Psychometric Evaluation of the Cybersecurity Concept Inventory

SETH POULSEN, University of Illinois at Urbana-Champaign
GEOFFREY L. HERMAN, University of Illinois at Urbana-Champaign
PETER A.H. PETERSON, University of Minnesota Duluth
ENIS GOLASZEWSKI, University of Maryland, Baltimore County
AKSHITA GORTI, University of Maryland, Baltimore County
LINDA OLIVA, University of Maryland, Baltimore County
TRAVIS SCHEPONIK, University of Maryland, Baltimore County
ALAN T. SHERMAN, University of Maryland, Baltimore County

We present a psychometric evaluation of a revised version of the *Cybersecurity Concept Inventory (CCI)*, completed by 354 students from 29 colleges and universities. The CCI is a conceptual test of understanding created to enable research on instruction quality in cybersecurity education. This work extends previous expert review and small-scale pilot testing of the CCI. Results show that the CCI aligns with a curriculum many instructors expect from an introductory cybersecurity course, and that it is a valid and reliable tool for assessing what conceptual cybersecurity knowledge students learned.

CCS Concepts: • **Social and professional topics → Student assessment**; **Computing education**; • **Security and privacy**;

Additional Key Words and Phrases: Classical Test Theory (CTT), concept inventories, Cybersecurity Assessment Tools (CATS), Cybersecurity Concept Inventory (CCI), cybersecurity education, Item Response Theory (IRT), psychometric evaluation.

ACM Reference Format:
Seth Poulsen, Geoffrey L. Herman, Peter A.H. Peterson, Enis Golaszewski, Akshita Gorti, Linda Oliva, Travis Scheponik, and Alan T. Sherman. 2021. Psychometric Evaluation of the Cybersecurity Concept Inventory. In *20XX Technical Symposium on Computer Science Education, March xx-xx, 2021, City, Country*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 Introduction

Knowledge of cybersecurity principles is critical for individuals and organizations to keep sensitive data secure and to avoid theft and other digital threats. Security breaches undermine the effectiveness of business, governments, and other organizations. Frequent news reports about major security breaches highlight the importance of cybersecurity [38]. Despite the paramount importance of digital security, there is a large and growing shortage of cybersecurity professionals [21, 26]. It is essential that we increase the efficiency and effectiveness of educational programs to fill this need.

To conduct reproducible research on the benefits and drawbacks of different methods and curricular structures for teaching cybersecurity, we should use validated assessment instruments to

SIGCSE 'XX, March xx-xx, 20XX, City, Country

ACM ISBN 978-1-4503-6874-2/20/06…$15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

minimize the amount of error with which we measure student knowledge. Until now, no such instrument has existed.

In this paper, we present the *Cybersecurity Concept Inventory (CCI)*, a validated instrument for assessing student knowledge of introductory cybersecurity concepts. After briefly reviewing the creation of the CCI, we give statistical evidence of its reliability and validity as an assessment of cybersecurity knowledge for students who have taken an introductory cybersecurity course.

2 Background

For context, we will provide a brief history of assessment instruments used in computer science education, and then we explain how we created the CCI.

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 physics by creating a meaningful way to compare the results of different pedagogical techniques [13, 17].

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)* [16], the *Multilanguage Assessment of CS1 Knowledge (SCS1)* [12, 29, 41], and the *Basic Data Structures Inventory (BDSI)* [30]. For a more extensive review of assessment instruments used in computing education research, see [10, 22].

The effect of CIs has not yet been as far reaching in computer science (CS) as it has been in other disciplines, such as physics, likely because CIs have not been in use in CS for as long. Despite their recent creation they have already been useful for many purposes in computing education, including but not limited to: examining the relationship between spatial ability and learning programming [3], comparing outcomes between digital logic courses which use differing pedagogical approaches [14], evaluating novel instructional practices in CS1 [23, 46], evaluating the effectiveness of teaching students using both block- and text-based programming languages [2], and understanding the impact of students' educational background on learning topics in computer science [1]. This activity confirms the utility of creating concept inventories to the computing education community.

2.2 Cybersecurity Assessment Exams

There are a number of cybersecurity assessment exams in broad use already, but we are unaware of any scientific study that characterizes the properties of any of these tests, necessitating the development of assessments that can be used as research instruments for pedagogical research. For example, there are several existing certification exams, including ones listed by NICCS as relevant [9].

CASP+ [5] comprises multiple-choice and performance tasks items including enterprise security, risk management, and incident response. OSCP [35] (offensive security) is a 24-hour practical test focusing on penetration testing. Other exams include CISSP, Security+, and CEH [4, 6, 43], which are mostly informational, not conceptual. Global Information Assurance Certification (GIAC) [7] offers a variety of vendor-neutral MCQ certification exams linked to SANS courses; for each exam type, the gold level requires a research paper, and none of them are suitable for use as a pedagogical research instrument.

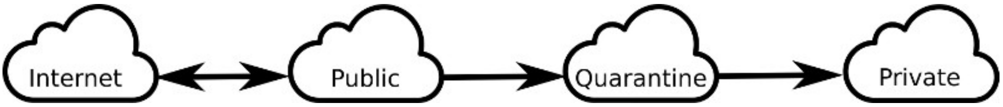
Additionally, the ACM, IEEE, and ABET have been working on curricular guidance for cybersecurity [11, 19], and the NICE Cybersecurity Workforce Framework [24] establishes a common lexicon for explaining a structured description of professional cybersecurity positions in the workforce with detailed documentation 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. For more details, see Sherman, et al. [37].

2.3 The Cybersecurity Assessment Tools (CATS) Project

Given this lack of validated assessment tools, the authors founded the *Cybersecurity Assessment Tools (CATS)* project in the interest of creating validated educational assessment tools for cybersecurity [25, 28, 33, 34, 36, 37, 39, 40, 42].

Unlike some areas of computer science, 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 what may be poor, mediocre, or ideal solutions to a given security problem. The questions on the CCI were designed to encourage this type of thinking. The CCI presents a series of scenarios, and asks questions about the scenario that force students to weigh their options and select the best solution choice to the security problem. Figure 1 gives an example test item from the CCI.

Scenario. An enterprise with highly sensitive data needs to be able to retrieve information from the internet. To support this requirement while protecting its sensitive data, the enterprise partitions its internal computer network into three segments: Public, Quarantine, and Private. In this system, data can flow ONLY from Internet to Public, Public to Internet, Public to Quarantine, and from Quarantine to Private.



Question. Choose the most effective method to ensure that, pertaining to the section of the network involving Public, Quarantine, and Private, data flow only from Public to Quarantine, and from Quarantine to Private:

A. Authenticate all flows of data.
B. Restrict access to authorized users only.
C. Encrypt all flows of data.
D. Install software firewalls between the segments.
E. Use only one-way physical connections between the segments.

Fig. 1. CCI Question 6 probes the concept "Devise a defense."

We now explain how we created the CCI; for more details, see [37]. First, our team engaged 33 cybersecurity experts in a Delphi process to identify the core concepts of cybersecurity that should be tested [28], which can be seen in Table 1. Next the team developed cybersecurity scenarios. We used these scenarios in a series of open-ended interviews with students to identify common misconceptions [34]. We then used these misconceptions to aid in constructing compelling distractors for multiple-choice questions. Table 2 shows the topic of each question on the CCI, and which of the five cored concepts it addresses.

During fall 2018 we had a group of experts review the CCI to ensure that they believed the test questions were sound and the assessment covers the topics that cybersecurity educators would

[1] CIA Triad (Confidentiality, Integrity, Availability).

1 (V)	Identify vulnerabilities and failures
2 (C)	Identify attacks against CIA triad[1] and authentication
3 (D)	Devise a defense
4 (G)	Identify the security goals
5 (T)	Identify potential targets and attackers

Table 1. The five core concepts underlying the CCI and CCA embody aspects of adversarial thinking.

Question	Topic	Core Concepts
1	Message Authentication Codes	T
2	Message Authentication Codes	G
3	Non-Repudiation	T
4	Input Validation	D
5	Network Design	G
6	Network Design	D
7	Network Design	V
8	Two-Factor Authentication	C
9	Replay Attacks	C
10	Integrity	C,V
11	Physical Attack	C
12	Insider Threats	T
13	Security Theater	V
14	Public-Key Cryptography	D
15	Replay Attacks	G
16	Authentication	V
17	Public-Key Cryptography	V
18	Authentication	T,C
19	Authorization	G
20	Encryption	D
21	Social Engineering	T
22	Biometric Authentication	C
23	Network Design	D
24	Physical Attack	G
25	Protocols	C,V

Table 2. The cybersecurity topic and core concepts tested by each of the items on the CCI (see Table 1 for concept abbreviations).

expect it to. We found that most experts approved of most of the questions, and agreed that the questions on the CCI covered the knowledge that they would want their students to have after a first course on cybersecurity [25]. We also pilot tested the CCI with 142 students, showing that the CCI has some desirable psychometric properties [25], and gaining insight into which questions did not work as well as we had hoped for assessing knowledge. Since then, we revised the CCI to improve items that were too hard or did not discriminate well between lower- and higher-performing students.

Using our revised version of the test, we started a more comprehensive round of data collection from fall 2019 through spring 2020. In this paper, we analyze these new data to understand the statistical evidence for the reliability and validity of the CCI. More specifically, we answer the following research questions:

> **RQ1**: What does the statistical evidence say about the reliability and validity of the CCI?
> **RQ2**: What levels of cybersecurity knowledge does the CCI measure well?
> **RQ3**: How do the statistical properties of the CCI compare with other concept inventories in use?
> **RQ4**: What can we learn by examining the response patterns to questions with desirable psychometric properties?

3 Methods

We explain how we collected and analyzed data.

3.1 Data Collection

We pursued multiple avenues for recruiting subjects to take the assessment, including emailing professors who do research in cybersecurity, talking to colleagues, and contacting institutions involved with cybersecurity education programs such as Scholarship for Service [8] and institutions qualifying as Centers for Academic Excellence in Cyber Defense (CAEs) [18]. By far the most effective recruitment strategy was making use of the professional connections of the members of our research team, who are embedded in the cybersecurity research and teaching communities [37].

We hosted the CCI on PrairieLearn, an online, open source homework and exam platform, to facilitate the administration of the assessment to students at a range of institutions [44].

For most students, their instructor offered some extra credit to complete the CCI. We collected data from September 2019 through May 2020. The institutional review board at the University of Maryland, Baltimore County approved our protocol.

A total of 574 students started the CCI in PrairieLearn. After we discarded test instances where the student did not complete the assessment, or spent less than 15 minutes on the assessment, our data set consists of scores from 354 students from 29 colleges and universities. Since it takes about 15 minutes just to read all the assessment questions, no student could complete the assessment in good faith in under 15 minutes. In the sanitized data set, the mean time to finish the test is 45 minutes, with 272 out of 354 (77%) test takers finishing in under an hour.

Our participants came from a range of institutions including private and public universities and community colleges, with the full list shown in Table 3. Institutions were geographically diverse within the United States, with a few data points coming from other countries as well. We collected data from both research-focused and teaching-focused institutions. The majority of students came from large, public research universities as they were often able to provide more subjects for testing. Subjects also came from courses with a variety of titles including Computer Security, Cybersecurity Concepts, and Information Assurance and Security, all of which covered most or all of the core cybersecurity topics that the CCI seeks to assess. Most students who took the CCI were CS majors in the latter half of completing their bachelor's degree.

3.2 Item Response Theory vs. Classical Test Theory

Classical Test Theory (CTT) and *Item Response Theory (IRT)* are two commonly used analytical frameworks for showing statistical support for the validity of assessment instruments, and for gauging the skill of students taking an assessment [20]. Both CTT and IRT give a measurement of each question's *difficulty*, that is, how hard it is to answer a question correctly, and its *discrimination*, how well a question differentiates between students of lower and higher skill levels. These

Institution	Number of Participants	Location	Highest Degree Granted	Ownership	Size
University of Illinois at Urbana-Champaign*	95	Midwest	Doctoral	Public	Large
University of Maryland College Park	61	East Coast	Doctoral	Public	Large
Texas A& M San Antonio	31	South	Masters	Public	Small
University of Maryland, Baltimore County*	20	East Coast	Doctoral	Public	Medium
Universidad de Alcalá de Henares	16	Spain	Doctoral	Public	Large
Penn State University	12	East Coast	Doctoral	Public	Large
Towson University	11	East Coast	Doctoral	Public	Medium
Tufts University	11	East Coast	Doctoral	Private	Medium
Glendale Community College	10	Southwest	Associates	Public	Medium
University of Georgia	10	South	Doctoral	Public	Large
University of Minnesota Duluth*	10	Midwest	Doctoral	Public	Medium
Portland Community College	9	West Coast	Associates	Public	Large
University of San Francisco	9	West Coast	Doctoral	Private	Medium
Florida State College at Jacksonville	8	South	Bachelors	Public	Large
Morgan State University	7	East Coast	Doctoral	Public	Small
Tennessee Technological University	6	South	Doctoral	Public	Small
University of Wisconsin Green Bay	6	Midwest	Masters	Public	Small
University of Mississippi	4	South	Doctoral	Public	Medium
Northern Kentucky University	3	South	Doctoral	Public	Medium
Marymount University	3	East Coast	Doctoral	Private	Small
University of Arkansas at Little Rock	2	South	Doctoral	Public	Small
University of California Santa Barbara	2	West Coast	Doctoral	Public	Large
Idaho State University	2	Mountain West	Doctoral	Public	Medium
Purdue University	1	Midwest	Doctoral	Public	Large
Polytechnic University of Puerto Rico	1	Puerto Rico	Bachelors	Private	Small
University of Melbourne	1	Australia	Doctoral	Public	Large
University of New Mexico	1	Southwest	Doctoral	Public	Large
Arizona State University	1	Southwest	Doctoral	Public	Large
Virginia Tech	1	East Coast	Doctoral	Public	Large

Table 3. Schools who participated in the study, number of participants from the school who gave valid responses to the assessment, and demographic data about the school. Institutions with less than 10,000 are listed as small, institutions with between 10,000 and 25,000 students are considered medium, and institutions with greater than 25,000 students are considered large. We use * to denote authors' institutions. Schools in the U.S. are listed by region, and schools outside the U.S. are listed by country.

metrics are defined differently between CTT and IRT, and therefore they should not be compared across frameworks. A robust assessment will have test questions with a range of difficulty levels to obtain information about students at a range of ability levels. It is desirable to have questions with high discrimination, because questions that do a better job differentiating between students of higher or lower ability can measure student ability more accurately.

CTT can be used on samples of any size and is useful for obtaining a simple measurement of the reliability of assessment instruments. Some strengths of IRT that CTT does not have are:

(1) Falsifiable assumptions: the assumptions of CTT must be taken as a given, whereas the assumptions of IRT can be tested using the data set and appropriate statistical tests.
(2) CTT assumes that the measurement error is the same for any student taking the test, where IRT allows us to see if there is a different measurement error for students of different ability levels.
(3) IRT enables us to estimate, for each question, how much information the question provides about each student.

IRT requires a larger sample size than CTT because CTT determines the difficulty and discrimination of each item as independent parameters or as a simple correlation between each independent item and the test as a whole, respectively. In contrast, IRT jointly determines the difficulty and discrimination of every item simultaneously. This simultaneous determination, requires fitting a model that grows increasingly complex as the number of items increases. We use both CTT and IRT to demonstrate the reliability and validity of the CCI, providing us an answer to **RQ1**.

3.3 Classical Test Theory

Due to small sample size, our team used only CTT to analyze the results of the pilot testing [25]. Here we use CTT as a means of comparing the current properties of the test to the properties of the earlier draft of the test, and as a way to verify basic properties of the test, such as whether the CTT difficulty and discrimination fit into the accepted ranges.

CTT assumes that each student has a true score (T) which, together with some error term (E), gives the student's actual score (X), so that $X = T + E$.

3.3.1 Reliability. As part of the overall evaluation of the test's reliability, we calculate the Cronbach's α, the most common measure of internal consistency for tests. 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 generally accepted value of Cronbach's α to denote a reliable assessment, but most people agree that for a test to be used in high-stakes scenarios, such as assigning grades, the Cronbach's α should be at least 0.7 or 0.8 [20, 27].

One method for evaluating the quality of items in an assessment is to compare the Cronbach's α with what it would be if the particular item was removed [20]. If removing a test item from the test increases the reliability of the test, the item may be poor quality, and would become a potential candidate for removal, especially if it has other undesirable properties. Items which do not increase the reliability of the test but have a positive discrimination are usually kept, as additional questions allow us to obtain more information about student knowledge (see Section 3.4.2).

3.3.2 Difficulty and Discrimination. In CTT, the *difficulty* is the percentage of students who correctly answered a given item, and the *discrimination* is the point biserial correlation between a student's score on the question and their score on the test [16].

3.4 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. Along with following accepted methods for concept inventory evaluation [20], we used a data driven approach to selecting an IRT model for our data. We fit the Rasch, *two-parameter logistic (2PL)*, and *three-parameter logistic (3PL)* models using the R package ltm [31, 32], which estimates the item parameters using marginal maximum likelihood estimation. 3PL did not reach a stable solution. This result is not surprising because it usually takes a large amount of data to fit a model with that many parameters (i.e., 2PL has $2N$ parameters, while 3PL has $3N$ parameters, where N is the number of items). A likelihood ratio test shows that 2PL does a significantly better job explaining the data than does the Rasch model ($p < 0.001$, LRT = 72.07). Since the 2PL is standard in concept inventory evaluation, and it fits our data the best, we will focus on this model.

The 2PL 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:

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

The distribution of student ability parameters θ_n is given a mean of 0 and standard deviation of 1.

3.4.1 *Item Response Functions.*

Inserting the difficulty and discrimination parameters for each test item into Equation 1 gives the *item response functions*, 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. Some example item response functions for items with different parameters are shown in in Figure 2. The difficulty of the item determines the ability level at which a student has a 50% chance of answering a question correctly. For example, the solid line in Figure 2, which has difficulty 0, represents a test item which 50% of students with mean ability level will answer correctly (recall that student ability levels are normalized around 0, so an ability level 0 is mean ability level). The dotted line represents an item which is slightly easier, with a difficulty of -0.5, meaning that 50% of students whose ability level is half a standard deviation below the mean will answer it correctly, and more than 50% of students with mean ability level will answer it correctly. The dashed line represents an item which has the same difficulty as the solid line, but has a greater discrimination. This means that as a student's ability level rises above the mean, their chance of getting the questions right increases more quickly than for questions with a lower discrimination. As a result, a question with higher discrimination will measure student ability with less error.

3.4.2 *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 item information function of the instrument.

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

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

where $I(\theta)$ is the information function of the test. As stated, the possibility of calculating the standard error of measurement at different abilities is one of the great strengths of IRT. We will

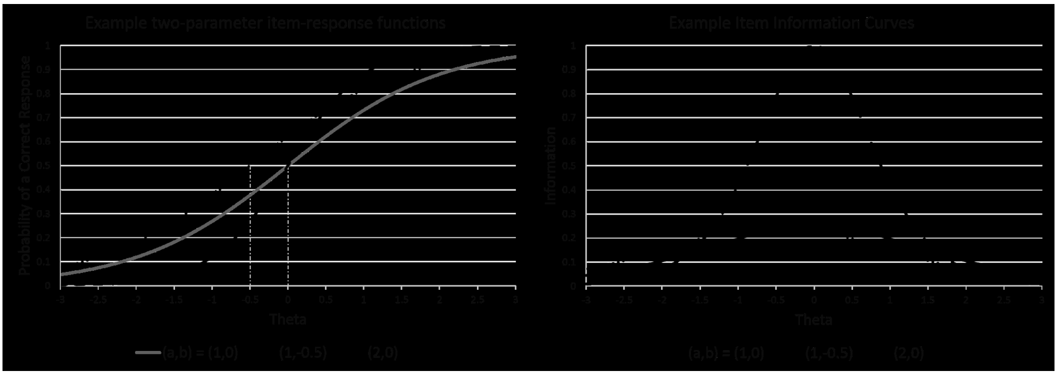


Fig. 2. Left: Three example item response functions with varying discrimination (a_i) and difficulty (b_i). Right: Item information curves for the same example items.

use this property to quantify the student ability levels at which the CCI can measure student knowledge with low error, giving us the answer to **RQ2**: What levels of cybersecurity knowledge does the CCI measure well?

Figure 2 shows some example item information functions. The solid item collects more information about higher performing students than the dotted item, due to having higher difficulty. It is desirable for an assessment to have questions with a range of difficulties so that student ability is measured at multiple levels. The high discrimination of the dashed item allows it to provide much more information across a range of ability levels than either of the other two items. It is always desirable for assessment items to have higher discrimination and thus provide measurements with higher information and lower standard error.

4 Results

We analyze CCI test data from CTT and IRT perspectives.

4.1 Classical Test Theory Results

CTT reveals information about the CCI's reliability, difficulty, and discrimination.

4.1.1 Reliability. The Cronbach's α of the CCI is 0.78, putting it in the acceptable range for CIs, and comparable to other commonly used CIs in CS, as shown in Table 7. As shown in Table 4, removing each item of the CCI individually results in the same or lower reliability. This property, along with the other psychometric properties of the questions, show that none of the items on the test need to be considered for removal.

4.1.2 Difficulty and Discrimination. Figure 3 shows a comparison of the classical test theory difficulty and discrimination of an earlier version of the CCI (Plot A) compared to those of the current version (Plot B). Table 5 displays the same information in tabular form. This table shows that our revisions to the test successfully strengthened its validity as a measurement of cybersecurity knowledge. This comparison demonstrates the value of continuing to revise and develop an assessment instrument past the initial validation phase.

4.2 Item Response Theory Results

Table 6 shows the difficulty and discrimination parameters for each question as predicted by fitting our data to the 2PL model shown in Equation 1.

Item	Change in α with item removed	Item	Change in α with item removed
Q1	−0.01	Q14	0.00
Q2	−0.01	Q15	−0.01
Q3	−0.01	Q16	−0.01
Q4	−0.01	Q17	0.00
Q5	−0.01	Q18	0.00
Q6	−0.01	Q19	−0.01
Q7	0.00	Q20	0.00
Q8	−0.01	Q21	−0.01
Q9	−0.01	Q22	−0.01
Q10	−0.01	Q23	−0.01
Q11	−0.01	Q24	0.00
Q12	−0.01	Q25	0.00
Q13	−0.01		

Table 4. Change in the reliability of the test with each item removed. The overall Cronbach's α is 0.78, which is in the acceptable range, and is comparable or better than those of many accepted concept inventories (see Table 7). Removing each item individually results in either the same or lower reliability, suggesting that none of the items on the test are weak enough to be considered for removal.

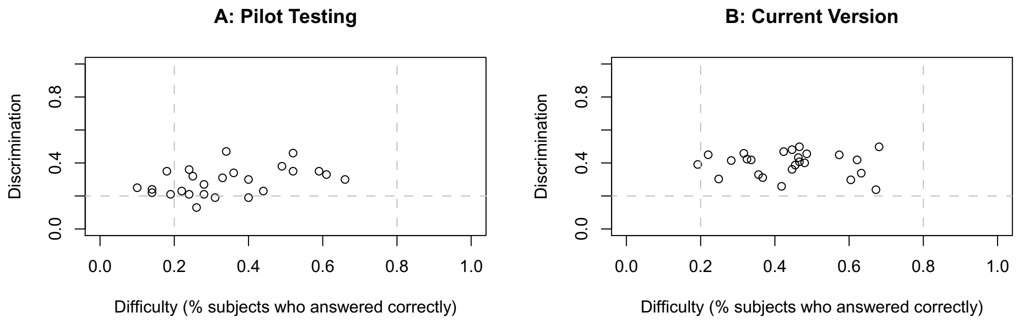


Fig. 3. Comparison of the classical test theory difficulty and discrimination of an earlier version of the CCI (Plot A) to that of the current version (Plot B). The revisions to the test successfully strengthened its validity as a measurement of cybersecurity knowledge. All CCI items are now in the accepted discrimination range for Classical Test Theory (above 0.2), and all but one are in the accepted difficulty range (between 0.2 and 0.8) [20]. We created Plot A from data given in [25].

4.2.1 Item Response Functions. Figure 4 shows the item response functions for the CCI, which represent the probability that a student of a given ability will answer a given item correctly. A steep curve, such as that of Q23, shows that a question has high discrimination, and a more shallow curve, such as that of Q7, shows that a question has lower discrimination. If a question's item response function has a negative slope (or discrimination), that would mean students with lower ability are more likely to answer the item successfully than are students with higher ability. This outcome would cause serious concern about the usefulness of the test item. None of the questions of the CCI have this problem.

Item	Diff. (Pilot)	Disc. (Pilot)	Item	Diff.(Pilot)	Disc. (Pilot)
Q1	0.28 (0.24)	0.42 (0.21)	Q14	0.42 (0.25)	0.26 (0.32)
Q2	0.43 (0.33)	0.47 (0.31)	Q15	0.32 (0.10)	0.46 (0.25)
Q3	0.34 (0.26)	0.42 (0.13)	Q16	0.33 (0.59)	0.43 (0.35)
Q4	0.45 (0.52)	0.48 (0.46)	Q17	0.63 (0.52)	0.33 (0.35)
Q5	0.46 (0.18)	0.40 (0.35)	Q18	0.60 (0.31)	0.29 (0.19)
Q6	0.45 (0.22)	0.36 (0.23)	Q19	0.49 (0.28)	0.46 (0.27)
Q7	0.67 (0.66)	0.24 (0.30)	Q20	0.36 (0.14)	0.33 (0.22)
Q8	0.19 (0.19)	0.39 (0.21)	Q21	0.46 (0.44)	0.43 (0.23)
Q9	0.47 (0.61)	0.50 (0.33)	Q22	0.57 (0.34)	0.44 (0.47)
Q10	0.62 (0.40)	0.42 (0.19)	Q23	0.68 (0.49)	0.50 (0.38)
Q11	0.46 (0.36)	0.39 (0.34)	Q24	0.37 (0.40)	0.31 (0.30)
Q12	0.48 (0.24)	0.40 (0.36)	Q25	0.25 (0.14)	0.31 (0.24)
Q13	0.22 (0.28)	0.45 (0.21)			

Table 5. Difficulty and discrimination of each item in Classical Test Theory, compared to its difficulty and discrimination at Pilot testing time.

Item	Diff. (a_i)	Disc. (b_i)	Item	Diff.(a_i)	Disc. (b_i)
Q1	1.14	0.94	Q14	0.85	0.39
Q2	0.32	1.10	Q15	0.82	1.13
Q3	0.87	0.88	Q16	0.90	0.92
Q4	0.23	1.14	Q17	-0.90	0.65
Q5	0.19	0.80	Q18	-0.94	0.46
Q6	0.34	0.66	Q19	0.04	1.04
Q7	-2.00	0.37	Q20	1.06	0.58
Q8	1.73	0.98	Q21	0.18	0.89
Q9	0.11	1.38	Q22	-0.37	0.95
Q10	-0.63	0.94	Q23	-0.72	1.47
Q11	0.23	0.82	Q24	0.99	0.57
Q12	0.09	0.78	Q25	2.04	0.57
Q13	1.29	1.23			

Table 6. Difficulty and discrimination of each question in the 2PL item response theory model.

4.2.2 Item Information Functions. Figure 5 shows the item information functions for each of the test items under 2PL. Some test items provide a great deal of information about the student's ability, such as Q23, while other items provide very little, such as Q7 and the other questions with near-flat item information functions. All the information curves are concave down, showing that none of the test items decrease the amount of information we know about a student's ability.

Figure 6 shows the item information function for the CCI, along with test information function for other accepted CIs in computer science. The CCI provides peak information about students with $\theta = 0.26$ (students whose ability is 0.26 standard deviations above the mean). The test information can be used to quantify the error of measurement of student ability at all ability levels, providing the answer to **RQ2**. For example, the test information curve is greater than 4 on the interval $-0.61 < \theta < 1.17$, telling 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%.

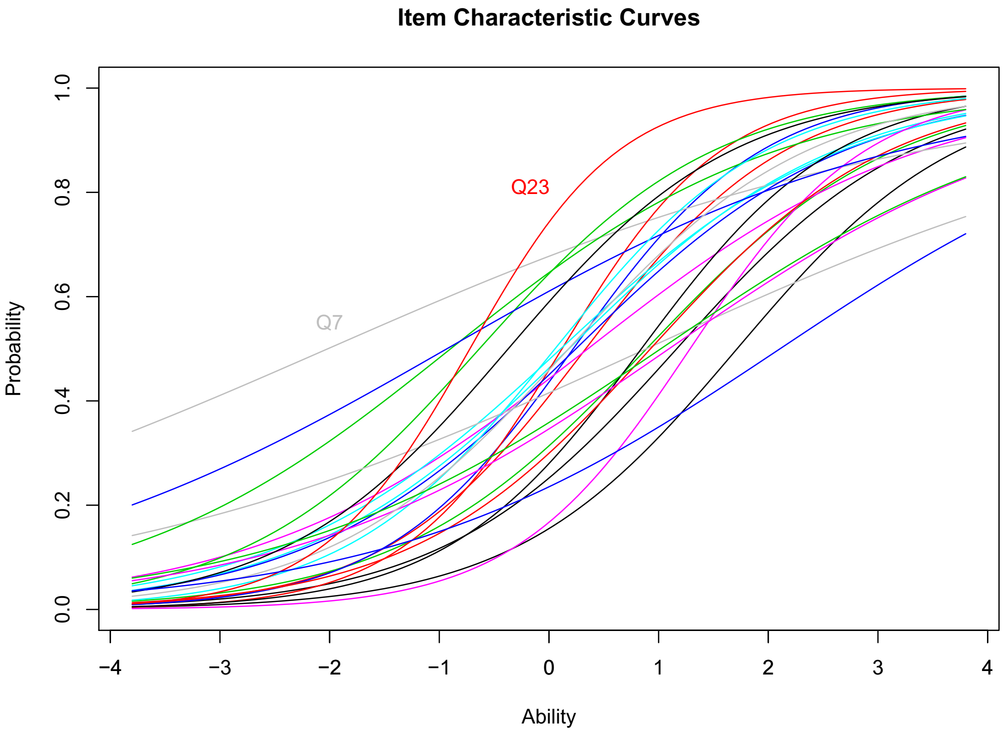


Fig. 4. Item characteristic curves from the 2PL IRT model. The differences in slopes of the lines show the variance of discrimination between parameters. These data confirm that the Rasch model's assumption, that each item has the same discrimination, is not a good fit for our data.

5 Discussion

To answer **RQ3**, we compare the psychometric evaluation results of the CCI to those of other accepted concept inventories in CS. The results are promising. Table 7 gives a quick overview showing that the statistical properties of the CCI are in the same general range as those for other CIs, and Figure 6 compares the information function for each CI. In comparison to the SCS1, the CCI does an excellent job providing information about students both above and below mean ability level, whereas the SCS1 is a very difficult test, providing much more information about students above the mean than below [45].

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

Table 7. Comparison of the reliability and 2PL model parameters of the CCI with those of other CIs. Parameters for other CIs come from [16, 30, 45].

In comparison with the DLCI, however, the CCI has some questions that provide relatively little information. We theorize that the nature of cybersecurity questions is such that measuring

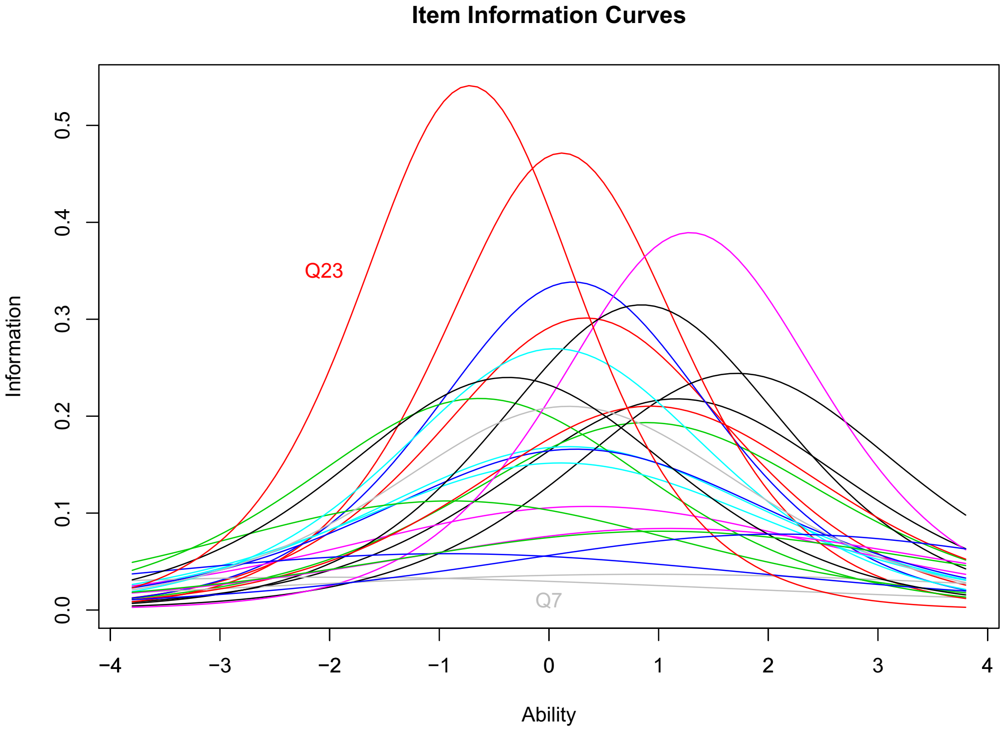


Fig. 5. Item information curves from the 2PL IRT model. Some test items provide a great deal of information about the student's ability, while other items provide very little. All the information curves are concave down, showing that none of the test items remove information.

student ability with low error is more difficult than in some other domains. For example, in the DLCI, most questions have an answer that is clearly correct, and answers that are clearly wrong. On the contrary, many questions on the CCI have answers that are good, better, and best, and we expect a student to pick the "best" answer to be awarded any points (with no partial credit awarded for other answers).

5.1 Expert Analysis of Interesting Items

To answer **RQ4**, we now provide an expert analysis of some of the highest-performing items on the CCI. For many items, we attribute their high discrimination in part to having high-quality distractors—answers that many students pick when they have a common misconception about the cybersecurity scenario. For six items (Q1, Q3, Q8, Q13, Q16, Q20), there was a distractor that subjects found more attractive than the correct answer. For each of these questions, we know that the distractor was appealing because the discrimination of the question was quite high, showing that stronger students were much more likely to answer the question correctly than were weaker students. Removing each of these questions individually from the test decreased the Cronbach's α of the test as a whole, showing that these questions did help strengthen the reliability of the test. We have selected a few of these strong questions to discuss in detail.

5.1.1 Question 3. For Question 3 (Figure 7), more students selected Distractor A (122 of 354: 34%) than selected the correct, and somewhat unusual, answer D (119 of 354: 34%). Alternative A is a compelling distractor.

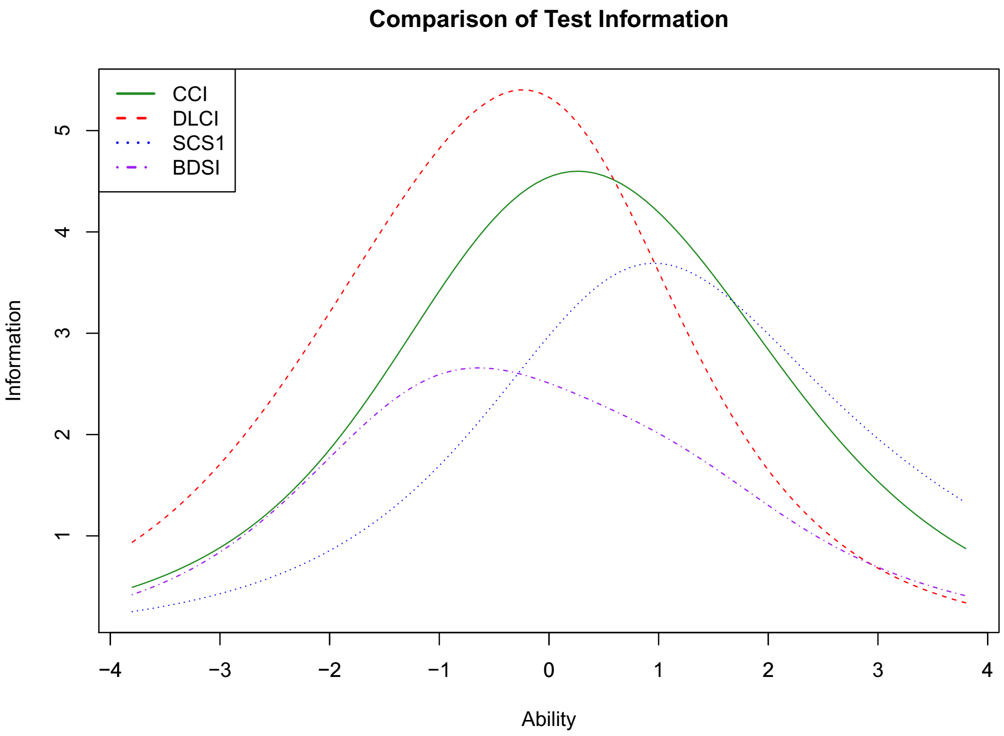


Fig. 6. Information curve for the CCI, compared to the test information curves for other accepted CIs in computer science. Information curves for other CIs calculated from 2PL model fit parameters provided in [16, 30, 45]. The CCI gives a reasonable amount of information about students of all ability levels. The slight skew of the test information curve means that the test gives slightly more information about students who are above average.

Question 3 probes the important adversarial-thinking concept “Identify the attacker.” The scenario explains why Alice might be motivated to lie by denying having sent the purchase order, which supports D, even though usually Alice would not normally intentionally reveal her signing key.

Each of the other alternatives can be excluded as implausible. With a strong signature system: signatures cannot be transferred from one document to another (A); the key cannot be deduced from the signature (B); signatures cannot be forged (C); and distinct documents cannot be found that produce identical signatures (E). Distractor A reflects an egregious misconception about digital signatures. Selecting the correct answer requires adversarial thinking and knowledge of fundamental properties of digital signatures. We conjecture that students who picked A may have engaged in improper analogizing, assuming that digital signatures can be copied in the same way that physical signatures can be copied. This conjecture aligns with prior research findings on students using improper analogies while trying to transfer knowledge from one domain to another, as when they assume the properties of an if-then statement in programming and the if-then construct in Boolean logic to be the same [15].

Scenario. Alice wants to send a file to Bob over an Internet connection. Bob receives a file digitally signed with Alice's private (signature) key, using a secure digital signature algorithm. The file specifies an electronic order to purchase a large number of shares for a new public offering. Contrary to expectation, the value of the stock plummets. Following this incident, Alice denies having signed the purchase order, pointing out that Charlie has been caught forging her signature.
Question. Choose the most likely explanation for how Charlie forged Alice's signature:

A. Copied Alice's digital signature from an older electronic purchase order.
B. Mathematically analyzed Alice's signature to deduce her private key.
C. Changed bits in Alice's signature to sign another electronic document.
D. Received Alice's private key from Alice.
E. Created a new document producing the same digital signature.

Fig. 7. CCI Question 3 probes the concept "Identify the attacker." More subjects selected Distractor A than the correct answer.

5.1.2 Question 23. Question 23, with four appealing distractors, had the highest discrimination (see Figure 8). The correct answer is Alternative B: to disconnect the local network from the internet. Many students, however, selected Alternative D: protecting the network with a state-of-the-art firewall. This explanation aligns with prior research findings from student interviews that students tend to prefer a digital solution over a physical solution, even in situations where the physical solution does a better job solving the cybersecurity problem [42].

Scenario. A law firm stores sensitive client records in a database on their local network.
Question. Choose the action that is the MOST likely to prevent an opposing law firm from reading the records:

A. Require fingerprint scans to access the law offices.
B. Disconnect their local network from the Internet
C. Use only trusted vendor software.
D. Protect the network with a state-of-the-art firewall and intrusion-detection system.
E. Secure the law offices 24/7 with strong locks and security cameras.

Definitions
24/7: Twenty-four hours a day, seven days a week.

Fig. 8. CCI Question 23 probes the concept "Devise a Defense." This question had the highest discrimination of all the questions on the test.

5.2 Using the CCI

We will continue to host the CCI on PrairieLearn [44] for the foreseeable future, and we invite educators to use it for research and to participate in our ongoing evaluation. The authors can forward test results for students who take the CCI through PrairieLearn. The authors are also willing to provide a PDF copy, or provide instructions on how someone might host the CCI through PrairieLearn on their own servers.

5.3 Limitations

Many CIs are poorly suited for use as pre-tests, and the CCI has not been administered as a pre-test. Therefore, we have no evidence for whether or not it can be use as a reliable pre-test. Also, we are unable to comment on the performance of particular demographic groups, because we did not collect this information. Therefore we are not able to address potential biases that questions may have toward individuals of certain demographic groups.

Another limitation of our data set is that the majority of our data come from computer science programs at large public research universities in the United States. We do not have enough data from different types of institutions to comment on differences in student performance based on university, or properties of test questions as applied to students from different kinds of institutions. We believe and hope that the CCI has equally desirable psychometric properties for assessing student knowledge of cybersecurity at a range of institutions types and in a range of degree programs including those in systems administration, business, information technology, and others. The data we have collected thus far, however, do not allow us to reach this conclusion with certainty.

One limitation of the 2PL model which we use is that it does not explicitly model student guessing as the 3PL model does, but allows the effects of guessing to manifest in the difficulty and discrimination parameters. However, due to the design of the questions, the population which we chose to test, and the fact that we discarded any test attempts which took less than 15 minutes, we believe that outright guessing was rare in our data set. If students did guess, they would have been making an educated guess where their chance of answering correctly scaled with their ability, as assumed by the 2PL.

Although there are many other cybersecurity assessments, none are focused on conceptual knowledge or have been previously validated (See Section 2.2). Consequently, we cannot evaluate the quality of our assessment by comparing students' performance on other related metrics.

6 Conclusion

Our psychometric evaluation provides evidence that the CCI is a reliable and valid assessment for classifying the strength of student understanding of basic cybersecurity concepts. Therefore, the CCI can and should be used to compare pedagogic approaches to teaching of cybersecurity.

We plan to apply the CCI to compare the effectiveness of various approaches to teaching and learning cybersecurity. We will also complete our evaluation of a second CI that we developed—the *Cybersecurity Curriculum Assessment (CCA)*, for students completing an undergraduate degree or track in cybersecurity. The CCA targets the same five concepts as does the CCI, but assuming greater technical depth.

CIs are useful tools for promoting change in education through valid and reliable measurement of student knowledge. We have shown that the CCI it is a valid and reliable instrument to measure the cybersecurity knowledge of students who have completed a first course in cybersecurity. We hope that its use will help instructors diagnose the knowledge of their students, and that it will lead to rigorous research in comparing pedagogic practices in cybersecurity education.

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] 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.

[2] 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 (SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 818–824. https://doi.org/10.1145/3328778.3366865

[3] 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 (SIGCSE '20)*. Association for Computing Machinery, New York, NY, USA, 766–772. https://doi.org/10.1145/3328778.3366829

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

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

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

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

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

[9] 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.

[10] 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 (SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 558–564. https://doi.org/10.1145/3287324.3287393

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

[12] Mark Guzdial. 2019. We should stop saying 'language independent.' We don't know how to do that. https://cacm.acm.org/blogs/blog-cacm/238782-we-should-stop-saying-language-independent-we-dont-know-how-to-do-that/fulltext

[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, Lisa Kaczmarczyk, Michael C Loui, and Craig Zilles. 2008. Proof by incomplete enumeration and other logical misconceptions. In *Proceedings of the Fourth International Workshop on Computing Education Research*. 59–70.

[16] 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.

[17] 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

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

[19] 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

[20] 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

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

[22] 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 _eprint: https://doi.org/10.1080/08993408.2018.1562145.

[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] Spencer Offenberger, Geoffrey L Herman, Peter Peterson, Alan T Sherman, Enis Golaszewski, Travis Scheponik, and Linda Oliva. 2019. Initial validation of the cybersecurity concept inventory: pilot testing and expert review. In *2019 IEEE Frontiers in Education Conference (FIE).* IEEE, 1–9.

[26] 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].

[27] 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

[28] 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.

[29] 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.

[30] 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.

[31] R Core Team. 2020. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[32] Dimitris Rizopoulos. 2006. LTM: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software* 17, 5 (2006), 1–25. http://www.jstatsoft.org/v17/i05/

[33] 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.

[34] Travis Scheponik, Alan T Sherman, David DeLatte, Dhananjay Phatak, Linda Oliva, Julia , and Geoffrey L Herman. 2016. How students reason about Cybersecurity concepts. In *2016 IEEE Frontiers in Education Conference (FIE).* IEEE, 1–5.

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

[36] 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.

[37] 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.

[38] 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].

[39] 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).

[40] 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.*

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

[42] 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).

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

[44] 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.

[45] 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

[46] 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.