AFRL-RH-WP-TR-2018-0120

CYBER TEST DEVELOPMENT

Amanda J. Koch
D. Matthew Trippe
Adam S. Beatty
Oren R. Shewach

Human Resources Research Organization

December 2018
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS 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.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and 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-2018-0120 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

_____//signature//_____
THOMAS R. CARRETTA
Work Unit Manager
Supervisory Control and Cognition Branch

__//signature//_____
JOSEPH P. NALEPKA
Chief, Supervisory Control and Cognition Branch
Warfighter Interface Division

____//signature//_____
LOUISE A. CARTER
Chief, Warfighter Interface Division
Airman Systems Directorate

This report is published in the interest of scientific and technical information exchange, 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)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
18-12-18	Interim	0 JAN 18 – 09 JAN 19

4. TITLE AND SUBTITLE
Cyber Test Development

5a. CONTRACT NUMBER
FA8650-14-D-6500, TO0007

5b. GRANT NUMBER
Not applicable

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Amanda J. Koch, D. Matthew Trippe, Adam S. Beatty, and Oren R. Shewach

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
H0SA (532909TC)

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
2018 No. 88

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
711th Human Performance Wing,
Airman Systems Directorate,
Warfighter Interface Division,
Supervisory Control & Cognition Branch
Wright-Patterson AFB, OH 45433
Air Force Materiel Command

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

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2018-0120

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
88ABW-2019-3572 cleared 5 August 2019

14. ABSTRACT

Since 2008, various 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 Services and has been shown to predict success in entry-level training in cyber-related military occupations. The goal of the current research project was to transition the existing static Cyber Test forms to a computer adaptive test (CAT) platform. Toward this end, 251 experimental items were pilot tested, calibrated, equated, screened, and added to the existing Cyber Test item pool. The items remaining in the pool were assembled into multiple parallel forms (or pools) from which the CAT algorithm will draw items. Additionally, the test blueprint was updated with the assistance of cyber subject matter experts across Services, and 215 new items were developed and are ready to be pilot tested.

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

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	29	19b. TELEPHONE NUMBER *(Include Area Code)* (937) 713-7143

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This report was prepared under Subcontract to Infoscitex (IST) (Subcontract NO. HIRT FPH02-S022, Prime Contract N0. FA8650-14-D-6500).

There are individuals not listed as authors who made important contributions to the work described in this report. We thank Chris Huber for his assistance with item reviewing and screening. We are extremely grateful to Suzanne Clark, Russell Fenton, 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 to update the Cyber Test blueprint. Finally, we thank our technical point of contact, Dr. Thomas R. Carretta, at the Air Force Research Laboratory for his guidance on this project.

INTRODUCTION AND BACKGROUND

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude test battery used by all U.S. military services for selection and classification of enlisted trainees. 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, Kucinkas, & Curran, 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, Embretson, Kyllonen, & Schmitt, 2006), which included 22 recommendations. One of the panel's recommendations was that research should be conducted to develop and evaluate a test of information and communications technology literacy (ICTL).[1] Many military jobs 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 test is now known as the Cyber Test. The Cyber Test was designed to predict success in entry-level training in cyber-related military occupations. Subsequent research on the Cyber Test confirmed that Cyber Test scores are predictive of cyber training success (Russell & Sellman, 2009, 2010; Trippe & Russell, 2011).

Transitioning the existing static Cyber Test forms to a computer adaptive test (CAT) platform is the next logical step in the evolution of this increasingly important test. The current item pool contains nearly 170 items in four sub-content areas (Networking and Telecommunications, Computer Operations, Security and Compliance, and Software Programming and Web Development). This item pool is roughly half the size of that for a typical ASVAB technical test. The existing item pool functions quite well with respect to measurement precision at the higher end of the ability distribution, where selection decisions are made. However, the relatively weak measurement precision available in the middle and low end of the distribution will create problems for a CAT application because (a) the middle of the ability distribution has the highest population density and (b) item selection algorithms choose items based on (among other factors) item information or measurement precision for a given ability level. Establishing a well-functioning CAT in this domain requires an item development effort focused on items that provide information on a larger number of applicants. HumRRO previously worked with the Air Force to develop 251 experimental Cyber Test items. These items have been pilot tested to Service applicants and require further review and screening to be suitable for operational use. This report documents the psychometric evaluation of these experimental items, the development of parallel two- and three-form CAT solutions (including these experimental items and the existing 170 items), and the development of 215 new items that are ready to be pilot tested.

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

ITEM ADMINISTRATION, CALIBRATION, AND EQUATING

Item Administration

The 251 experimental items (developed under a previous contract; see Koch, Trippe, Beatty, & Purl, 2017) were provided to the Defense Personnel Assessment Center (DPAC) for administration on the ASVAB platform. It was subsequently determined that nine of the items could not be administered on the ASVAB platform (e.g., images as response options), bringing the total number of items pilot tested to 242. Experimental items were "seeded" within existing Cyber Test forms in a manner similar to that of experimental ASVAB items. More specifically, 10 randomly selected experimental items from the item pool were administered to 86,623 Service applicants between August of 2016 and April of 2018. This kind of randomization effectively controls for many potential extraneous factors (e.g., order effects) encountered in traditional pilot testing.

An analysis sample, used to conduct all calibration and equating analyses, was created by limiting the full data set in several ways. First, we eliminated invalid records by identifying those with testing time 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. Finally, we further identified and removed corrupt or otherwise invalid records to include (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		
Army Guard	55	0.06
Army Regular	1,260	1.48
Army Reserve	49	0.06
Air Force Guard	6,004	7.06
Air Force Regular	39,812	46.84
Air Force Reserve	3,912	4.60
Marine Regular	43	0.05
Marine Reserve	4	0.00
Navy Regular	31,310	36.85
Navy Reserve	2,537	2.99
Coast Guard Regular	2	0.00
Gender		
Female	17,630	20.74
Male	45,635	53.70
Unknown/missing	21,723	25.56
Race		
American Indian	623	0.73
Asian	3,077	3.62
African American	13,262	15.60
Caucasian/white	41,802	49.19

(continued)

Table 1: Demographic Characteristics of the Calibration Sample (continued)

Characteristic	*n*	% of Sample
Race (continued)		
Hawaiian/Pacific	647	0.76
Other	2,421	2.85
Decline to Respond	1,461	1.72
Unknown/missing	21,695	25.53
Ethnicity		
Hispanic or Latino	10,123	11.91
Not Hispanic or Latino	53,091	62.47
Decline to Respond	79	0.09
Unknown/missing	21,695	25.53
Total	84,988	100.00

Note. Gender, race, and ethnicity were not included in Cyber Test data extracts and had to be subsequently merged with limited success, which resulted in a larger than normal proportion of missing demographic data. Due to rounding, sum may not be 100%.

Item Calibration and Equating

All Cyber Test items were analyzed 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 test item responses by examinees are the result of underlying levels of ability possessed by those individuals. IRT provides a seamless approach to a variety of test analysis, development, and reporting activities. 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. Calibration was executed via the software program MULTILOG (Thissen, 2003).

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

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i\left(\theta - b_i\right)\right]}$$

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 point of inflection on the ICC is higher or lower (i.e., to the right or to the left) along the ability (theta) scale. For example, the point of inflection of the curve for the sample item in Figure 1 is centered at zero, the mean on the ability scale. An efficient test will be composed of items with ICCs similar to that depicted, but with varying difficulties (“b” parameter) that discriminate along the entire ability scale, which is typically called “theta.” Item characteristic curves also differ in their lower asymptotes (related

to how easy it is to get the item correct by guessing, or the "c" parameter) and the gradient of their slopes at the inflection point (i.e., "a" parameter).

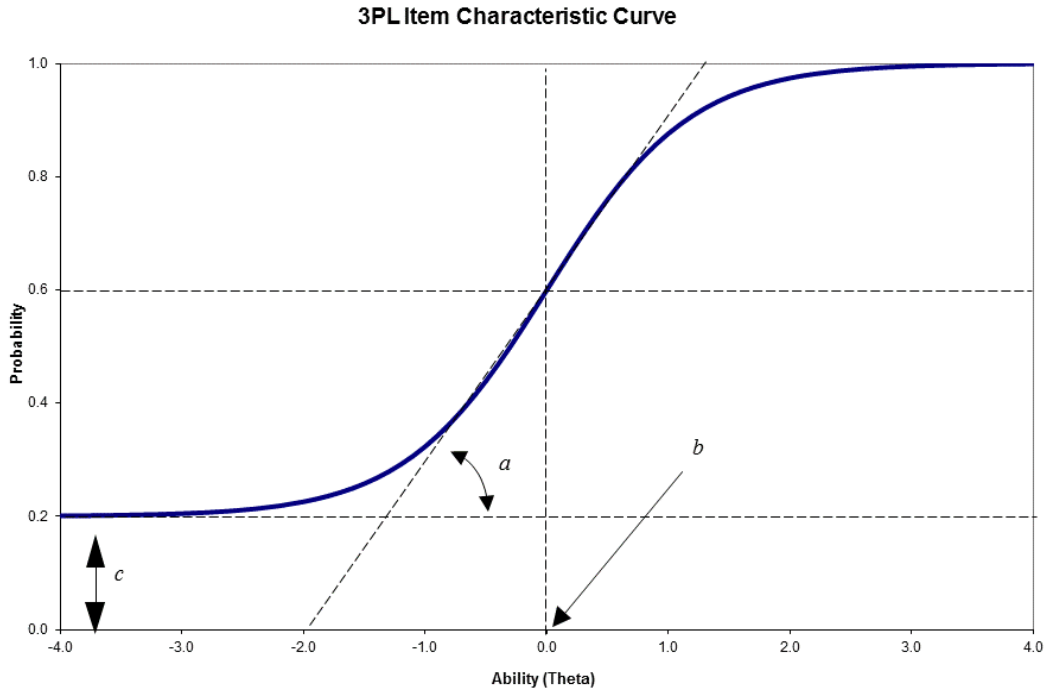

Figure 1: Example item characteristic curve in the 3-parameter logistic model.

Each individual Service applicant in the calibration sample was administered one of two 29-item operational Cyber Test forms (see Trippe & Russell, 2011) and 10 randomly seeded experimental items. Each of the 242 experimental items was administered to an average of 3,512 individuals 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 calibrated item parameters for the 58 operational items in the traditional Marginal Maximum Likelihood (MML) framework in which algorithms search for parameter values as well as ability values 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). Parameter estimates for the 242 experimental items were then calibrated in the fixed theta framework. The fixed theta framework has a few advantages related to the stability of the calibration. 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.

Item parameter estimates calibrated in the analyses just described were on a somewhat arbitrary scale that needed to be linked back to the original operational scale established in 2011 by an equating process. Item parameter estimates were previously equated in 2014 (Trippe, Moriarty,

Beatty, & Diaz, 2014), and those 2014 values were used as the "base" scale for equating in this project. The equating process involves using the operational items administered in both 2014 and in this effort as "anchor items." We applied the Stocking-Lord (1983) procedure to establish a common scale. The Stocking-Lord procedure uses item parameters from the current effort and the 2014 calibration to calculate test characteristic curves (TCCs) for each set of parameters. A transformation multiplier and additive constant (M1 and M2) are then calculated to transform the current TCC to match the original TCC as closely as possible.

Operational item parameters that served as anchor items in this procedure were evaluated for potential parameter drift from 2014. Anchor parameters were placed on a common scale. Then, values of the squared differences were calculated at 31 quadrature points (the same used in the Stocking/Lord procedure) and the mean of the 31 squared differences was computed for each item. Items were flagged if their mean squared difference (or mean d-square) exceeded the 90^{th} percentile as a relative quantitative indicator of parameter drift. We progressively removed items with the largest mean d-square values from the equating solution, one at a time, and observed how this affected mean d-squared 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. 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 has changed over time or formally generally unknown concept has become mainstream) independent of quantitative evidence of drift. We based the decision to remove an anchor item from the final equating solution on a combination of the content review results and quantitative evidence of parameter drift. We removed nine anchor items that were both identified as obsolete and demonstrated empirical evidence of drift. We removed two additional anchor items with exceptionally large mean d-square values based on quantitative evidence alone. The Stocking-Lord (1983) procedure was then implemented using a total of 47 operational items as anchors. The resulting constants (M1 = 0.90678, M2 = 0.014708) were then used to transform all parameters to the original operational scale.

TECHNICAL AND SENSITIVITY REVIEW

Post Hoc Sensitivity Review

A small subset of the 242 experimental items underwent a *post hoc* sensitivity review based on statistical evidence of differential item functioning (DIF). We conducted 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.

Table 2: Subgroup Comparisons in DIF 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

We computed an Empirical Bayes Mantel-Haenszel statistic (Zwick, Thayer, & Lewis, 1999) for each item and subgroup comparison. Table 3 summarizes the classification framework described in Zwick et al. (1999).

Table 3: Classification of DIF 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	$EB_{MH} < 0.0$

Table 4 contains the DIF analyses results summarized according to the DIF classification framework. We did not necessarily conclude that an item was biased based on statistical evidence alone. Differences in relative difficulty may, in fact, represent construct-relevant variance that is not necessarily bias. That is, there may be true differences in the construct across groups. An item or test cannot be said to be truly biased unless the source of the differential functioning is determined to be construct irrelevant. This requires logical analysis of the item or test content (Camilli & Shepard, 1994). We reviewed the three items identified as demonstrating category C DIF in a mixed-gender group. 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: DIF Results

Comparison	*n* A Items	*n* B+ Items	*n* C+ Items	*n* B- Items	*n* C- Items
F/M	235	1	0	3	3
B/W	238	0	0	4	0
H/W	242	0	0	0	0
Total	715	1	0	7	3

Post Hoc Item Quality Review

Our goal was to assemble CAT forms from sets of items developed over several project phases. Sets of items available include: (a) 58 original operational items developed in 2011, (b) 118 items developed in 2014, and (c) 242 experimental items developed in 2017. Current empirical item response data was available for the 58 operational items and 242 experimental items. No new data were available for the 118 items developed in 2014. Nevertheless, we also reviewed the content of items developed in 2014 for evidence of potential content obsolescence.

First, two psychometric subject matter experts (SMEs) with working knowledge of the content areas covered in the test blueprint independently reviewed the operational Cyber Test items and flagged any item suspected to be subject to potential obsolescence. Any item flagged by either of the psychometric SMEs was sent to an IT SME for further content review. The psychometric SMEs reviewed the IT SME's comments and decided to remove three potentially obsolete items from form assembly. Additionally, nine items had been previously identified as potentially obsolete (Trippe et al., 2014). The same process was followed for the 118 items developed in 2014, with two psychometric SMEs identifying potentially obsolete items, and the IT SME

providing further content reviews. Based on this process, 11 items from the 2014 item pool were removed from form assembly. In most cases for the 12 operational items and 11 items from 2014 that were excluded, the items were still relevant and functioning as intended but would likely become less effective over time or be subject to parameter drift because of references to outdated software versions (e.g., Windows Vista, XP) or technology concepts that were "common" at the time the item was written but are now less common (e.g., wired connection of peripheral devices). That is, CTT- and IRT-based indices of item quality suggested that most of these items were still of acceptable quality but concerns over the content were the primary driver of the decision to remove them from the pool.

Development of the 242 experimental items was described in detail in Koch et al. (2017). Great care was taken to ensure the quality of the content in the 242 experimental items. Nevertheless, item quality must also be evaluated in terms of psychometric indicators. First, psychometric SMEs independently reviewed the content of each experimental item in the context of available psychometric indicators of item quality, which included (a) the p-value or proportion of applicants who endorsed the keyed response, (b) the biserial item-total correlation, (c) the proportion of examinees endorsing each distractor response, (d) the distractor-total correlation, and (e) 3PL IRT item parameters. The psychometric SMEs, who also have working knowledge of the item content, independently rated each experimental item as an item to "keep" or "drop" from the final item pool or an item needing further review. If the SME indicated the item should be dropped or reviewed, s/he provided an explanation for the decision (e.g., content flaw, needs IT SME review, obsolete, psychometric issues). Content flaws included issues such as two possible correct answers, which are often revealed by positive distractor-total correlations or typographical errors. Items rated as needing technical review were often highly technical in nature and showed some ambiguous psychometric properties. A few items were flagged as needing further review and were discussed with an IT SME to confirm content quality. Items rated as obsolete were those that referred to content that had become dated since the item was written or were likely not to remain current in the foreseeable future. An IT SME also reviewed these items to confirm they were obsolete. Items rated as having psychometric issues demonstrated poor statistical evidence of item quality such as (a) a low or negative item-total correlation, (b) an extremely high or low p-value, (c) extreme or out of bounds IRT parameters, or (d) positive distractor correlation(s). The reasons for dropping items from the pool were not necessarily mutually exclusive. It is often the case that content flaws are reflected in psychometric indices. Undesirable psychometric characteristics may also simply indicate that an item is inappropriate for the applicant population and are not necessarily indicative of poor item content.

After the two psychometric SMEs rated each item independently, a third psychometric SME reviewed the decisions and weighed in on items for which the first two SMEs had disagreed. Feedback from an IT SME was obtained for items that had potential content or obsolescence issues. A fourth psychometric SME reviewed all of the decisions and comments (across the three psychometric SMEs and the IT SME) and made a final decision, consulting with two other psychometric SMEs in borderline cases (e.g., raters were split on the keep/drop decision) until a consensus was reached. After all discrepancies were resolved, SMEs agreed that 117 (48%) of the experimental items were acceptable for the next step of form assembly. The most frequent reason for dropping an item was psychometric issues; many of these items had extreme "b"

parameter values and reflected near-chance response patterns, suggesting that the items were too difficult for test takers. It may be that some of the content areas are inherently difficult for this applicant population, making it hard for item writers to develop easy to moderate items. That is, some content areas (e.g., software programming concepts) may be so unfamiliar to most applicants that even items assessing basic knowledge are high in difficulty. Items identified as having content flaws were often removed because of a distractor that could be plausibly correct.

FORM ASSEMBLY

After obtaining final equated parameters for all items and finalizing the set of items that would be dropped (due to psychometric or content issues), the next step was to attempt to develop parallel item pools, or forms, that the CAT ASVAB can use to draw items from.[2] The Air Force requested that we attempt to develop a two-form solution and a three-form solution. 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 measurement properties (e.g., average difficulty, information provided at a given theta level) will not differ across CAT forms. Although 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 a number of psychometric and content characteristics of 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" (i.e., items that address highly similar content) should be accounted for.

However, it is difficult to balance several objectives simultaneously (e.g., two forms manually constructed to have equal content and key distributions will likely differ dramatically on their difficulty and discrimination). Therefore, in order to determine the optimal assignment of items to forms to balance the competing test specifications, we used Automated Test Assembly (ATA; van der Linden, 2005). 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. Specifically, 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 constraints the set of possible solutions. In order 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 using PROC OPTMODEL. The objective function we minimized was an equally weighted sum of the distance between the test information functions (TIFs) and test characteristic curves (TCCs) of the forms. We also specified the number of desired items per form, content area, item key, and item enemy targets as the constraints on the solution set.

Even though the use of ATA is an extremely helpful aid for creating parallel forms, there is still some iteration and judgment involved in deciding on a solution. For instance, a solution that provides an exact content distribution match to the blueprint might result in forms that are reasonably close in terms of difficulty and discrimination, but allowing for a one-item deviation from optimal on content area might result in forms that are practically identical with respect to

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

difficulty and discrimination. Therefore, our approach was to generate many ATA solutions with different constraint settings in order to evaluate the trade-offs implied by different specifications. The general process for developing parallel forms was as follows:

1. ATA was conducted with optimal form specifications, such as using the maximum number of items (i.e., 133 items on each form for the two-form solution and 88 items on each form for the three-form solution), content distributions that matched the blueprint, and equal key distributions. This resulted in solutions that were not feasible. In other words, given the characteristics of the overall item pool, there were not solutions that matched the blueprint with equal key distributions.
2. ATA was then conducted with a series of deviations from the optimal form specifications. Examples include lowering the number of items included on each form and allowing key and content distributions to differ by one or more items from optimal.
3. HumRRO analysts came to a consensus on which solution to choose. This largely involved a judgment of the trade-off between adherence to optimal specifications and how parallel the forms were psychometrically.

The remainder of this section presents results of ATA for the two-form and three-form solutions.

Two-Form Solution

For the two-form solution, 130-item forms were chosen. In order to evaluate parallelism of the forms, we compared plots of the TIF and TCC curves and looked for overlap. To the extent that these curves appear to be roughly the same, they can be viewed as functionally parallel. Figure 2 presents the TIF and TCC curves for the final two-form solution. As can be seen from the plots, there is a high degree of overlap between the two forms in both plots. Virtually no separation is detectable in the TCC plot, and there is only a small degree of separation in the TIF plot, mostly in the extreme tails. The TIF is more likely to show separation due to its sensitivity to items at a given theta level with high discrimination values; to the extent that one form has items targeted towards a particular theta with a higher average discrimination than the other form, this will tend to show up in the TIFs. These differences tend to be smoothed out more in the TCC. Nonetheless, the two plots taken together provide strong evidence that the forms can be treated as functionally parallel.

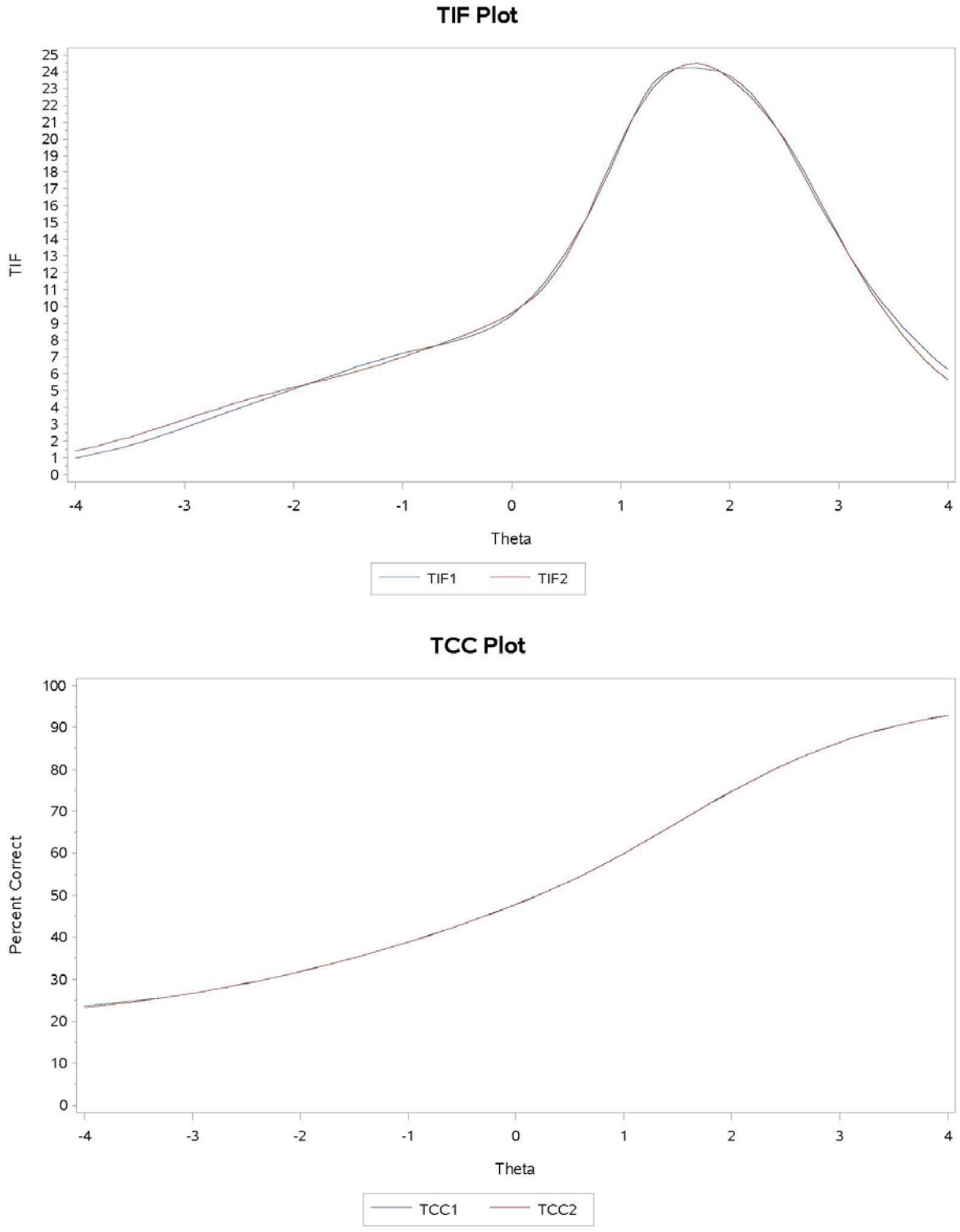

Figure 2: Overlaid test information functions (TIFs) and test characteristic curves (TCCs) for the final two-form solution.

Tables 5 and 6 present the content and key distribution results for the final two-form solution. The content distributions for each form are equal, and they are close to the blueprint values. The deviations from blueprint content are due to the characteristics of the eligible overall item pool. Specifically, there are 103 Computer Operations (CO)items, 69 Networking and Telecommunications (NT) items, 66 Security and Compliance (SC) items, and 28 Software Programming (SP) items (39%, 26%, 25%, and 11% of the overall pool, respectively[3]). As can be seen, achieving blueprint specifications with NT and SP items is not possible, especially because almost all of these items were used, and this difference was made up by the inclusion of more CO items. These are fairly minor differences, however. As for the key distribution, all are within 5% of an equal key distribution, which we viewed as close enough given the higher importance of including more items and psychometric characteristics for parallelism.

Table 5: Two-Form Parallel Solution – Content Distribution

	Target %	Form 1 - *n*	Form 1 - %	Form 2 - *n*	Form 2 - %
CO	30	49	37.66	49	37.66
NT	30	34	26.15	34	26.15
SC	25	33	25.38	33	25.38
SPWD	15	14	10.77	14	10.77

Note. CO = Computer Operations; NT = Networking & Telecommunications; SC = Security & Compliance; SPWD = Software Programming & Web Design. Due to rounding, sum may not be 100%.

Table 6: Two-Form Parallel Solution – Key Distribution

	% A	% B	% C	% D
Form 1	22.31	26.92	26.92	23.85
Form 2	23.08	26.15	20.77	30.00

Three-Form Solution

For the three-form solution, 87-item forms were chosen. Figure 3 presents the TIF and TCC curves for the final three-form solution. As with the two-form solution, there is a high degree of overlap with the TCC plot, and separation between the forms cannot be seen. The TIF plot is slightly less clean than with the two-form solution, largely due to spreading the items across an additional form. There are simply fewer options for balancing the test constraints while keeping the forms as close to parallel as possible. Even so, the largest differences are still at the extreme tails, and the TIFs are still quite similar as a whole.

[3] Sum is not 100% due to rounding.

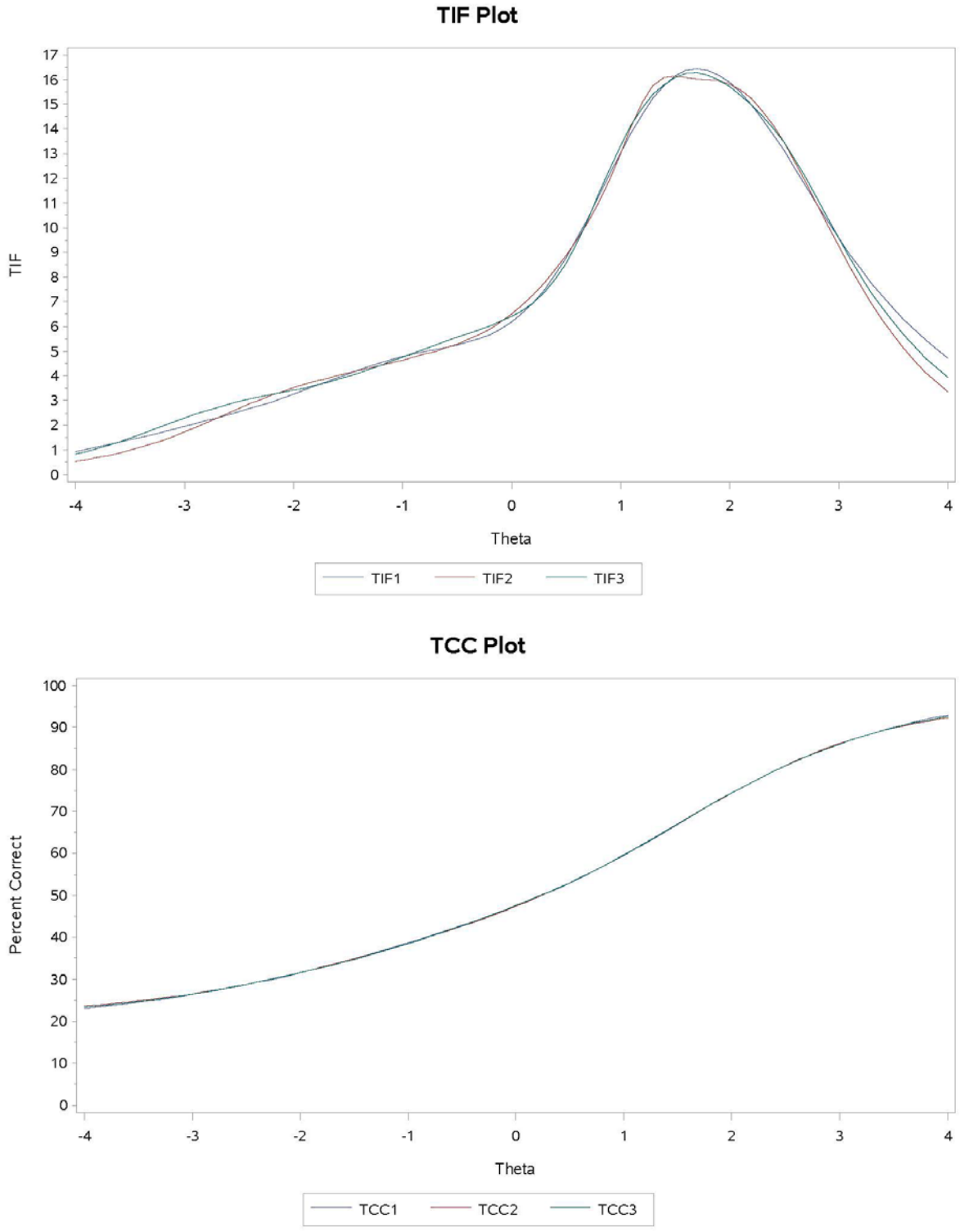

Figure 3: Overlaid test information functions (TIFs) and test characteristic curves (TCCs) for the final three-form solution.

Tables 7 and 8 present the content and key distribution results for the final three-form solution. Like the two-form solution, the three forms are fairly close to one another in terms of the content distributions and tend to be quite close to the blueprint values (though still exhibit the same small deviations from the blueprint due to the properties of the overall item pool). As with the two-form solution, we placed less emphasis on equalizing the key distributions and obtaining a perfect 25% in each category in order to focus more heavily on including as many items as possible and having the forms appear as parallel as possible, particularly given the increased difficulty in making the forms parallel with an additional form included. The percentage of keyed responses were generally still between 20% and 30%, with one outlier at 32%.

Table 7: Three-Form Parallel Solution – Content Distribution

	Target %	Form 1 - *n*	Form 1 - %	Form 2 - *n*	Form 2 - %	Form 3 - *n*	Form 3 - %
CO	30	33	37.93	32	36.78	34	39.08
NT	30	23	26.44	23	26.44	23	26.44
SC	25	22	25.29	22	25.29	21	24.14
SPWD	15	9	10.34	10	11.49	9	10.34

Note. CO = Computer Operations; NT = Networking & Telecommunications; SC = Security & Compliance; SPWD = Software Programming & Web Design.

Table 8: Three-Form Parallel Solution – Key Distribution

	% A	% B	% C	% D
Form 1	20.69	25.29	21.84	32.18
Form 2	20.69	27.59	25.29	26.44
Form 3	25.29	26.44	22.99	25.29

Note. Due to rounding, sum may not be 100%.

Form Assembly Summary

Solutions for two- and three-form solutions were developed using ATA that were as close as possible to parallel while also attempting to balance other test specifications such as content and key distributions. This was generally successful, with almost all of the available items being used in both solutions, TIF and TCC plots that appeared largely parallel, and content and key distributions that were quite close to test specifications. These functionally parallel forms are ready for the next step of assessing how well they perform under a CAT simulation.

NEW ITEM DEVELOPMENT

There is a need to continue to refresh the Cyber Test item pool (e.g., due to obsolescence, item exposure). The final task in this project involved updating the test blueprint and developing a set of new items that are ready to be pilot tested.

Blueprint Validation

The test blueprint upon which the Cyber Test is based was originally developed in 2008 (see Russell & Sellman, 2009) and updated in 2012 (see Trippe et al., 2014) and again in 2015 (see Koch et al., 2017). The blueprint is organized hierarchically, with four broad content areas at the highest level. 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, skill, and ability (KSA) statements that serve as the basis for item development. Because of the potentially mercurial nature of some content within the scope of the blueprint, it is wise to review the existing blueprint and KSA statements for both obsolescence and job relevance.

With assistance from the Air Force, HumRRO convened a military SME panel to conduct a blueprint review and validation workshop. The purpose of this review was to determine relevance of the previous blueprint to contemporary entry-level training for cyber-related occupations. We held a Joint-Service teleconference to (a) introduce the Cyber Test research program and (b) solicit input on the blueprint from SMEs. Forty-eight NCO SMEs in cyber related occupations from the Air Force (8), Navy (10), Army (20), Marine Corps (8), Coast Guard (1), and National Security Agency (1) were invited to participate in the teleconference and provide input to the blueprint via a follow up survey. We received 23 responses to the blueprint validation survey. The responding SMEs were affiliated with the Air Force (8), Navy (5), Army (7), and Marine Corps (3). SMEs reported an average of 6.70 years of experience in cyber-related occupations, with a range between 2 and 13 years of experience. SMEs were provided the broad content areas, sub-content areas, and 49 KSA statements from the most recent Cyber Test blueprint. Additionally, we asked SMEs to rate 61 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, SMEs were asked 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 required 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 9 displays the 10 KSA statements with the highest importance ratings. Overall, the most important KSA statements tend to capture fundamental concepts that often serve as the basis for higher level learning within one or more lower-level, specific content categories.

Table 9: Most Important KSAs

Category	KSA Statement	*M*	*SD*
NT	Knowledge of common network terminology.	4.17	1.09
NT	Knowledge of the purpose, capabilities, and functions of network hardware (e.g., routers, switches, hubs, bridges, servers, transmission media).	3.96	1.04
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).	3.70	1.40
NT	Knowledge of common network tools (e.g., ping, traceroute, nslookup) and interpretation of the results.	3.65	1.24
CO	Knowledge of basic computer concepts (bit, byte, CPU).	3.65	1.24
SC	Knowledge of Internet, website, and email vulnerabilities.	3.65	1.27

(continued)

Table 9: Most Important KSAs (continued)

Category	KSA Statement	*M*	*SD*
SC	Knowledge of system and application security threats, risks, and vulnerabilities.[a]	3.64	1.11
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]	3.57	1.31
SC	Knowledge of network security features (e.g., firewalls).	3.57	1.10
SPWD	Knowledge of Windows command line (e.g., ipconfig, netstat, dir, nbtstat).	3.57	1.25

Note. NT = Networking and Telecommunications; CO = Computer Operations; SC = Security and Compliance; SPWD = Software Programming and Web Design. 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, the SMEs were asked to provide “Needed at Entry" ratings for each KSA statement. Specifically, SMEs were instructed to indicate whether each KSA should be acquired prior to enlistment (needed at entry) or following enlistment (not needed at entry). Table 10 shows the 10 KSA statements that received the highest needed at entry ratings. Like the importance ratings, this list includes KSAs that are basic, simple concepts that provide the building blocks for more advanced skills.

Table 10: KSAs Receiving Highest Needed at Entry Ratings

Category	KSA	% Indicating Needed at Entry
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]).	78
CO	Knowledge of word processing software (e.g., Microsoft Word, OpenOffice Writer).	78
CO	Ability to connect PC hardware components (e.g., monitor, printer).	70
NT	Knowledge of common network terminology.	65
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).	65
CO	Knowledge of spreadsheet software (e.g., Microsoft Excel, OpenOffice Calc).	65
CO	Knowledge of basic computer concepts (bit, byte, CPU).	61

(continued)

Table 10: KSAs Receiving Highest Needed at Entry Ratings (continued)

Category	KSA	% Indicating Needed at Entry
CO	Knowledge of presentation software (e.g., Microsoft Powerpoint, OpenOffice Impress).	61
NT	Knowledge of the purpose, capabilities, and functions of network hardware (e.g., routers, switches, hubs, bridges, servers, transmission media).	57
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).	57

Note. NT = Networking and Telecommunications; CO = Computer Operations; SC = Security and Compliance; SPWD = Software Programming and Web Design. 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 asked to estimate the rate of obsolescence for each KSA statement using the scale shown in Table 11. Higher scores indicate a slower rate of obsolescence.

Table 11: 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 12 displays the 10 KSA statements that received the highest obsolescence ratings. Overall, the KSAs rated as the most stable tend to capture fundamental concepts (e.g., common terminology, tools, commands, or components).

Table 12: KSAs Receiving Highest Obsolescence Ratings

Category	KSA Statement	*M*	*SD*
CO	Knowledge of basic computer concepts (bit, byte, CPU).	4.70	0.62
SPWD	Understanding of different numbering systems such as hex and binary	4.70	0.75
NT	Knowledge of common network terminology.	4.61	0.77
NT	Knowledge of common network tools (e.g., ping, traceroute, nslookup) and interpretation of the results.	4.61	0.57
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]).	4.57	0.66

(continued)

Table 12: KSAs Receiving Highest Obsolescence Ratings (continued)

Category	KSA Statement	*M*	*SD*
NT	Knowledge of the purpose, capabilities, and functions of network hardware (e.g., routers, switches, hubs, bridges, servers, transmission media).	4.52	0.83
SPWD	Knowledge of Unix command line (e.g., mkdir, mv, ls, passwd, grep).	4.52	0.88
SPWD	Knowledge of basic language constructs (e.g., arrays, do-loops, if/then statements).	4.52	0.88
SPWD	Knowledge of computer programming principles such as object-oriented design.[a]	4.50	0.94
SPWD	Knowledge of Windows command line (e.g., ipconfig, netstat, dir, nbtstat).	4.48	0.77

Note. NT = Networking and Telecommunications; CO = Computer Operations; SC = Security and Compliance; 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.

SME ratings of the importance, stability, and "needed at entry" status of each KSA statement were used to exclude non-essential KSAs and identify the final list of KSAs for the blueprint. We chose to err on the side of retaining KSAs that appeared on the most recent blueprint. That is, previous SME groups identified these KSAs as being important and needed at entry, and we did not want to place too much weight on the judgments of the current, small sample; therefore, we sought relatively strong evidence to justify the exclusion of these KSAs. KSAs appearing on the most recent Cyber Test blueprint were retained if they met the following thresholds:

- More than 30% of respondents agreed that the KSA is needed at entry,
- The mean importance rating was greater than or equal to 2 (where 2 = a little important), and
- The mean obsolescence rating was greater than or equal to 1.5 (where 1 = likely to change in 2 years or less and 2 = likely to change in 2 to 5 years).

Higher standards were used to determine which KSAs from other sources to retain because these KSAs had not been identified as critical by previous SME groups. We planned to retain KSAs that were not on the most recent cyber blueprint if:

- More than 50% of respondents agreed that the KSA is 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, so no new KSAs were added to the blueprint. The final blueprint consisted of 41 KSAs, all from the previous cyber blueprint.

SMEs were also offered the opportunity to add important content areas or KSAs that were not included in the KSA list. SMEs suggested 11 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., critical thinking, data analysis), (b) the KSA was subsumed by other KSAs on the list, or (c) the KSA was determined to be too advanced for the test taker population (e.g., computer forensics and image analysis). Table 13 summarizes the final blueprint.

Table 13: Final Cyber Test Blueprint

Broad/Sub-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. $N = 23$.

SMEs also were asked to participate in a weighting exercise to determine the proportion of test items that should be devoted to each content area. They were asked 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 are presented in Table 14.

Table 14: Category Weights

Category	*M*	*SD*	Min	Max	Final Weight	2015 Weight	2012 Weight
Networking and Telecommunications	30.65	11.99	0	65	30	30	35
Computer Operations	28.91	9.65	5	50	30	30	35
Security and Compliance	25.00	9.05	0	40	25	25	20
Software Programming and Web Design	15.43	17.51	0	90	15	15	10

Note. $N = 23$.

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 Networking and Telecommunications, Computer Operations, Security and Compliance, and Software Programming and Web Design, respectively. These category weights are the same as the weights derived in 2015 vary only slightly from 2012 category weights. 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.

Item Development

We recruited civilian information technology (IT) experts to serve as item writers. We contacted four experts with cyber-related experience in the Army who had developed item content for previous cyber tests, and all agreed to participate. Item writers signed a non-disclosure agreement, which set out rules for saving and destroying items. One HumRRO researcher also participated as an item writer and focused on assessing more basic KSAs (e.g., related to basic networking, spreadsheet software, file extensions)

Cyber Test item developers underwent training in item development in May 2018, in a 1-hour teleconference that used slides and handouts. Due to their prior experience, all item writers were familiar with the process, so the training focused on reviewing several important aspects to developing quality items, including the purpose of the test, the demographics of the target population, and best practices in test item development.

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 difficulty in a non-expert population (i.e., 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. We provided multiple example items from each content category that are representative of low, moderate, and high levels of difficult in the applicant population. The example items included items that these item writers developed under the previous contract (i.e., the experimental item pool of 251). We stressed that the new items should fall into the easy to moderate difficulty range.

Item Review

Even with such training, item review is still necessary to help mitigate the effects of construct-irrelevant factors on test reliability and validity. Item review is typically an iterative process involving many steps and people. The primary purpose of the item review is to confirm that the items are (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.

The goal was to develop 200 new items. Due to the possibility of dropping items during reviews, we developed additional items to make sure 200 items remained after all reviews. The item

writers developed 213 items (distributed according to the content area weights reported in Table 14) and submitted them via email for an editorial review (by a HumRRO researcher) from May through August 2018. Those with significant edits or comments were returned to the author for revision; a second editorial review was performed after edits were made. The editorial review was primarily concerned with grammar, reading level, appropriateness for the test and population, and adherence to HumRRO's Guidelines for Sensitivity and Bias Review (Waters, 2008). One item was found to belong to a different test content category than those originally indicated. 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 notion could be conveyed without introducing unnecessary verbal load.

After the editorial review was completed, each item underwent a technical review performed by an item writer who was not the item's author. Part of the item writing training included how to review test items. When reviewing each other's items, the 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 and whether there was only one correct response, and plausibility of the distractors. Two items were dropped during the technical review process due to redundancy (i.e., different item writers developed items assessing the same content). Additionally, four items from the existing and experimental item pools were revised (with the same editorial and technical review processes undergone by new items) and included in the new item pool.

Another editorial review was performed on the items following the technical review if edits were made. If substantial edits were made during the technical review, the item was sent to an additional reviewer (one who had not written or reviewed the item) for another technical review. Finally, another HumRRO reviewer conducted a final editorial review of all of the items; comments and suggested edits were sent to one of the item writers for final edits. This resulted in a total of 215 items remaining for pilot testing (see Table 15).

Table 15: New Item Pool

Content Area	Number of New Items
Networking and Telecommunications	64
Computer Operations	64
Security and Compliance	53
Software Programming and Web Design	34

Item Preparation

The 215 new items were formatted according to guidelines provided by DPAC for administration on the ASVAB platform. The intent is for the new items to be "seeded" within existing Cyber Test forms in a manner similar to that of experimental ASVAB items.

SUMMARY AND CONCLUSION

The Cyber Test is a valuable component of the pre-enlistment assessment tools available to the Services. Transitioning the existing static Cyber Test forms to an adaptive test is an important step in the evolution of this increasingly important test. Toward this end, HumRRO calibrated and evaluated 251 experimental items, equated the items to prior item pools, and created parallel sets of two and three forms of items to be used in a CAT. Additionally, HumRRO conducted a thorough review and validation of the blueprint upon which the Cyber Test is based. We then developed 215 new items to expand the Cyber Test item pool and prepared the items for pilot testing. The next steps are to conduct a CAT simulation to estimate how well the forms will perform under operational conditions, implement the CAT, collect pilot data on the new items from Service applicants, and then evaluate the items' psychometric properties and functioning.

REFERENCES

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 (2009). *CAT-ASVAB forms 5 - 9* (Technical Bulletin No. 3), Author, Seaside, CA.

Defense Manpower Data Center (2014). *ASVAB fairness information*, 2014, from Official site of the ASVAB website: http://officialasvab.com/fairness_res.htm Retrieved 1 June.

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., & Purl, J. (2017). *U. S. Air Force enlisted selection and classification (S&C) research: CAT Cyber Test development final report (2017-022),* 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. (1982), 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., & 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.

Trippe, D. M., Moriarty, K. O., Beatty, A. S., & Diaz, T. E. (2014). *Cyber test form development and follow-on cyber applications* (FR 2014-041), Human Resources Research Organization, Alexandria, VA.

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.

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