Psychometric Evaluation of the Cybersecurity Curriculum Assessment

Geoffrey L. Herman
Shan Huang
glherman,sh69@illinois.edu
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA

Peter A. Peterson
pahp@d.umn.edu
University of Minnesota, Duluth
Duluth, Minnesota, USA

Linda Oliva
Enis Golaszewski
Alan T. Sherman
oliva,golaszewski,sherman@umbc.edu
University of Maryland, Baltimore County (UMBC)
Baltimore, Maryland, USA

ABSTRACT

We present a psychometric evaluation of the *Cybersecurity Curriculum Assessment (CCA)*, completed by 193 students from seven colleges and universities. The CCA builds on our prior work developing and validating a *Cybersecurity Concept Inventory (CCI)*, which measures students' conceptual understanding of cybersecurity after a first course in the area. The CCA deepens the conceptual complexity and technical depth expectations, assessing conceptual knowledge of students who had completed multiple courses in cybersecurity. We review our development of the CCA and present our evaluation of the instrument using Classical Test Theory and Item-Response Theory. The CCA is a difficult assessment, providing reliable measurements of student knowledge and deeper information about high-performing students.

CCS CONCEPTS

• **Applied computing → Education**; **Collaborative learning**; **Distance learning**; **Computer-assisted instruction**.

KEYWORDS

Classical Test Theory, concept inventories, Cybersecurity Assessment Tools (CATS), Cybersecurity Curriculum Assessment, cybersecurity education, Item Response Theory, psychometrics

ACM Reference Format:
Geoffrey L. Herman, Shan Huang, Peter A. Peterson, Linda Oliva, Enis Golaszewski, and Alan T. Sherman. 2023. Psychometric Evaluation of the Cybersecurity Curriculum Assessment. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2023), March 15–18, 2023, Toronto, ON, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3545945.3569762

1 INTRODUCTION

We face a dangerous shortage of cybersecurity professionals who possess the breadth of skills and knowledge needed to create and maintain safe and secure computing systems [20, 25]. Without these professionals, businesses, governments, and many other organizations are left vulnerable to the host of nefarious actors who seek to deny access to these systems, steal information, or spread disinformation [37]. We cannot address this shortage of cybersecurity professionals unless we have an accurate understanding of which cybersecurity curricula and teaching strategies are providing students with a strong foundation. The *Cybersecurity Assessment Tools (CATs)* project addresses this need by developing instruments that can robustly measure how well courses and curricula are forming deep conceptual knowledge in students [27, 30, 32, 35, 36, 38, 39, 41].

The CATs project first developed and validated the *Cybersecurity Concept Inventory (CCI)* to assess students' conceptual knowledge of cybersecurity after a first course in the area [30]. Building on the CCI, we created the *Cybersecurity Curriculum Assessment (CCA)* to assess students' conceptual knowledge of cybersecurity after they had completed a multi-course curriculum in cybersecurity. The CCA covers the same core topics as the CCI, and deepens the conceptual complexity and the technical details in the questions. We created these assessments with the goal of infrastructure for future research that can help us better understand the benefits and drawbacks of different teaching methods (CCI) and curricular structures (CCA) for teaching cybersecurity. These are the first validated conceptual assessments tools for cybersecurity [30].

After briefly discussing the foundational work for developing the CCA, we first present evidence for the validity of the CCA for its stated purpose based on reviews from a panel of experts. We then describe our work administering the CCA to a diverse population of students and provide evidence for the reliability and other desirable statistical properties of the CCA. We provide recommendations for how to use the CCA and thoughts on avenues for continuing to improve the CCA based on the psychometric analysis.

2 BACKGROUND

We provide a brief history of assessment instruments in *computer science (CS)* education. We then explain how we created the CCA.

2.1 Concept Inventories

A *concept inventory (CI)* is a validated, criterion-referenced assessment for a given set of topics that enables researchers and instructors to gauge what their students have learned about a given subject. One of the first CIs, the Force Concept Inventory, is credited with helping to realize the active learning revolution in introductory

SIGCSE '23, March 15–18, 2023, Toronto, ON, Canada.

ACM ISBN 978-1-4503-9431-4/23/03…$15.00
https://doi.org/10.1145/3545945.3569762

physics by creating a meaningful way to compare results of different pedagogical techniques [13, 16].

Over the last ten years, computing education researchers have been creating CIs, so our discipline can also benefit from them. Examples include the *Digital Logic Concept Inventory (DLCI)* [15], the *Multilanguage Assessment of CS1 Knowledge (SCS1)* [28, 40], the *Basic Data Structures Inventory (BDSI)* [29], and the CCI. For a more extensive review of assessment instruments used in computing education research, see [11, 21].

Despite their recent creation, CIs have been used in a variety of ways in computing education research, including but not limited to: examining the relationship between spatial ability and learning programming [4], comparing outcomes between digital logic courses using differing pedagogical approaches [14], evaluating novel instructional practices in CS1 [23, 45], evaluating the effectiveness of teaching students using block- and text-based programming languages [3], and understanding the impact of students' educational background on learning topics in CS [2].

2.2 Cybersecurity Assessment Exams

There are many existing cybersecurity assessments. For example, several certification exams, including ones listed by NICCS as relevant [10]. CASP+ [6] comprises multiple-choice and performance tasks items including enterprise security, risk management, and incident response. OSCP [33] (offensive security) is a 24-hour practical test focusing on penetration testing. Other exams include CISSP, Security+, and CEH [5, 7, 42], which are mostly informational, not conceptual. Global Information Assurance Certification (GIAC) [8] offers a variety of vendor-neutral *multiple-choice question (MCQ)* certification exams linked to SANS courses; for each exam type, the gold level requires a research paper. None of these have been rigorously validated, necessitating the development of assessments that can be used as research instruments for pedagogical research.

Additionally, the ACM, IEEE, and ABET have been working on curricular guidance for cybersecurity [12, 18], and the NICE Cybersecurity Workforce Framework [24] establishes a common lexicon for explaining a structured description of professional cybersecurity positions in the workforce with detailed documentation of the knowledge, skills, and abilities needed for various types of cybersecurity activities. We use these resources to inform the definitions and terminology we use in the CCI and CCA. For more details, see [36].

2.3 The CATS Project and the CCA

In cybersecurity there often is not a clear right or wrong answer to a given problem. Cybersecurity professionals must think deeply about real world scenarios and differentiate poor, mediocre, or ideal solutions to a given security problem. We designed the 25 MCQs on the CCI and CCA (labelled Q1–Q25 and henceforth called "items" or "test items") to encourage effective application of conceptual understanding. The CCI and CCA present a series of scenarios and questions about these scenarios that force students to weigh their options and select the best solution choice to the security problem.

Figure 1 gives a typical example test item from the CCA: All alternatives are true, but only one is the best answer. The best answer is Alternative B, which highlights the core vulnerability: Alice could inadvertently authorize a fraudulent transaction. Most students chose Alternative C, but knowing who pushed the button would not mitigate the core vulnerability. Also, the bank requires each customer to maintain control of their device.

Scenario. To guard against potential man-in-the-middle attacks on a customer's home computer, a bank requires all remote (i.e., not at the physical bank) transactions to be authenticated by a trusted physically-secure physical device issued by the bank. The device has no clock. The bank verifies a transaction by requesting that the customer transmit the proposed transaction together with a signed token output from the device. Each token includes a unique sequence number. To output the token, the customer inserts the device into their home computer and pushes a physical button on the device. The device cryptographically signs the token using a unique secret key physically secured on the device, and outputs the signed token. The bank requires each customer to maintain possession of their device.
Alice logs into the bank's website and fills out a form to transfer $2000 from Account 1 to Account 2. When prompted, she pushes the button on her device to authorize the transaction.
Question. Choose the most significant security limitation of the device in this context: The device...

A. cannot produce a timestamp.
B. lacks a display to show Alice the details of the transaction being authorized.
C. cannot verify who pushed the button.
D. communicates with Alice's home computer through an unencrypted channel.
E. signs the token with its own secret key, not with Alice's secret key.

Figure 1: CCA Question 22 probes the concept "Identify vulnerabilities and failures."

See [27, 36] for more details about our development process for the CCI and CCA. By engaging a panel of 33 cybersecurity experts through a Delphi process, we identified five core cybersecurity topics to guide the development of CCI/CCA questions [36, 39].

(1) Identify vulnerabilities and failures
(2) Identify attacks against CIA (Confidentiality, Integrity, Availability) triad and authentication
(3) Devise a defense
(4) Identify the security goals
(5) Identify potential targets and attackers

We then constructed a series of cybersecurity scenarios (e.g., Figure 1) that we used in open-ended interviews with student to identify misconceptions [41]. We used these misconceptions to guide our construction of compelling distractors for MCQs. We wrote five test items per topic [36, 41].

Building on this prior foundation of best practices for creating assessments [31], we continue this process following recommended best practices [19] for evaluating the statistical properties of a conceptual assessment using psychometrics. We seek to answer the following research questions:

RQ1: Do other cybersecurity experts agree that the CCA assesses knowledge that students should know after completing multiple courses in cybersecurity?
RQ2: What does the statistical evidence say about the

reliability and validity of the CCA?
RQ3: What levels of cybersecurity knowledge does the CCA measure well?
RQ4: How do the statistical properties of the CCA compare with those of other concept inventories in use?

3 RQ1: EXPERT REVIEW METHODS AND RESULTS

In spring 2021, we asked experts, including from our Delphi study, to give their feedback on the quality of the CCA test items. 20 of the experts started the review process; 11 completed it. We include only the experts who completed the CCA in our analysis.

The panel of 11 cybersecurity experts comprised experts from academia (6) and government agencies (5). Three of the experts from government agencies also had affiliations with universities. We chose experts from academia to represent what faculty thought students should know based on their curricula. We included experts from industry and government agencies to represent experts with experience with what knowledge new gradates would need to be successful on the job. All experts had at least a master's degree focused on cybersecurity, with 10 having a PhD. All experts had at least five years of experience in the field, with nine having more than 20 years of experience.

We asked experts to complete the CCA and rate all items on a 4-point scale: "Accept as is," "Accept with minor revisions," "Accept with major revisions," and "Reject." We asked experts to provide open-ended feedback on each item, especially if they marked anything other than "Accept as is." The experts spent between 30 minutes and seven hours providing feedback with a mean time of two hours and 45 minutes. All but one expert spent at least an hour, indicating a high level of engagement from the panel members.

Eight of the 11 experts answered every test item correctly. One expert had one mistake; one expert had three mistakes; and one expert had four mistakes. For each item, at least 10 of the 11 experts answered them correctly. Despite the nuanced nature of the items, we saw a strong consensus for the correct answer. The high percentage of correct answers indicates that the items generally provided enough information to be understood by content domain experts.

Experts unanimously agreed that 22 of the 25 items should be included in the CCA. For the remaining three items, 10 of the 11 experts indicated that those items should be included in the CCA. The dissenting opinions on these items typically pointed out nuances in the test item or assumptions that should be clarified. Based on feedback from the experts, we revised and improved all CCA test items prior to administering the CCA to students.

4 PSYCHOMETRIC METHODS

We discuss how we administered the CCA to students and analyzed the results using Classical Test Theory and Item-Response Theory.

4.1 Data Collection

We pursued multiple avenues for recruiting subjects to take the assessment, including emailing professors who do research in cybersecurity, asking colleagues, and contacting institutions involved with cybersecurity education programs such as Scholarship for Service [9] and institutions qualifying as Centers for Academic Excellence in Cyber Defense (CAEs) [17].

We hosted the CCA on PrairieLearn, an online, open source homework and exam platform, to facilitate the administration of the assessment to students at a range of institutions [43].

Most instructors offered some extra credit to complete the CCA. We collected data primarily at the end of two consecutive academic school years: April–June 2021 and April–June 2022. The institutional review board at UMBC approved our protocol.

A total of 278 students started the CCA in PrairieLearn. We discarded test instances whenever the student answered less than 24 of the 25 questions or spent less than 15 minutes on the assessment (it takes about 15 minutes just to read the assessment items). The resulting valid data set had 193 responses from seven colleges and universities. 130 students (68%) identified as male, 39 (20%) identified as female, 3 identified as non-binary (1.5%), and 21 declined to provide gender information. 159 students (83%) identified as being between the ages of 18–25.

Our participants came from a range of institutions including large and small public universities as well as military academies. Institutions were geographically diverse within the United States, though most were from the East Coast. We collected data from research-focused and teaching-focused institutions. The majority of students came from large public research universities and the military academies, as they were often able to provide more subjects for testing. Most students who took the CCA were CS majors.

4.2 IRT and CTT

Classical Test Theory (CTT) and *Item Response Theory (IRT)* are commonly used analytical frameworks for showing statistical support for the validity of assessment instruments and for gauging the skill of students taking an assessment [19]. CTT and IRT measure each item's *difficulty* — how hard it is to answer a question correctly — and *discrimination* — how well an item differentiates between students of different ability levels. While both frameworks use the same terms, they define them differently and provide complementary perspectives on the quality of an assessment as a whole, the quality of items individually, and the abilities of the students.

For robust educational assessments, we want items to span difficulty levels to provide information about students across a range of ability levels [19, 31]. We also want items to have high discrimination so that we can measure student ability more precisely.

CTT is a more intuitive framework that produces results that are more limited to a particular context, whereas IRT is a more robust framework that creates more transferable information about an assessment by using logistic modeling.

4.3 Classical Test Theory

In CTT, a student's actual score, X, on an assessment (e.g., the number of questions they answered correctly) is the estimated ability of the student. A student's true score, T, is the theoretical measure of a student's ability level, which is the sum of the student's actual score and some measurement error E (i.e., $X = T + E$).

The *difficulty* of an item is the percentage of students who correctly answered that item. Difficulty values thus can range from 0 to 1. *Discrimination* is the point biserial correlation between a student's score on an item and their score on the test as a whole [15].

Discrimination can range from -1 (perfectly anti-correlated) to 1 (perfectly correlated). A discrimination of at least 0.2 is a suggested value for the minimum discrimination value for any item to be considered of sufficient quality for inclusion on an assessment [19].

4.3.1 Reliability. Cronbach's α is a measure of the internal consistency of an assessment based on the amount of correlation between scores on different items on the assessment. It ranges from 0 to 1, with a higher value indicating more internal consistency. There is no one universally accepted value of Cronbach's α to denote a reliable assessment, but commonly cited ranges of Cronbach's α for high-stakes assessments should be at least 0.7 or 0.8 [19, 26]. The internal reliability of an assessment can serve as a estimate of the unidimensionality of the assessment (i.e., do all items on the test measure roughly the same underlying construct, such as cybersecurity conceptual knowledge?).

We can evaluate the quality of items in a test by comparing Cronbach's α with and without a particular item included in the test [19]. If removing a test item from the test increases the reliability of the test, it may be measuring a different construct, indicating it may be an item to consider removing permanently.

4.4 Item Response Theory

We use IRT to gain greater insight into the properties of particular questions, and how much information individual questions and the test as a whole give about students of differing ability levels. Keeping in alignment with our prior work analysis with the CCI [30] and what is most commonly done with concept inventories more generally [19], we use the *two-parameter logistic (2PL)* model of IRT.

The 2PL model assumes that the probability of student n correctly responding to item i can be modeled as a function of the student's ability, θ_n, the discrimination of the item, a_i, and the difficulty of the item, b_i, as follows:

$$p_i(\theta_n) = \frac{1}{1 + e^{-a_i(\theta_n - b_i)}}. \quad (1)$$

4.4.1 Item Response Functions. Unlike CTT, a student's ability cannot be inferred directly from how many items they answer correctly, but rather it is estimated and normalized based on *which* items a student answered correctly or incorrectly and the relative difficulty and discrimination of those items. A student's ability θ_n can be interpreted as a parameter lying along a normal distribution with mean 0 and standard deviation 1.

The *difficulty* of an item determines the ability level at which a student has a 50% chance of answering that item correctly. The *discrimination* of the item determines the slope of the item response function when the difficult of the item and a student's ability are the same (see Section 4.4.1). Both parameters can take any positive or negative value.

Inserting the difficulty and discrimination parameters for each test item into Equation 1 gives the *item characteristic curves*, which help us visualize the difficulty and discrimination of test items, and the probability that a student with a given ability level will answer the question correctly. Figure 2 shows the item characteristic curves from the CCA to illustrate how to interpret these functions. Q13 has difficulty 3.14, and Q24 has difficulty 0.87. The dashed lines show that the Q13 and Q24 curves reach 50% probability at 3.14 and 0.87 on the Ability axis, respectively. Q13 has discrimination 0.46 and Q24 has discrimination 1.89. The curves show that Q13 increases gradually as ability increases because of its low discrimination while Q24 increases rapidly because of its high discrimination.

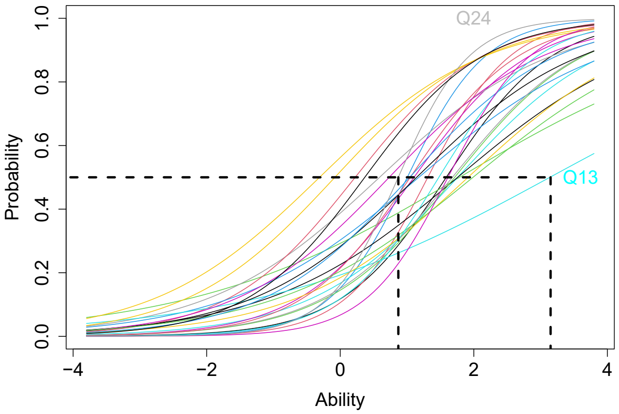

Figure 2: Item characteristic curves from the 2PL IRT model of the CCA. Q24 has high discrimination, 1.89 (steep slope), with moderate difficulty, 0.87. Q13 has low discrimination, 0.46 (shallow slope), with high difficulty, 3.14.

4.4.2 Item Information Functions. The *item information function* for an item is the derivative of the item response function for that item. It shows how much information that item gives about subjects taking the test. An item with higher discrimination will give more information about student knowledge and thus allow an assessment to measure student knowledge with less error. Summing the item information functions for all items on an instrument gives the *test information function* of the instrument.

Figure 3 graphs the item information functions from the CCA. Q24 has a high peak because of its high discrimination. Q13 has a proportionately lower peak because of its lower discrimination.

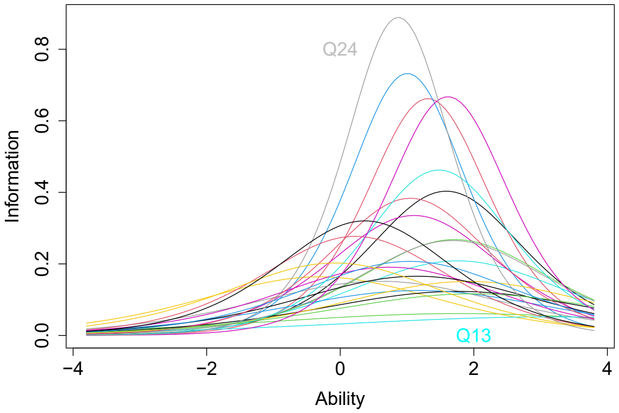

Figure 3: CCA item information curves from the 2PL IRT model. High discrimination items have high peak information.

Item response theory enables us to use the *standard error of measurement (SE)* for a student based on their ability level:

Item	$\Delta\alpha$	Item	$\Delta\alpha$	Item	$\Delta\alpha$
Q1	−0.009	Q10	−0.006	Q19	−0.003
Q2	−0.012	Q11	−0.007	Q20	−0.006
Q3	−0.001	Q12	−0.003	Q21	−0.009
Q4	−0.011	Q13	−0.000	Q22	−0.011
Q5	−0.006	Q14	−0.004	Q23	−0.003
Q6	−0.009	Q15	−0.005	Q24	−0.012
Q7	−0.002	Q16	−0.007	Q25	−0.006
Q8	−0.003	Q17	−0.004		
Q9	−0.002	Q18	−0.009		

Table 1: Change in Cronbach's α from removing each CCA item individually. Overall Cronbach's α is 0.83. Removing each item lowers reliability, suggesting that all items measure the same construct.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \qquad (2)$$

where $I(\theta)$ is the information function of the test. The possibility of calculating the standard error of measurement at different abilities is a great strength of IRT. For example, because Q24 provides more information about students across a wide range of ability levels than does Q13, it also has less measurement error across the broad spectrum of students. We will use this property to quantify the student ability levels at which the CCA can measure student knowledge with low error, giving us an answer to **RQ3**: What levels of cybersecurity knowledge does the CCA measure well?

5 RQ2, RQ3, RQ4: RESULTS

The median total CCA score was 6 out of 25 (24%), with mean 6.9 (27.6%) and quartiles 4 (16%) and 8 (32%). The median time to take the CCA was 52.6 minutes, with mean 65.7 and quartiles 31.3 and 88.5, including time to complete the demographic questions.

5.1 Classical Test Theory Results

The Cronbach's α of the CCA is 0.83, putting it in the acceptable range for CIs, and better than other commonly used CIs in CS (see Table 4). Table 1 shows that removing each item of the CCA individually results in the same or lower reliability, thus each item could potentially be kept in the CCA according to this property.

Table 2 shows the difficulty and discrimination values for each question on the CCA. The CCA is hard: for each item (except Q7 and Q23), less than half of the students answered correctly. Every item has discrimination values above the desired 0.2.

We conducted a simple test for gender bias in the CCA. A Shapiro-Wilk test [34] showed that the data significantly deviate from an assumption of normality ($W(193) = 0.778, p < 0.001$), so we used a Mann-Whitney U test [22]. We found a significant difference ($p = 0.02$) between the performance of male students ($\mu = 7.4, \sigma = 5.0$) and female students ($\mu = 5.5, \sigma = 3.3$) with moderate effect size ($d = 0.44$).

5.2 Item Response Theory Results

Table 3 shows the difficulty and discrimination parameters using the 2PL model. Figure 2 visualizes these values as item characteristic curves.

5.2.1 *Item Response Functions.*

All items have positive discrimination: a minimum requirement for item quality. All items except two (Q7 and Q23) are difficult, requiring that students be above average to answer them correctly with a 50% chance or better. One item, Q13 has troubling high difficulty, indicating that only students who are about 3 standard deviations above the mean have a 50% chance or better chance of getting that question correct.

Item	Diff.	Disc.	Item	Diff.	Disc.
Q1	0.17	0.52	Q14	0.36	0.41
Q2	0.17	0.61	Q15	0.21	0.41
Q3	0.30	0.33	Q16	0.19	0.46
Q4	0.22	0.57	Q17	0.30	0.41
Q5	0.20	0.45	Q18	0.25	0.51
Q6	0.25	0.52	Q19	0.22	0.36
Q7	0.55	0.37	Q20	0.3	0.45
Q8	0.39	0.39	Q21	0.17	0.52
Q9	0.24	0.36	Q22	0.13	0.59
Q10	0.44	0.45	Q23	0.51	0.39
Q11	0.18	0.47	Q24	0.23	0.6
Q12	0.32	0.39	Q25	0.40	0.46
Q13	0.20	0.27			

Table 2: Difficulty and discrimination of each CCA item in Classical Test Theory

Item	Diff. (b_i)	Disc. (a_i)	Item	Diff. (b_i)	Disc. (a_i)
Q1	1.59	1.27	Q14	0.73	0.87
Q2	1.31	1.63	Q15	1.91	0.78
Q3	1.79	0.50	Q16	1.68	1.03
Q4	1.00	1.71	Q17	1.14	0.81
Q5	1.76	0.91	Q18	1.05	1.24
Q6	1.11	1.16	Q19	1.99	0.68
Q7	-0.31	0.81	Q20	1.04	0.91
Q8	0.59	0.78	Q21	1.48	1.36
Q9	1.76	0.70	Q22	1.61	1.63
Q10	0.22	1.05	Q23	-0.08	0.90
Q11	1.71	1.04	Q24	0.87	1.89
Q12	1.18	0.71	Q25	0.37	1.13
Q13	3.14	0.46			

Table 3: Difficulty and discrimination of each CCA item using the 2PL Item Response Theory model.

5.2.2 *Test Information Function.*

To answer **RQ3**, we examine the CCA's test information function, computed by summing the item information functions in Figure 3. Figure 4 shows the CCA test information function compared with those for other validated CS CIs. The CCA provides peak information about students with $\theta = 1.23$. The test information function is greater than 4 on the interval $-0.20 < \theta < 2.46$, showing us that if a student's ability level is in that range, their ability level can be estimated within ±0.5 standard deviations with confidence 68%. The CCA provides precise information across a range of ability levels, primarily about students who are above average.

6 DISCUSSION AND CONCLUSIONS

In Section 3, we answered **RQ1** by showing that a panel of experts found the CCA to reflect knowledge that we expect students to know after completing a multi-course curriculum in cybersecurity. In Section 5, we answered **RQ2** by showing that the CCA has

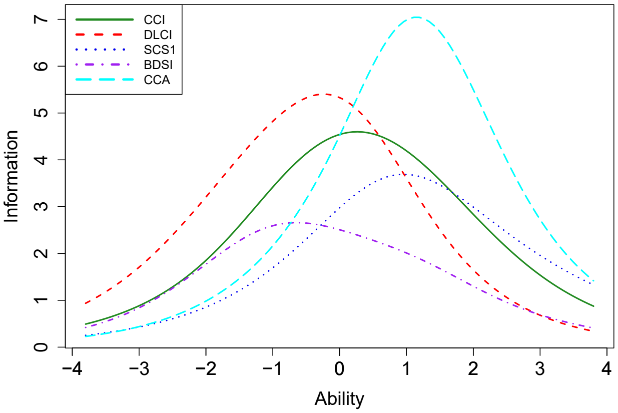

Figure 4: Test information curve for the CCA compared to the test information curves for other validated CIs in CS. Information curves for other CIs calculated from 2PL model fit parameters provided in [15, 29, 44]. Compared to other CIs, the CCA gives more information about students who performed above average.

excellent reliability and all items have generally desirable difficulty and discrimination parameters. For **RQ3**, we can see that the CCA provides detailed information about students who are about average or above average in their cybersecurity conceptual knowledge.

To answer **RQ4**, we compared the psychometric evaluation results of the CCA to those of other published or validated concept inventories in CS (see Table 4 and Figure 4). The items of the CCA all have a fairly high level of discrimination, contributing to a high level of information for the CCA relative to the other CIs. The CCA has the highest peak information of the CS CIs. Likewise, the CCA has the highest reliability of the CIs, suggesting that the CCA can provide consistent measurements of student knowledge.

The CCA is more difficult than the other CIs, but is quite similar to the SCS1, the most broadly used of the CIs [44]. The CCA is considerably harder than the CCI, which is not surprising given the fact that the CCA was designed to be harder, requiring greater technical depth. However, we did administer the CCI and CCA to different populations (students completing a first course versus students completing multiple courses, respectively), suggesting that the CCA may be have been made somewhat too difficult when compared to the CCI. Additionally, the CCA may be more difficult because of the nuanced nature of cybersecurity, where there is not always exactly one correct answer, but rather CCA distractors often have varying degrees of correctness or desirability. Future work could further investigate whether a polytomous scoring model (i.e., partial credit) that provides partial credit for "good but not best" choices might provide more useful information about lower performing students.

The IRT parameters suggest some possible future revisions to the CCA to optimize for student time or information. Eight items have difficulties approximately 1.7 or higher, meaning that several items are providing maximal information about the top 5% of students. For example, Q13 provides maximal information about students who are in the top 0.1%. Some of these items (particularly Q13) could either be dropped to make the CCA shorter and therefore easier to administer or be swapped for easier items.

Measurement	CCA	CCI	DLCI	SCS1	BDSI
Cronbach's α	0.83	0.78	0.80	0.70	0.68
Min. Difficulty	−0.31	−2.00	−1.84	0.08	−3.03
Max. Difficulty	3.14	2.04	0.55	5.07	1.25
Min. Disc.	0.46	0.37	0.28	0.49	0.33
Max. Disc.	1.99	1.47	1.68	1.53	2.03

Table 4: Comparison of the reliability and 2PL model parameters of the CCA with those of other CIs. Parameters for other CIs come from [15, 29, 44].

Based on these findings, we replaced CCA items Q3 and Q13, which have the lowest item information. We replaced them with two proven test items from the CCI that yield high information for average and below-average performing students.

These results suggest that the CCA can validly accomplish our goal of assessing students' cybersecurity conceptual knowledge after multiple courses in the field. In particular, the assessment can identify and study students who are particularly well prepared to face the diversity of nuanced challenges facing cybersecurity professionals.

6.1 Limitations

We administered the CCA only at the end of students' curriculum and cannot verify whether the CCA can be appropriately administered to students before that time in their studies. While we found a difference in performance based on gender, we do not have sufficient data to discuss whether the CCA is a biased instrument. We will need other measures, studies, and/or larger sample sizes to explore more deeply whether specific items, wording or context of questions, or other confounding variables lie at the root of the gender-performance disparity. Though we tried to recruit broadly, our sample over represents students from large, public research universities and military academies. We have not yet administered the CCI and CCA to the same populations, limiting our ability to provide deeper insights into the relative difficulty of the two inventories.

6.2 Using the CCA

We will continue to host the CCA on PrairieLearn [43] for the foreseeable future, and we invite educators to use it for research and to participate in our ongoing evaluation [1]. The authors can forward aggregated test results for students who take the CCA through PrairieLearn. The authors are also willing to provide a PDF copy of the CCA, or provide instructions on how someone might host the assessment on their own.

ACKNOWLEDGMENTS

We thank the many people who contributed to the CATS project as Delphi experts, interview subjects, Hackathon participants, expert reviewers, student subjects, and former team members, including Michael Neary, Spencer Offenberger, Geet Parekh, Konstantinos Patsourakos, Dhananjay Phatak, and Julia Thompson. Support for this research was provided in part by the U.S. Department of Defense under CAE-R grants H98230-15-1-0294, H98230-15-1-0273, H98230-17-1-0349, H98230-17-1-0347; and by the National Science Foundation under UMBC SFS grants DGE-1241576, 1753681, and SFS Capacity Grants DGE-1819521, 1820531.

REFERENCES

[1] 2022. UMBC Cyber Defense Lab. https://cisa.umbc.edu/

[2] Yifat Ben-David Kolikant and Sara Genut. 2017. The effect of prior education on students' competency in digital logic: the case of ultraorthodox Jewish students. *Computer Science Education* 27, 3-4 (2017), 149–174.

[3] Jeremiah Blanchard, Christina Gardner-McCune, and Lisa Anthony. 2020. Dual-modality instruction and learning: a case study in CS1. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 818–824. https://doi.org/10.1145/3328778.3366865

[4] Ryan Bockmon, Stephen Cooper, William Koperski, Jonathan Gratch, Sheryl Sorby, and Mohsen Dorodchi. 2020. A CS1 spatial skills intervention and the impact on introductory programming abilities. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 766–772. https://doi.org/10.1145/3328778.3366829

[5] Mark Ciampa. 2017. *CompTIA Security+ Guide to Network Security Fundamentals, Loose-Leaf Version* (6th ed.). Course Technology Press, Boston, MA, USA.

[6] CompTIA. [n.d.]. CASP (CAS-003) Certification Study Guide: CompTIA IT Certifications. https://www.comptia.org/training/books/casp-cas-003-study-guide

[7] International Information System Security Certification Consortium. [n.d.]. Certified Information Systems Security Professional. https://www.isc2.org/cissp/default.aspx. [accessed 3-14-17].

[8] International Information Systems Security Certification Consortium. [n.d.]. GIAC Certifications: The Highest Standard in Cyber Security Certifications. https://www.giac.org/.

[9] CyberCorps. 2019. Participating Institutions. https://www.sfs.opm.gov/ContactsPI.aspx

[10] Cybersecurity and Infrastructure Security Agency. [n.d.]. The National Initiative for Cybersecurity Careers & Studies. URL: https://niccs.us-cert.gov/featured-stories/take-cybersecurity-certification-prep-course.

[11] Adrienne Decker and Monica M. McGill. 2019. A topical review of evaluation instruments for computing education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 558–564. https://doi.org/10.1145/3287324.3287393

[12] CSEC2017 Joint Task Force. 2017. *Cybersecurity Curricula 2017.* Technical Report. CSEC2017 Joint Task Force.

[13] Richard R. Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics* 66, 1 (1998), 64–74. https://doi.org/10.1119/1.18809

[14] Geoffrey L Herman and Joseph Handzik. 2010. A preliminary pedagogical comparison study using the digital logic concept inventory. In *2010 IEEE Frontiers in Education Conference (FIE)*. IEEE, F1G–1.

[15] Geoffrey L Herman, Craig Zilles, and Michael C Loui. 2014. A psychometric evaluation of the digital logic concept inventory. *Computer Science Education* 24, 4 (2014), 277–303.

[16] David Hestenes, Malcolm Wells, and Gregg Swackhamer. 1992. Force concept inventory. *The Physics Teacher* 30, 3 (1992), 141–158. https://doi.org/10.1119/1.2343497

[17] CAE in Cybersecurity Community. 2019. CAE Institution Map. https://www.caecommunity.org/content/cae-institution-map

[18] Association for Computing Machinery (ACM) Joint Task Force on Computing Curricula and IEEE Computer Society. 2020. *Computing Curricula 2020 Paradigms for Global Computing Education.* Association for Computing Machinery, New York, NY, USA. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/cc2020.pdf

[19] Natalie Jorion, Brian Gane, Katie James, Lianne Schroeder, Louis V. DiBello, and James Pellegrino. 2015. An Analytic Framework for Evaluating the Validity of Concept Inventory Claims. *Journal of Engineering Education* 104 (10 2015), 454–496. https://doi.org/10.1002/jee.20104

[20] Martin C. Libicki, David Senty, and Julia Pollak. 2014. *Hackers wanted: an examination of the cybersecurity labor market.* RAND.

[21] Lauren Margulieux, Tuba Ayer Ketenci, and Adrienne Decker. 2019. Review of measurements used in computing education research and suggestions for increasing standardization. *Computer Science Education* 29, 1 (Jan. 2019), 49–78. https://doi.org/10.1080/08993408.2018.1562145 Publisher: Routledge.

[22] Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test.* John Wiley & Sons, Ltd, 1–1. https://doi.org/10.1002/9780470479216.corpsy0524 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470479216.corpsy0524

[23] Greg L Nelson, Benjamin Xie, and Amy J Ko. 2017. Comprehension first: evaluating a novel pedagogy and tutoring system for program tracing in CS1. In *Proceedings of the 2017 ACM Conference on International Computing Education Research.* 2–11.

[24] NIST. [n.d.]. NICE Framework. http://csrc.nist.gov/nice/framework/. [Online; accessed 8-October-2016].

[25] Jon Oltsik. 2020. The cybersecurity skills shortage is getting worse. https://www.csoonline.com/article/3571734/the-cybersecurity-skills-shortage-is-getting-worse.html [Online; accessed 21-August-2020].

[26] Panayiotis Panayides. 2013. Coefficient alpha: Interpret with caution. *Europe's Journal of Psychology* 9, 4 (11 2013). https://doi.org/10.5964/ejop.v9i4.653

[27] Geet Parekh, David DeLatte, Geoffrey L Herman, Linda Oliva, Dhananjay Phatak, Travis Scheponik, and Alan T Sherman. 2017. Identifying core concepts of cybersecurity: Results of two Delphi processes. *IEEE Transactions on Education* 61, 1 (2017), 11–20.

[28] M.C. Parker, M. Guzdial, and S. Engleman. 2016. Replication, validation, and use of a language independent CS1 knowledge assessment. ICER 2016 - Proceedings of the 2016 ACM Conference on International Computing Education Research (2016), 93–101.

[29] Leo Porter, Daniel Zingaro, Soohyun Nam Liao, Cynthia Taylor, Kevin C Webb, Cynthia Lee, and Michael Clancy. 2019. BDSI: A validated concept inventory for basic data structures. In *Proceedings of the 2019 ACM Conference on International Computing Education Research.* 111–119.

[30] Seth Poulsen, Geoffrey L. Herman, Peter AH Peterson, Enis Golaszewski, Akshita Gorti, Linda Oliva, Travis Scheponik, and Alan T Sherman. 2021. Psychometric evaluation of the Cybersecurity Concept Inventory. *ACM Transactions on Computing Education (TOCE)* 22, 1 (November 2021), 1–18.

[31] National Research Council Board on Testing, Center for Education Assessment, Division of Behavioral, Social Sciences, and Education. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment.* The National Academies Press, Washington, DC. https://doi.org/10.17226/10019

[32] Travis Scheponik, Enis Golaszewski, Geoffrey Herman, Spencer Offenberger, Linda Oliva, Peter A. H. Peterson, and Alan T. Sherman. 2020. Investigating Crowdsourcing to Generate Distractors for Multiple-Choice Assessments. In *National Cyber Summit (NCS) Research Track*, Kim-Kwang Raymond Choo, Thomas H. Morris, and Gilbert L. Peterson (Eds.). Springer International Publishing, Cham, 185–201. https://arxiv.org/pdf/1909.04230.pdf.

[33] Offensive Security. [n.d.]. Penetration Testing with Kali Linux (PWK). https://www.offensive-security.com/pwk-oscp/.

[34] S. S. Shapiro and M. B. Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3/4 (1965), 591–611. http://www.jstor.org/stable/2333709

[35] Alan T. Sherman, David DeLatte, Michael Neary, Linda Oliva, Dhananjay Phatak, Travis Scheponik, Geoffrey L. Herman, and Julia Thompson. 2018. Cybersecurity: Exploring core concepts through six scenarios. *Cryptologia* 42, 4 (2018), 337 – 377.

[36] Alan T. Sherman, Geoffrey L. Herman, Linda Oliva, Peter A. H. Peterson, Enis Golaszewski, Seth Poulsen, Travis Scheponik, and Akshita Gorti. 2021. Experiences and Lessons Learned Creating and Validating Concept Inventories for Cybersecurity. In *National Cyber Summit (NCS) Research Track 2020*, Kim-Kwang Raymond Choo, Tommy Morris, Gilbert L. Peterson, and Eric Imsand (Eds.). Springer International Publishing, Cham, 3–34.

[37] Dan Swinhoe. 2020. The 15 biggest data breaches of the 21st century. https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html [Online; accessed 21-August-2020].

[38] Alan T. Sherman, Linda Oliva, David DeLatte, Enis Golaszewski, Michael Neary, Konstantinos Patsourakos, Dhananjay Phatak, Travis Scheponik, Geoffrey Herman, and Julia Thompson. 2017. Creating a Cybersecurity Concept Inventory: A Status Report on the CATS Project. *2017 National Cyber Summit* (06 2017).

[39] Alan T. Sherman, Linda Oliva, Enis Golaszewski, Dhananjay Phatak, Travis Scheponik, Geoffrey Herman, Dong San Choi, Spencer Offenberger, Peter Peterson, Josiah Dykstra, Gregory Bard, Ankur Chattopadhyay, Filipo Sharevski, Rakesh Verma, and Ryan Vrecenar. 2019. The CATS Hackathon: Creating and Refining Test Items for Cybersecurity Concept Inventories. In *IEEE Security and Privacy.*

[40] A. E. Tew and M. Guzdial. 2011. The FCS1: A language independent assessment of CS1 knowledge. (2011), 111–116.

[41] Julia Thompson, Geoffrey Herman, Travis Scheponik, Linda Oliva, Alan T. Sherman, and Ennis Golaszewski. 2018. Student misconceptions about cybersecurity concepts: Analysis of think-aloud interviews. *Journal of Cybersecurity Education, Research and Practice* (07 2018).

[42] Matt Walker. 2011. *CEH Certified Ethical Hacker All-in-One Exam Guide* (1st ed.). McGraw-Hill Osborne Media.

[43] Matthew West, Geoffrey L. Herman, and Craig Zilles. 2015. PrairieLearn: Mastery-based Online Problem Solving with Adaptive Scoring and Recommendations Driven by Machine Learning. In *2015 ASEE Annual Conference & Exposition.* ASEE Conferences, Seattle, Washington, 26.1238.1–26.1238.14. https://peer.asee.org/24575.

[44] Benjamin Xie, Matthew J. Davidson, Min Li, and Amy J. Ko. 2019. An item response theory evaluation of a language-independent CS1 knowledge assessment. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19)*. Association for Computing Machinery, Minneapolis, MN, USA, 699–705. https://doi.org/10.1145/3287324.3287370

[45] Benjamin Xie, Greg L Nelson, and Amy J Ko. 2018. An explicit strategy to scaffold novice program tracing. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education.* 344–349.