needed to acquire more advanced knowledge and skills.

Table 9: KSAs Receiving Highest "Needed at Entry" Ratings

Category	KSA	% Indicating Needed at Entry
CO	Ability to connect PC hardware components (e.g., monitor, printer).	70
CO	Ability to search on-line and other resources to obtain information that will help solve a problem (e.g., using Boolean logic to customize searches).	67
NT	Knowledge of common network terminology.	64
CO	Knowledge of basic computer concepts (bit, byte, CPU).	64
CO	Knowledge of the functions and operation of typical PC hardware and peripherals (e.g., central processing units [CPUs], network interface cards [NICs], data storage).	61
CO	Knowledge of word processing software (e.g., Microsoft Word, OpenOffice Writer).	61
CO	Knowledge of presentation software (e.g., Microsoft PowerPoint, OpenOffice Impress).	61
CO	Knowledge of electronic devices (e.g., computer systems/components, access control devices, digital cameras, electronic organizers, hard drives, memory cards, modems, network components, printers, removable storage devices, scanners, telephones, copiers, credit card skimmers, facsimile machines, global positioning systems [GPSs]).	58
CO	Knowledge of spreadsheet software (e.g., Microsoft Excel, OpenOffice Calc).	55
SC	Knowledge of Internet, website, and email vulnerabilities.	52

Note. NT = Networking and Telecommunications; CO = Computer Operations; SC = Security and Compliance. Needed at entry was rated as yes = should be acquired prior to enlistment, and no = should not be acquired prior to enlistment.

The SMEs were also asked to estimate the rate of obsolescence for each KSA statement using the scale shown in Table 10. Higher scores indicate a slower rate of obsolescence. Table 11 displays the 10 KSA statements that received the highest scores, indicating a high degree of stability. Overall, the KSAs rated as the most stable tended to reflect fundamental concepts (e.g., common terminology, tools, commands, or components).

Table 10: Obsolescence Rating Scale

Rating	Assigned Score
Likely to change in 2 years or less	1
Likely to change in 2 to 5 years	2
Likely to change in 5 to 10 years	3
Likely to change in 10 years or more	4
Not likely to change at all	5

Table 11: KSAs Receiving Highest Obsolescence Ratings

Category	KSA Statement	*M*	*SD*
CO	Knowledge of basic computer concepts (bit, byte, CPU).	4.58	0.90
SPWD	Understanding of different numbering systems such as hex and binary	4.33	1.08
NT	Knowledge of telecommunication protocols (e.g., TCP/IP, OSI layers, token rings).	4.30	0.98
NT	Knowledge of different types of network communication (e.g., Local Area Network [LAN], Wide Area Network [WAN], Metropolitan Area Network [MAN], Wireless Local Area Network [WLAN], Wireless Wide Area Network [WWAN]).	4.21	0.89
NT	Knowledge of common network tools (e.g., ping, traceroute, nslookup) and interpretation of the results.	4.21	0.93
NT	Knowledge of different ports & their functions (e.g., SSL, port 80).	4.19	1.00
NT	Knowledge of network essentials (e.g., hub v. switch; types of networks).	4.18	1.04
NT	Knowledge of common network terminology.	4.15	1.06
CO	Knowledge of the functions and operation of typical PC hardware and peripherals (e.g., central processing units [CPUs], network interface cards [NICs], data storage).	4.15	1.15
NT	Knowledge of network protocols, standards, and directory services (e.g., Transmission Critical Protocol/Internet Protocol [TCP/IP], Dynamic Host Configuration Protocol [DHCP], Domain Name System [DNS]) and how they interact to provide network communications.[a]	4.15	1.09

Note. NT = Networking and Telecommunications; CO = Computer Operations; SPWD = Software Programming and Web Design. Obsolescence was rated on a scale from 1 to 5, where 1 = Likely to change in 2 years or less, 2 = Likely to change in 2 to 5 years, 3 = Likely to change in 5 to 10 years, 4 = Likely to change in 10 years or more, and 5 = Not likely to change at all.
[a]Not included in final blueprint due to low Importance and Needed at Entry ratings.

We used SME ratings of the importance, stability, and "needed at entry" status of each KSA statement to identify the final list of KSAs for the blueprint. In making these determinations, we chose to err on the side of retaining KSAs that appeared on the most recent blueprint. Our rationale for doing so was that (a) previous SME groups identified these KSAs as being important and needed at entry, and (b) the aggregated judgments of the current SME sample are partially a function of sampling error. In other words, differences in KSA ratings between SME groups can occur simply due to idiosyncrasies of the specific individuals sampled in each group (e.g., general rating biases or unequal representation of specific cyber occupations), as opposed to reflecting true changes in the cyber domain over time. In addition, the KSAs from the most recent Cyber Test blueprint were still generally ranked among the top-rated KSAs in the new survey; therefore, we retained all 41 KSAs from the previous blueprint.

We used higher standards to determine which KSAs from other sources to retain because these KSAs had not been identified as critical by previous SME groups. Consistent with previous blueprint exercises for this test (Koch et al., 2018), we planned to retain KSAs that were not on the most recent cyber blueprint if:

- More than 50% of respondents agreed that the KSA was needed at entry,
- The mean importance rating was greater than or equal to 3 (where 3 = somewhat important), and
- The mean obsolescence rating was greater than or equal to 2.

However, no KSAs met these thresholds. As such, no new KSAs were added to the blueprint.

SMEs were also offered the opportunity to add important content areas or KSAs that were not included in the KSA list. SMEs suggested 17 new KSAs. Upon review of the suggested KSAs, we determined that these newly-suggested KSAs would not be assessed on the Cyber Test for reasons including: (a) the KSA represented abilities or skills not specific to cyber occupations (e.g., logical thinking, detail orientation, study skills), (b) the KSA was subsumed by other KSAs on the list, and (c) the KSA was determined to be too advanced for the test taker population (e.g., Kubernetes clusters).

The final blueprint consisted of 41 KSAs, all from the previous Cyber Test blueprint. See Table 12 for a summary of the final blueprint.

Table 12: Final Cyber Test Blueprint

Broad/Subordinate Content Area	Number of KSAs
Networking and Telecommunications	12
Networking	6
Telecommunications	6
Computer Operations	15
PC Configuration and Maintenance	10
Using IT Tools/Software	5
Security and Compliance	9
System Security	4
Offensive Methods	5
Software Programming and Web Design	5
Software Programming	3
Numbering Systems	1
Database Development and Administration	1

Note. $n = 33$.

We also asked SMEs to participate in a weighting exercise to determine the proportion of test items that should be devoted to each content area. Specifically, we asked them to determine how many test items should measure each content area by assigning weights (totaling 100) to the content areas, using multiples of five percentage points. Results of the weighting exercise are presented in Table 13.

Table 13: Blueprint Category Weights from the SME Weighting Exercise

Category	*M*	*SD*	Min	Max	Final Weight	2018 Weight	2015 Weight	2012 Weight
Networking and Telecommunications	28.64	7.83	15	50	30	30	30	35
Computer Operations	31.21	8.48	15	55	30	30	30	35
Security and Compliance	26.82	8.46	10	40	25	25	25	20
Software Programming and Web Design	13.18	8.08	0	30	15	15	15	10

Note. $n = 33$.

There was considerable variability in the weight estimates, likely reflecting the different occupational perspectives of the SMEs. We rounded the mean weights to increments of five percentage points and used weights of 30%, 30%, 25%, and 15% for NT, CO, SC, and SPWD, respectively. These category weights are the same as the weights derived in 2018 and 2015 and vary only slightly from the 2012 category weights. This likely reflects an anchoring effect since SMEs were informed of all previous weights before completing the exercise. However, this anchoring is desirable to maintain continuity in the test blueprint while still allowing for updates due to recent changes in the Cyber domain. In other words, it is beneficial for SMEs to consider the existing blueprint as a reference point so that any changes they recommend are deliberate, rather than a product of random variation in rating tendencies from one SME group to the next. Outcomes of the blueprint review (both the KSA review and weighting exercise) provided specifications to guide the item writing process, which is described in the next section.

6.2 Item Writing

The target number of new items to draft was 200. Based on that total, we established a target number of items for each KSA category within each content area. In addition to the blueprint category weights described above, the Air Force asked us to consider the test's ability to assess Computational Thinking to support compliance with a recent Congressional mandate (William M. [Mac] Thornberry National Defense Authorization Act for Fiscal Year 2021, 2020; for a review of the Computational Thinking construct, see Adams & Oppler, 2021). In discussions with the Air Force, we determined that two content areas within SPWD —namely, Understanding Numbering Systems and Knowledge of Basic Language Constructs—were particularly aligned with Computational Thinking.

To enhance the Cyber Test's coverage of the Computational Thinking construct, we adjusted the initial blueprint weights to emphasize these two content areas. Specifically, we increased the target items for each of these areas by five and decreased the target number of CO items by 10. This adjustment was also useful because our existing pool of operational and pilot tested items had more CO items and fewer Software Programming & Web Design items than suggested by the blueprint weights.[3] As a result, the adjusted item development targets should bring future Cyber Test item pools into closer alignment with the test blueprint. Table 14 provides the initial test blueprint weights, the adjusted weights, and the final item development targets corresponding to the adjusted weights.

[3] The same was true of our final eligibility pool, although this pool was not yet established at the time item development targets were set.

Table 14: Adjusted Blueprint Weights and Item Development Targets

Category	Initial Blueprint Weights	Adjusted Blueprint Weights	Target Number of Items
Networking & Telecommunications	30	30	60
Computer Operations	30	25	50
Security & Compliance	25	25	50
Software Programming & Web Design	15	20	40

Note. Blueprint weights are given as percentages of the total test.

We recruited civilian information technology experts to serve as item writers. Specifically, we contacted five experts with prior cyber-related experience in the Army. Four of the experts had developed item content for previous Cyber Test efforts, whereas one had not written test items before. All agreed to participate. Item writers signed a non-disclosure agreement, which described rules and procedures for saving and destroying items.

We provided item development training for all item writers during a three-hour teleconference. The training covered the purpose of the test, the demographics of the target population, important aspects of developing quality Cyber Test items, general best practices in test item development, instructions for drafting items in the HumRRO item banking platform, and the process for item review.

Item writing efforts focused on developing items of "easy" and "moderate" difficulty to address gaps in the existing item pool. SMEs often have trouble estimating item difficulty for a non-expert population (in this case, applicants for enlisted service) precisely because of their expertise. That is, what may be perceived as an easy item for a SME may in fact be quite difficult in the target population. Therefore, item writer training included a calibration session designed to orient SMEs to the target population. During this session, we provided multiple example items from each content category that were representative of low, moderate, and high levels of difficulty in the applicant population; these examples included items that the same item writers had developed under the previous contract. We emphasized that the new items should fall into the easy to moderate difficulty range. For each item drafted by the SMEs, they were asked to indicate the item's level of difficulty by estimating the proportion of entry-level test takers they believed would answer the item correctly.

6.3 Item Review

Once the new items were drafted, we conducted a multistage review process to assess the items' quality and make improvements as needed. The primary purpose of the item review was to confirm that the items were (a) content valid, (b) appropriate for the test's purpose, (c) appropriate for the target population, (d) current in their content, and (e) correctly keyed. Each of the newly-developed Cyber Test items underwent two levels of review – editorial and technical.

As previously mentioned, our goal was to develop 200 new items. Due to the possibility of dropping items during reviews, we developed additional items to ensure 200 items remained

following all reviews. Between February and August of 2022, the item writers developed 207 items and submitted them via HumRRO's item banking platform. An initial editorial review was conducted for grammar, reading level, appropriateness for the test and population, and adherence to HumRRO's Guidelines for Sensitivity and Bias Review (Waters, 2008). No items were found to be in violation of sensitivity guidelines regarding (a) offensive or exclusionary language, (b) stereotypes, or (c) ethnocentrism. There were many instances where syntax and vocabulary had to be simplified when the same concept could be conveyed without introducing unnecessary verbal load.

After the initial editorial review was completed, each item underwent a technical review performed by an item writer who was not the item's original author. Item writers were provided guidance on how to review items as part of the item writing training. When reviewing each other's items, item writers were asked to address the following questions and make specific suggestions:

- Is this item appropriate for its content area? Would it be better suited for another content area?
- Is the item based on trivial or obscure knowledge?
- Is the item in danger of becoming obsolete?
- Does any component of the item need to be revised? If so, how?
    - Is the stem valid?
    - Is the key correct?
    - Are the distractors plausible?
    - Are the distractors incorrect?
- Is the item appropriate for the target population (i.e., entry-level enlisted applicant)?

The items were then revised based on the results of the technical review. These revisions primarily concerned the preciseness and clarity of the stem and response options, correctness of the key, whether there was only one correct response, and plausibility of the distractors. Three items were found to belong to a different test content category than originally indicated. Two items were dropped during the technical review—one due to redundancy (i.e., different item writers developed items assessing the same content) and the other because it did not fit within the test blueprint.

If an item was edited during the technical review, we performed another editorial review on the revised item. If the technical edits were substantial, the item was reviewed again by another item writer. Finally, a HumRRO reviewer conducted a final editorial review of all items; comments and suggested edits were sent to one of the item writers for final edits. After all reviews were completed, a total of 205 items remained for pilot testing. See Table 15 for characteristics of the pilot-ready item pool.

Table 15: New Item Pool for Pilot Testing

Content Area	Target #	Cluster	Final Item Count	Mean Difficulty Rating
Networking & Telecommunications	60	Networking	30	2.9
		Telecommunications	30	2.9
Computer Operations	50	PC Configuration and Maintenance	30	3.1
		Using IT Tools/Software	21	2.8
Security & Compliance	50	Offensive Methods	26	3.1
		System Security	25	3.2
Software Programming & Web Design	40	Numbering Systems	13	2.5
		Software Programming	23	3
		Database Development and Administration	7	3.4
		Total Items / Grand Average Estimated Difficulty	205	3.0

Note. Mean Difficulty Rating is the average SME-rated difficulty level of items within clusters, rated on a 5-point scale: 1 = very easy (more than 75% will get correct); 2 = easy (about 75% will get correct); 3 = medium (about 50% will get correct); 4 = hard (about 25% will get correct); 5 = very hard (fewer than 25% will get correct).

6.4 Item Preparation

We formatted the 205 new items according to guidelines provided by DTAC for administration on the ASVAB platform. The intent is for the new items to be experimentally "seeded" within existing Cyber Test forms in a manner similar to that of experimental ASVAB items. Once sufficient experimental data are collected, these items can be calibrated and incorporated as operational items in a future iteration of the Cyber Test.

7.0 CONCLUSION

The Cyber Test is an increasingly important component of the pre-enlistment assessment tools available to the Armed Services. As the cyber knowledge domain rapidly evolves, it is important to refine and expand the existing Cyber Test forms to keep abreast of recent developments in the field. To this end, HumRRO calibrated and evaluated 215 experimental items, equated their item parameters to the existing operational item pool, and created a set of three parallel CAT forms for operational use. In addition, HumRRO conducted a thorough review and validation of the blueprint upon which the Cyber Test is based. Using this blueprint, we then developed 205 new items to expand the Cyber Test item pool and prepared these items for pilot testing. The next steps are to conduct a CAT simulation to estimate how the new forms will perform under operational conditions, collect pilot data on the newly developed items from Service applicants, and evaluate the pilot items' psychometric properties and functioning.

8.0 REFERENCES

Adams, K. A., & Oppler, S. H. (2021). *Computational thinking literature review*, Human Resources Research Organization, Alexandria, VA.

Camilli, G., & Shepard, L., (1994). *Methods for identifying biased test items*, Sage Publications, Thousand Oaks, CA.

Campbell, J. P., & Knapp, D. J.,(Eds.) (2001). *Exploring the limits in personnel selection and classification.* Lawrence Erlbaum Associates, Mahwah, NJ.

Defense Manpower Data Center (2008). *CAT-ASVAB forms 5 - 9* (Technical Bulletin No. 3). Seaside, CA.

Defense Manpower Data Center (2014). *ASVAB fairness information*, 2014, from Official site of the ASVAB website: http://officialasvab.com/fairness_res.htm

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in *R*. *Applied Psychological Measurement*, *35*, 398-409.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25), Human Resources Research Organization, Alexandria, VA.

Embretson, S. E., & Reise, S. P. (2000), *Item response theory for psychologists,* Lawrence Erlbaum Associates, Mahwah, NJ.

Garmire, E., & Pearson, G. (Eds.) (2006). *Tech tally: Approaches to assessing technology literacy.* National Academy of Engineering and National Research Council, The National Academy Press, Washington, DC.

Koch, A. J., Trippe, D. M., Beatty, A. S., & Shewach, O. R. (2018). *Cyber Test Development* (2017-088), Human Resources Research Organization, Arlington, VA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems,* Erlbaum, Hillsdale, NJ.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores,* Addison-Wesley, Reading, MA.

Ree, M. J., & Earles, J. A. (1992). *Subtest and composite validity of ASVAB forms 11, 12, and 13 for technical training courses* (AFHRL-TR-81-55), U.S. Air Force Human Resources Laboratory, Brooks AFB, TX.

Russell, T. L., & Sellman, W. S. (Eds.) (2009). *Development and pilot testing of an information and communications technology literacy test for military enlistees: Volume 1 final report* (FR 08-128), Human Resources Research Organization, Alexandria, VA.

Russell, T. L., & Sellman, W. S. (Eds.) (2010). *Information and communication technology literacy test training school validation: Phase II final report* (FR 09-89), Human Resources Research Organization, Alexandria, VA.

Stocking, M., & Lord, F. M. (1983), Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thissen, D. (2003). *MULTILOG* 7: *Multiple categorical item analysis and test scoring using item response theory* [computer program]. Scientific Software, Chicago, IL.

Trippe, D. M., Moriarty. K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology, 26,* 182-198.

Trippe, D. M., & Russell, T. L. (Eds.) (2011). *Information and communications technology literacy test norming study: Phase III final report* (AFCAPS-FR-2011-00xx), Air Force Personnel Center, Randolph AFB, TX.

van der Linden, W. J. (2005). *Linear models for optimal test design,* Springer, NY.

Waters, S. D. (2008). *Guidelines for item sensitivity and bias review,* Human Resources Research Organization, Alexandria, VA.

Welsh, J. R., Jr., Kucinkas, S. K., & Curran, L. T. (1990). *Armed Services Vocational Aptitude Battery (ASVAB): Integrative review of validity studies* (AFHRL-TR-90-22), U.S. Air Force Human Resources Laboratory, Brooks AFB, TX.

William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, 134 Stat. 3666 (2020). https://www.govinfo.gov/content/pkg/PLAW-116publ283/html/PLAW-116publ283.htm

Zwick, R., Thayer, D.T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analyses. *Journal of Educational Measurement, 36*, 1-28.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

%	Percent
Θ	Theta (IRT ability scale)
3PL	Three-Parameter Logistic
a_i	IRT item discrimination parameter
ASVAB	Armed Services Vocational Aptitude Battery
ATA	Automated Test Assembly
b_i	IRT item difficulty parameter
c_i	IRT pseudo-guessing parameter
B/W	Black/White
CAT	Computer Adaptive Test
CO	Computer Operations
DIF	Differential Item Functioning
DMDC	Defense Manpower Data Center
DTAC	Defense Testing Assessment Center
F/M	Female/Male
H/W	Hispanic White/Non-Hispanic White
HumRRO	Human Resources Research Organization
ICC	Item Characteristic Curve
ICTL	Information and Communications Technology Literacy
IIF	Item Information Function
IRT	Item Response Theory
KSA	Knowledge, skills, and abilities
M1	Transformational multiplier
M2	Additive constant
n	Sample size
NT	Networking and Telecommunications
SC	Security and Compliance
SME	Subject Matter Expert
SPWD	Software Programming and Web Design
TCC	Test Characteristic Curve
TIF	Test Information Function