Initial Validation of the Cybersecurity Concept Inventory: Pilot Testing and Expert Review

Spencer Offenberger, Geoffrey L. Herman
Computer Science
University of Illinois at Urbana-Champaign
Champaign, IL 61820
{so10, glherman}@illinois.edu

Peter Peterson
Computer Science Department
University of Minnesota Deluth (UMD)
Duluth, MN 55812
pahp@d.umn.edu

Alan T Sherman,[1] Enis Golaszewski,[1] Travis Scheponik,[1] Linda Oliva[2]
1 Cyber Defense Lab, Department of Computer Science and Electrical Engineering
2 Department of Education
University of Maryland, Baltimore County (UMBC)
Baltimore, MD 21250
{sherman, golaszewski, tschep1, oliva}@umbc.edu

***Abstract*— We analyze expert review and student performance data to evaluate the validity of the Cybersecurity Concept Inventory (CCI) for assessing student knowledge of core cybersecurity concepts after a first course on the topic. A panel of 12 experts in cybersecurity reviewed the CCI, and 142 students from six different institutions took the CCI as a pilot test. The panel reviewed each item of the CCI and the overwhelming majority rated every item as measuring appropriate cybersecurity knowledge. We administered the CCI to students taking a first cybersecurity course either online or proctored by the course instructor. We applied classical test theory to evaluate the quality of the CCI. This evaluation showed that the CCI is sufficiently reliable for measuring student knowledge of cybersecurity and that the CCI may be too difficult as a whole. We describe the results of the expert review and the pilot test and provide recommendations for the continued improvement of the CCI.**

Keywords*—Cybersecurity education, Cybersecurity Assessment Tools (CATS) Project, Cybersecurity Concept Inventory (CCI), assessment validation.

I. Introduction

The United States is facing an unfulfilled demand for cybersecurity professionals, leaving it ill-prepared for conflicts in cyberspace [1]. With American universities stretched thin by this demand, current proposals have involved outsourcing cybersecurity work and recruiting students at younger ages to be cybersecurity professionals [2]. Because this demand is expected to continue increasing, we need rigorous methods to identify effective ways to educate students in cybersecurity. Creating a valid and broadly used conceptual assessment tool for cybersecurity is a vital resource for supporting rigorous research on the efficacy of various teaching methods for cybersecurity education. Unfortunately, no such validated research instrument exists to assess student conceptual knowledge of cybersecurity.

Sherman et al. began the *Cybersecurity Assessment Tools (CATS)* Project to meet this need for validated research instruments for cybersecurity education [3]–[9]. The CATS Project is developing two *Concept Inventories (CIs)* to evaluate how well teaching practices help students learn core cybersecurity concepts: the *Cybersecurity Concept Inventory (CCI)* and *Cybersecurity Curriculum Assessment (CCA)*. The CCI assesses how well a student has learned the basic concepts of cybersecurity after one cybersecurity course. The CCA assesses how well a student has learned cybersecurity concepts after completing a full cybersecurity curriculum. In this paper, we report on our evaluation of the first draft of the CCI using a panel review by 12 cybersecurity experts and psychometric evaluation of 142 student responses to the CCI. In the pilot study, students taking a first cybersecurity course completed the CCI online or in class proctored by the course instructor.

A. Validity and Concept Inventories

CIs have been applied to show that students regrettably succeed in traditional assessments through fact memorization rather than conceptual knowledge [10]–[12]. With a deeper conceptual knowledge, students learn more efficiently in the future and transfer their knowledge across contexts [12]. CIs have been effectively used to promote the adoption of evidence-based teaching practices across STEM [10], [11], [13].

A CI can be powerful and useful only if it is deemed as a valid assessment tool by the education community that will use the tool. A valid CI effectively evaluates targeted concepts and can be used to draw a reasonable inference of student knowledge [14]. The validity of the instrument is established by a set of evidence and arguments about whether the assessment tool can be appropriately used to draw these inferences. To establish the validity of our assessment tool, we are following the design and evaluation framework recommended by the National Research Council [15], [16].

B. Outline of Paper

We review the development process of the CCI and how that process compares to the development of other CIs. We then describe the framework we use to evaluate whether the CCI can be used validly to assess student knowledge of cybersecurity concepts. We then describe the research methods for the expert panel review and pilot test with students. We analyze the results of this pilot test using *Classical Test Theory (CTT)*. We then discuss these findings to identify the strengths of the CCI and to recommend future improvements for the CCI.

II. Background

The National Research Council recommends establishing a cognitive framework for the design of an assessment tool [16]. This cognitive framework defines what knowledge of a topic should be assessed and the ways in which students reveal their knowledge, or lack of knowledge, about that topic. Prior work on the CATS Project has focused on establishing this cognitive framework, providing baseline arguments for the validity of the CCI.

Because a test cannot be universally valid for every population or use, we need to define carefully the contexts, populations, and uses for which the CCI is valid. We intend the CCI to measure the cybersecurity conceptual knowledge of students who have completed a first course in cybersecurity. Cybersecurity is taught to an increasingly wide range of stakeholders, including policy makers, computer scientists, medical professionals, and business professionals, whose courses vary in focus and depth. Because of this high variance, we have chosen to optimize the CCI for the largest population of cybersecurity professionals—computer scientists. While the CCI may provide useful insights about the conceptual knowledge of policy makers or others, our goal is to have the tool provide the most insight about computer science students.

A. Previous Development of the CCI

In accordance with the recommendations of the National Research Council, we based the design of the CCI on the consensus opinions of a panel of experts [3] and on documented student misconceptions [6].

Parekh et al. [3] began the CATS Project development by identifying the core concepts of cybersecurity using a Delphi process. A Delphi process is a rigorous and structured method for creating consensus among experts about potentially contentious issues, such as what subset of concepts should be included on the CCI [17]. A Delphi process has been used to identify the cognitive framework of several previous CIs [18]. as shown in Table I, this process identified five concepts all related to adversarial thinking to include in the CCI [3]. From these concepts, Sherman et al. [5] developed cybersecurity scenarios that require students to understand these concepts. For example, one scenario explores the concept "Identify attacks against Confidentiality Integrity Authentication (CIA) triad and authentication (C)." It involves a hypothetical government facility where we define defenses and biometric authentication methods, allow questions on potential attacks.

Using these scenarios, Scheponik et al. [4] performed think-aloud interviews to discover student misconceptions and problematic reasoning about cybersecurity [6]. Example forms of problematic reasoning include student beliefs that encryption protects against most any cybersecurity threat and the belief that cybersecurity threats come only from outside an organization.

TABLE I. Five core concepts of cybersecurity

Identify vulnerabilities and failures (V)
Identify attacks against CIA triad and authentication (C)
Devise a defense (D)
Identify the security goals (G)
Identify potential targets and attackers (T)

Using findings from these interviews, we created the CCI multiple-choice questions, called *items*, using the same scenarios and others developed later. Each CCI item comprises a scenario, a stem (i.e., a question about the scenario), and five answer choices. We created the wrong answers (distractors) based on the interview findings. We created five stems for each of the five concepts. By grounding the design of the CCI in the Delphi process and student interviews, we have established baseline arguments for the validity of the CCI.

In this paper, we continue the National Research Council's recommended development process. We use a panel of 12 experts to review whether the draft CCI indeed matches the targeted cognitive framework. Once an assessment tool is created, it should be administered to its targeted demographic and be statistically evaluated [16]. We administered a pilot test of the CCI to a group of 142 students from six universities to evaluate whether students responded to questions on the CCI according to our expectations from the interviews. We use statistical analysis of student responses to determine what inferences can be validly drawn from administrations of the CCI.

B. Classical Test Theory (CTT)

Jorian et al. [19] outline three basic criteria of a valid CI: CI indicates overall understanding of the concepts; CI indicates understanding of a specific concept; and CI indicates misconceptions or student errors. Jorion et al. recommend using a series of statistical tests to demonstrate whether a CI meets these criteria. CTT is often the first evaluation paradigm used to evaluate an instrument because it is useful with smaller sample sizes [20]. CTT is more practical than more exhaustive analytics, such as *Item Response Theory (IRT)*, because CTT allows us to find problematic questions and distractors and suggest modifications with a smaller number of students. This analysis enables more rapid iteration and improvement of the CI.

CTT argues that an assessment tool should minimize error. An assessment should also possess items that all test a single construct (the core concepts), that are neither too hard nor too easy, and that each provide an accurate estimate of a student's overall ability.

C. Reliability

Reliability is a measure of the likelihood that repeated measurements of the same student will yield the same score. If an assessment tool is not reliable, it cannot be valid.

In CTT, the core assumption is that a student's *observed score (X)* consists of two hypothetical values: a student's *true score (T)* and some random *error (E)* [20]. The student's true score would be the score of an infinite number of independent administrations of the test [21]. This model is expressed symbolically as $X = T + E$ [18]. A reliable assessment tool minimizes the error, so the observed score best reflects the student's understanding.

The conventional measurement used for internal reliability is Cronbach's α. Cronbach's α is "an estimate of the correlation between two random samples of items from a universe of items like those in the test" [22]. We can determine Cronbach's α without taking the CCI multiple times if two conditions are met. The conditions are: (1) the assessment tool measures a single trait, (2) each item is either correct or incorrect [18]. A reliable instrument will lead to α values that are close to 1.

There is no universally acceptable Cronbach value, but 0.8 is considered good and 0.7 is the minimum value considered satisfactory, according to Panayiotis [23] and Jorion et al. [19].

The standard error is a function of α and defines a confidence interval for each student's true score. We calculate *standard error* using $SE = S_x\sqrt{1-\alpha}$, where S_x is the standard deviation of the sample and α is Cronbach's α. When the standard error is small, we can be confident students with different observed scores have different true scores.

D. Difficulty and Discrimination

Reliability alone does not indicate the instrument provides a valid representation of student knowledge. The validity of the instrument can be further established by each item's difficulty and discrimination. The *difficulty* of an item is the fraction of students with the correct response [20]. Each item of the instrument should have a balanced range of difficulties falling within 0.2 to 0.8 [18], [19]. When the difficulty is outside this range, it does not effectively separate students of a different understanding.

The *discrimination* of an item is the point-biserial correlation between the item and the overall performance [19]. An item with low discrimination has weaker students (low total scores) perform similarly to stronger students (high total scores) on that item. A good item will have a discrimination of at least 0.2 [18].

E. Topic Agreement and Distractor Analysis

Distractor analysis is used to identify items in which their inclusion does not improve α or has a difficulty and discrimination outside the accepted range. To analyze distractors we partitioned the students into tertiles (thirds) according to total scores. We computed the proportion of test takers selecting each response [20]. There are certain trends we expect to see: (1) The percentage of students selecting the correct answer should increase from the bottom third to the top third, (2) The item's difficulty for the top third of students should be near the upper range of accepted difficulty, (3) Each distractor should have a negative discrimination value [24]. We calculate a *distractor's discrimination* value by setting it as the correct answer and re-grading the student's responses.

F. Concept Subtests

Cronbach's α can be applied to a group of items called a *subtest*. In our case, we propose that there may be five subtests in the CCI, aligning with the five concepts identified in the Delphi process, each consisting of five items designed to cover those concepts. We evaluate these subtests separately to assess reliability to determine whether we can interpret understanding of the concepts from these subtests alone. Ideally, each subtest should have a reliability similar to that of the overall assessment tool. In practice, having a similar reliability to the entire assessment tool is difficult because each subtest has fewer items.

III. Methods

We validated the CCI in two parts. First, experts reviewed and refined the CCI. Second, students took the current CCI as a pilot test.

A. Expert Panel

The *initial CCI* comprised 32 items developed using the processes described in Section IIA. We gave these items to an expert panel for review. The expert panel consisted of 11 instructors with backgrounds in cybersecurity and one cybersecurity professional. Each expert received the initial CCI in the form of an online exam containing each of the items to complete. We asked experts to provide comments and rank each item on the scale: Accept, Accept with Minor Revisions, Accept with Major Revisions, and Reject. After answering, experts were shown the correct answer and given the option to provide additional comments on the correct answer.

We selected 25 items with a range of difficulties based on our best estimation: six easy, 16 medium, and three hard. The actual performance of students would likely differ from our estimations. Each item focuses on one of the five major concepts shown in Table I.

B. Pilot Test

The goal of the pilot test was to administer the current CCI to a small group of 100–200 students and then use the results of this pilot test to refine the instrument. We concluded the pilot test in December 2018 by 142 students from six universities.

Instructors at each university had the option of administering a paper version or online version of the CCI. Both versions included instructions at the beginning of the exam and identical scenarios, questions, and distractors.

The instructor proctored the paper version of the CCI by allocating 50 minutes for students to take the 25-item CCI in class. Students completed the CCI. The instructor collected the exam papers and sent them to us where we recorded all student responses.

If the instructor decided to administer the online version, we provided a link to the exam. The online exam differed from the paper version in three ways. First, the online version had a random ordering of distractors. Second, items that shared a scenario were randomly ordered within that scenario. For example, if Q1 and Q2 are the two items in the one scenario, Q1 can appear before or after Q2 but always together. The reason for randomizing the online version was to dissuade

collusion between students and to minimize any possible effect of item ordering on student performance. Because students who had access were all in the same course, they may have attempted to work together even if they received no benefit from receiving a better score. Third, students were told to spend 50 minutes but this limit was not strictly enforced. Each student completed the exam and then selected a submit button to save and submit their exam.

C. Pilot Demographics

The universities included in the pilot trial have diverse locations and populations. Universities A and D are large Midwestern public universities and have over 40 thousand students enrolled. University E is a large public university from the Southwest with over 40 thousand students enrolled. Universities B, C, F are smaller universities from the Midwestern and Eastern part of the country. These Universities have 10 thousand or less students enrolled.

Table II lists the demographics of the study including institution and number of responses. All universities administered the online exam except for University A.

TABLE II. Breakdown of students by university.

University	Institution Type	Number of Subjects
University A	Large, Midwest, Public	91
University B	Small, East, Public	14
University C	Small, Midwest, Public	1
University D	Large, Midwest, Public	6
University E	Large, Southwest, Public	17
University F	Small, East, Public	12
Not Specified		3
Total		142

IV. Results

We present results from the expert review of the CCI and our psychometric analysis of student responses to the CCI. To help the reader interpret our findings, we compare our results with three CIs evaluated with the same techniques. These CIs are the Concept Assessment Tool for Statics (27 questions and 1,372 students), the Statistics Concept Inventory (38 questions and 402 students), and the Dynamics Concept Inventory (29 questions and 5,966 students) [19]. We chose these CIs because they are the few technical CIs that have been analyzed using similar techniques.

A. Expert Panel

Figure 1 summarizes the results of the expert review process. Although experts reviewed 8 more items, the figure presents the results for the 24 items in the current CCI. Q25 was not finished in time for expert review but was included in the pilot test as it effectively represents topic C. We selected items for the CCI from those that experts reviewed positively receiving a vast majority of Accept and Accept with Minor Revisions.

Additionally, experts wrote comments for each item, which we used to revise the items. For example, we show how we used expert reviews to revise item Q4. Q4 covered a potential SQL injection vulnerability and the means of defending against it. The initial wording of the Q4 scenario is below.

> **Scenario A3.** When a user Mike O'Brien registered a new account for an online shopping site, he was required to provide his username, address, first and last name, and a password. Immediately after Mike submitted his request, you—as the security engineer—receive a database input error message in the logs.

Experts commented that this wording is imprecise because an error in the logs is not something you "receive" but rather written into the log on the server. The word "receive" implies the error was noticeable and could lead students to infer that the error came from the client side. We modified this item by replacing "receive a database input error" with "observe a database input error." The change makes it clear that the user input did not cause an alert, instead logging on the server side. This clarification will lead students away from client side solutions such as "more thoroughly test the software before deploying it" and toward server side solutions such as the correct response, "sanitize input at the server side." The expert review process strengthened clarity, which is critical to measuring a student's conceptual knowledge. Whenever expert review led to a disagreement with another expert, we removed that item from the CCI, or if the item had support from other experts, directly discussed the problem with the experts who had conflicting feedback. We found a resolution for all disagreements; but this outcome may not always be possible when using an expert panel.

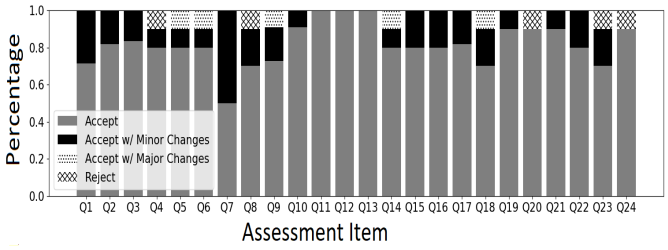

Fig. 1. Expert response to items. Almost all experts approved of all the items.

B. Reliability and Standard Error

The Cronbach's α of the CCI in our pilot test is 0.78. As seen in Table III, this value is close to Jorian et al.'s recommendation for good reliability and above Panayiotis's minimum recommendation. The reliability of the CCI is strong when compared to published values of other CIs. The CCI is sufficiently reliable to be a valid CI.

The standard measurement error of the CCI is 2.13 for our pilot test. A 2.13 standard error implies a 68% confidence interval for a student's true score, given a mean observed score of 8.61 points is from 6.48 to 10.74.

We recalculated Crobach's α with each item excluded. Each item should increase the quality of the instrument indicated by that item's exclusion decreasing the overall reliability. There are no low-quality items that decrease the overall reliability, indicating that each item is reliable enough for inclusion in a valid CI.

TABLE III. Comparison of the quality of the CCI with that of other concept inventories. Recommended ranges for values dervied from [19].

Measurement	CCI	Statics	Statistics	Dynamics	Recommended Value
Cronbach's α	0.78	0.84	0.64	0.74	≥ 0.80
Minimum Difficulty Value	0.10	0.16	0.03	0.06	≥ 0.20
Maximum Difficulty Value	0.66	0.78	0.87	0.91	≤ 0.80
Minimum Discrimination Value	0.16	0.18	-0.13	0.01	≥ 0.20
Maximum Discrimination Value	0.47	0.65	-0.57	0.56	None

TABLE IV. Descriptive statistics of the CCI.

Cronbach's α	0.78
Standard Error of Measurement	2.13
Mean (out of 25)	8.61
Standard Deviation	4.58

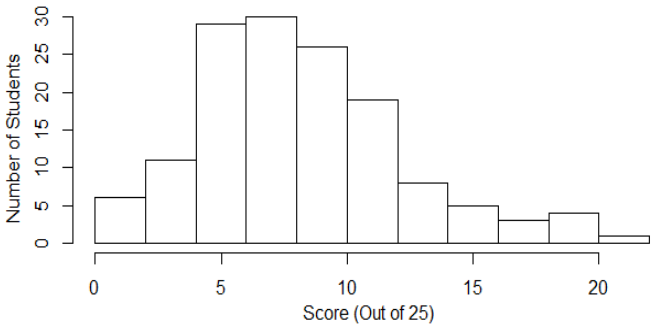

Fig. 2. Histogram of student scores on the CCI. Most students scored between 5 and 10 with mean of 8.61 and standard deviation of 4.58.

C. Difficulty and Discrimination

If an item is too hard or too easy, it cannot effectively differentiate students. Figure 3 shows the acceptable range of difficulty, and Table V shows the difficulty of each item. The range of difficulty for the CCI is 0.10 to 0.66. Figure 2 also shows that majority of the students score within five to ten correct answers. When compared to the other instruments shown in Table III, the CCI is too difficult and will have less discriminatory power.

A high discrimination indicates that a student performance on a given item is highly correlated to overall performance. Figure 3 and Table V show the discrimination for each item. The range of discrimination is 0.16 to 0.47. The discrimination range is not as high as those of other CIs in Table III, but the bottom of the range is encouraging. Those CIs had one, ten, and five items fall below the 0.2 minimum values, compared to the CCI with three items that fall below. Most of the items being above the minimum value is an encouraging indicator for the validity of the instrument.

D. Concept Subtests

We group the individual items within a concept to evaluate the reliability of that concept subtest. Table VI shows the α's of the concept subtests. When evaluating the concepts, it is notable that all of the values are significantly less than 0.7, considered minimal [23]. These findings suggest that the concept subtests of the CCI cannot be validly used as standalone instruments.

TABLE V. Difficulty and discrimination of each CCI item. Difficulty should be between 0.2 and 0.8. Discrimination should be greater than 0.2 [19]. Bolded values indicate potentially problematic values.

Item	Discrimination	Difficulty	Item	Discrimination	Difficulty
Q1	0.21	0.24	*Q14*	0.32	0.25
Q2	0.31	0.33	*Q15*	0.25	**0.10**
Q3	**0.13**	0.26	*Q16*	0.35	0.59
Q4	0.46	0.52	*Q17*	0.35	0.52
Q5	0.35	**0.18**	*Q18*	**0.19**	0.31
Q6	0.23	0.22	*Q19*	0.27	0.28
Q7	0.30	0.66	*Q20*	0.22	**0.14**
Q8	0.21	**0.19**	*Q21*	0.23	0.44
Q9	0.33	0.61	*Q22*	0.47	0.34
Q10	**0.19**	0.40	*Q23*	0.38	0.49
Q11	0.34	0.36	*Q24*	0.30	0.40
Q12	0.36	0.24	*Q25*	0.24	0.14
Q13	0.21	0.28			

TABLE VI. Cronbach's α by concept subtest. Values for all subtests fall below desired thresholds [19].

Subconcept	Cronbach's α	Items Included
V	0.22	Q1, Q3, Q11, Q17, Q21
C	0.45	Q2, Q5, Q14, Q18, Q24
D	0.47	Q4, Q6, Q13, Q19, Q23
G	0.36	Q8, Q9, Q10, Q22, Q25
T	0.50	Q7, Q12, Q15, Q16, Q20

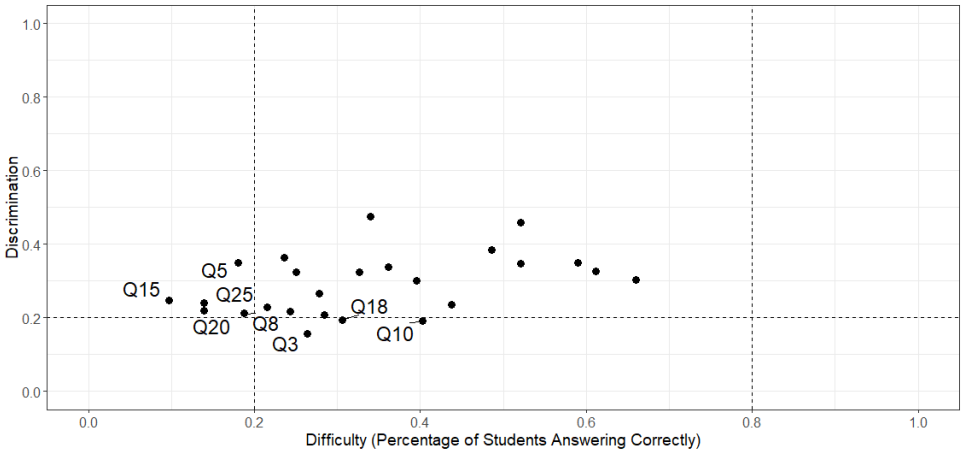

Fig. 3. Difficulty vs. discrimination. CCI items skew towards being too difficult with most of the items toward the bottom of the accepted range.

E. Deeper Analysis of Specific Items

Our psychometric analysis suggests that the instrument has too many difficult items. We analyze the distractor distribution and distractor discrimination to understand why some items are so difficult. We present an example of analyzing and improving one of these items, Q15. Q15 had a low difficulty score of 0.10 (i.e., too difficult) and a relatively low discrimination of 0.25 in the pilot trial. We compare Q15 to a stronger item Q4 which had a moderate difficulty of 0.52 and an acceptable discrimination of 0.46 in the pilot trial.

Table VII shows the distractor analysis for Q15 and Q4. In Q4, which has a desirable distribution, the percentage of students selecting the correct answer increases from the bottom tertile to the top and the top tertile scores near the top of the acceptable range (0.8). Although in Q15 the percentage of students selecting the correct answer increases from bottom tertile to top, there is little separation between the top and middle tertile. The top tertile students answer Q15 correctly 18% of the time and select distractor A 59% of the time. The preference for distractor A among the top tertile is causing the item to be too difficult.

TABLE VII. Example distractor discriminations (An asterisk identifies correct alternative) with regard to tertile scores. Upper students should pick each distractor less than lower students

Q4			
Response	*Lower*	*Middle*	*Upper*
A	0.02	0	0.03
*B	0.28	0.62	0.85
C	0.11	0	0.26
D	0.37	0.14	0
E	0.22	0.24	0.10
blank	0.02	0	0

Q15			
Response	*Lower*	*Middle*	*Upper*
A	0.22	0.36	0.59
B	0.39	0.26	0.08
C	0.15	0.17	0.15
D	0.22	0.05	0
*E	0	0.17	0.18
blank	0.02	0	0

Table VIII shows the discrimination of each distractor for items Q4 and Q15. We expect the distractors to have negative discrimination values. Q4 has negative or zero values for each distractor, as well as a large positive discrimination for the correct answer. Q15 has a large positive discrimination for the correct answer and is above the minimum acceptable value, but distractor A has a larger discrimination value. If distractor A were the correct answer, Q15 would have better discrimination and the item's difficulty would be in the acceptable range.

TABLE VIII. Example distractor discrimination. The correct answer should have positive discrimination. Distractors should have negative discrimination. An asterisk identifies correct alternative.

Q4	
Alternative	*Discrimination*
A	0
*B	0.46
C	-0.14
D	-0.26
E	-0.04

Q15	
Alternative	*Discrimination*
A	0.35
B	-0.19
C	0
D	-0.26
*E	0.24

V. Discussion

Our validation study reveals the instrument could be used to evaluate cybersecurity but would benefit from minor modifications. The CCI has many desirable properties: high reliability and strong expert consensus on the suitability of all items. Unfortunately, our findings reveal a few weaknesses of the CCI as currently constructed: low cohesion for individual concepts, items that are too difficult, and too many difficult items on the instrument.

A. Reliability and Validity

From the results of the pilot trial, the CCI had very high reliability, especially when compared to other CIs. The Cronbach's α is 0.78, which is considered good for a CI. In addition to the instrument's reliability, no items decrease the overall α, indicating that each item measures the same construct of cybersecurity conceptual knowledge [18]. The reliability of the instrument is necessary but not sufficient for the instrument to be valid.

Experts positively reviewed each item and provided feedback to improve the items. In addition to our goal

of covering each of the five core concepts, we considered this feedback to select the 25 items that had the strongest consensus of quality from the experts. The expert reviews provide evidence for the content validity of the CCI: multiple cybersecurity instructors believe that the CCI items represent conceptual knowledge that students should have after a first course in cybersecurity. The content validity provides further evidence for the overall instrument validity.

B. Concept Cohesion

The strengths of the CCI indicate that the collection of items and individual items are well designed from an instructor perspective and reliable from a student performance perspective. The student response data, however, reveal that there is still room for improvement. Notably, while we designed the CCI to assess five concepts, the student performance data did not align well with these five concepts. For example, there is no consistent correlation of the items within each concept. Additionally, the items that evaluate the concepts have low reliability; each α for the individual concept is below 0.5 [19]. Because of the low reliability of the concepts, we cannot recommend using the concept subtests to assess student knowledge of each concept individually.

There are two possible interpretations for this lack of cohesion and reliability within the concept subtests. First, it is possible that the items were poorly designed and do not reflect the core concepts. Second, it is possible that the concepts themselves are poorly bounded, interconnected, or too complex. Given that the expert reviewers did not express any concerns about the content of the items, we argue that the second interpretation is more likely.

Our finding of low cohesion among concept subtests is a common finding among previously published CIs [19]. The commonality of this finding suggests that it is generally difficult for designers of an instrument to design effective concept subtests. While most items may primarily engage students in one concept, the concepts are likely interconnected. Students need to use multiple concepts to answer each item correctly. We believe that this fact may be especially true in cybersecurity, which requires individuals to consider the motivations or capabilities of attackers, constraints or goals of defenders, and the technologies or techniques needed to mitigate risk.

Additionally, the concepts discovered in the Delphi process may be too complex and are really combinations of similar, but separate, concepts [3]. For example, concept "Identify attacks against CIA triad and authentication (C)" involves four unique forms of attack. A confidentiality attack could cover attacking a secure message protocol, and an availability attack could cover a denial-of-service attack. Each example is a form of attack and each is very relevant to cybersecurity. A student may understand mechanisms that enable secure communications yet still have very little idea about denial-of-service attacks. Because each item of the CCI may be multifaceted, creating subtests will be difficult, if not intractable.

If we want to create reliable and valid concept subtests, we may need to consider other models for creating them. For example, we could try narrowing the scope of concept C to just one attribute (e.g., confidentiality). This option may not be desirable because it ignores the complexity of an attacker's varied motivations. Alternatively, we could create multiple instruments that more fully explore each of the five core concepts, but this option would dramatically increase the work and cost of creating assessment tools for cybersecurity. As currently constructed, the CCI provides a reliable instrument for measuring a student's overall understanding of cybersecurity, which is a much-needed first step. Future work can explore which types of future development are needed for creating these subtests.

C. Difficulty

Unlike the alignment of the concepts, an appropriate range of difficulty is often achieved in published CIs and necessary for the instrument to be valid. The CCI is skewed to be too difficult: five items are more difficult than the recommended difficulty, and for 21 out of 25 items, fewer than 50% of students answered each item correctly. This degree of difficulty suggests that some items need to be made easier to improve our ability to distinguish between students with varying abilities and knowledge. Future work on the CCI must explore how to make some items easier to improve the quality of the CCI.

D. Limitations

There are a number of limitations in the pilot trial. The most notable limitation is the depth of analysis performed on the pilot trial results. IRT is not practical with the number of students in the trial. Additionally, we did not perform measurements such as *Confirmatory Factor Analysis (CFA)* and *Exploratory Factor Analysis (EFA)*, because the Cronbach's α for each concept subtest was so low. These limitations are acceptable because this study is a pilot test.

We did not obtain permission to collect institution demographics such the percentage of women or underrepresented minorities enrolled in relevant degree programs. We cannot provide meaningful analysis of how different subpopulations performed on the assessment.

One institution administered the CCI on paper. Due to a lack of statistical power, we are not able to determine the extent different media may have had on student performance.

There were also limitations in the number of students from each university. Ideally, the representation from the different universities would be even so that the results would not be skewed toward University A. The localization may have biased the findings to one university.

E. Future Work: Refining Items

We will take Q15 as a specific example of the type of modification we will make to the difficult items. Less than 10% of students answered Q15 correctly, below what is acceptable for a CI.

The item challenges subjects to find vulnerabilities in a defense and falls under concept "Identify vulnerabilities and failures (V)." The scenario describes a hypothetical nuclear treaty between two countries that requires a method of securely transmitting a message from a monitoring device. Neither country trusts the other, and the design must be fair to each

country. There are certain properties the solution must hold. Each party wants assurance that the message is not modified. Country A wants to ensure that the message originates from the device. Country B wants to monitor the message data in real time. The premise is: "The sender applies a keyed cryptographic hash function to each message using a key distributed only to the sender, Country A, and Country B." Students are expected to find potential vulnerabilities in the suggested outputs of the device.

Option A is the message with a hash of the message and the current time. Options B, C, and D are the key and a hash of the message, the message and hash of the message, and the hash of the message, respectively. Option E, the correct answer, is that the design cannot satisfy the system requirements.

Our distractor analysis revealed that the best students chose distractor A in much the same way that they choose the correct answer for other problems. This finding reveals that as student knowledge increased, this wrong answer choice became more compelling. When constructed well, each item should lead students to pick the correct answer more often as their knowledge increases.

The preference for Option A is understandable given that it is more reasonable than are the other three distractors. Options B and D do not even send the original message, so the message cannot be verified. Option A and Option C do not guarantee that the source sent the message, and since each party has the key, they can modify the message and attach a new hash. Because A has the same structure as C with the addition of time being sent, it appears to be strictly superior to C, making it the best options among the distractors A-D. Students must see the problems with each distractor and select Option E, which serves as a "none of the above." Including a "none of the above" usually makes assessments harder [25], especially with Options A and C satisfying some of the desired properties.

The problem with the item, and "none of the above" in general, is: Option E makes no assertion. This fact leads students to pick the most reasonable of the other choices. We have modified this item, changing Option E to make an assertion: "The design does not work because Countries A and B can modify the message." This edit allows students a definitive assertion to test and come to the same conclusion that the other options do not satisfy the requirements. We anticipate that this change, while minor, will make the item easier and differentiate more students.

After making similar modifications to other items, our next work is to administer the instrument to more students and reanalyze the results. With easier items, the difficulty will cover a better range and better separate students. The range of difficulties and modification of items that are too difficult should increase the discriminatory power of the CCI and improve the CCI's validity and usefulness. Seperately, we are beginning to validate the CCA.

VI. Conclusion

The expert review and pilot testing of the CCI revealed the CCI reliably tests student knowledge of cybersecurity. Currently, we could use the CCI as an evaluation instrument but the scores would be low, reducing the discriminatory power of the CCI. By making the CCI easier, it will be more broadly applicable and provide useful measurements of a broad range of cybersecurity students. Further research will explore the modifications of the items and testing of more students.

VII. Acknowledgments

We thank all of the experts and pilot trial participants. This work was supported 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, and H98230-17-1-0347; and by the National Science Foundation under SFS grant 1241576 and DGE grant 1820531.

REFERENCES

[1] M. Libicki, D. Senty, and J. Pollak, *Hackers Wanted: An Examination of the Cybersecurity Labor Market*. The RAND Corporation, 01 2014.

[2] F. D. Michael Suby, "The 2015 $(ISC)^2$ global information security workforce study," *Frost & Sullivan*, 04 2015.

[3] G. Parekh, D. DeLatte, G. Herman, L. Oliva, D. Phatak, T. Scheponik, and A. T. Sherman, "Identifying core concepts of cybersecurity: Results of two Delphi processes," *IEEE Transactions on Education*, pp. 1–10, 07 2017.

[4] T. Scheponik, A. T. Sherman, D. DeLatte, D. Phatak, L. Oliva, J. Thompson, and G. L. Herman, "How students reason about cybersecurity concepts." *IEEE Frontiers in Education Conference (FIE)*, 10 2016.

[5] A. T. Sherman, D. DeLatte, M. Neary, L. Oliva, D. Phatak, T. Scheponik, G. L. Herman, and J. Thompson, "Cybersecurity: Exploring core concepts through six scenarios." *Cryptologia*, vol. 42, no. 4, 09 2018.

[6] J. Thompson, G. Herman, T. Scheponik, L. Oliva, A. T. Sherman, and E. Golaszewski, "Student misconceptions about cybersecurity concepts: Analysis of think-aloud interviews," *Journal of Cybersecurity Education, Research and Practice*, 07 2018.

[7] A. T. Sherman, L. Oliva, D. DeLatte, E. Golaszewski, M. Neary, K. Patsourakos, D. Phatak, T. Scheponik, G. Herman, and J. Thompson, "Creating a cybersecurity concept inventory: A status report on the CATS project," *2017 National Cyber Summit*, 06 2017.

[8] A. T. Sherman, L. Oliva, E. Golaszewski, D. Phatak, T. Scheponik, G. Herman, D. S. Choi, S. Offenberger, P. Peterson, J. Dykstra, G. Bard, A. Chattopadhyay, F. Sharevski, R. Verma, and R. Vrecenar, "The CATS hackathon: Creating and refining test items for cybersecurity concept inventories," in *IEEE Security and Privacy*, 2019.

[9] S. Offenberger, "Validating a concept inventory for cybersecurity," Master's Thesis, University of Illinois, 306 N Wright St, Urbana, IL 61801, 2019.

[10] R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American Journal of Physics*, vol. 66, 01 1998.

[11] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *The Physics Teacher*, vol. 30, pp. 141–158, 03 1992.

[12] T. Litzinger, P. Van Meter, C. Firetto, L. J. Passmore, C. B. Masters, S. R. Turns, G. L. Gray, F. Costanzo, and S. Zappe, "A cognitive study of problem solving in statics," *Journal of Engineering Education*, vol. 99, pp. 337–353, 10 2010.

[13] D. Evans, G. Gray, S. Krause, J. Martin, C. Midkiff, B. Notaros, M. Pavelich, D. Rancour, T. Reed, P. S. Steif, R. Streveler, and K. Wage, "Progress on concept inventory assessment tools," *IEEE Frontiers in Education Conference (FIE)*, 12 2003.

[14] K. Douglas and S. Purzer, "Validity: Meaning and relevancy in assessment for engineering education research: Assessment validity for engineering education research," *Journal of Engineering Education*, vol. 104, no. 2, pp. 108–118, 04 2015.

[15] J. Libarkin, "Concept inventories in higher education science," in *BOSE Conf*, 2008.

[16] National Research Council Board on Testing, C. f. E. Assessment, Division of Behavioral, Social Sciences, and Education., *Knowing What Students Know: The Science and Design of Educational Assessment*, J. W. Pellegrino, N. Chudowsky, and R. Glaser, Eds. Washington, DC: The National Academies Press, 2001.

[17] B. B. Brown, "Delphi process a methodology used for the elicitation of opinions of experts," 1968.

[18] G. Herman, C. Zilles, and M. C. Loui, "A psychometric evaluation of the Digital Logic Concept Inventory," *Computer Science Education*, vol. 24, pp. 277–303, 10 2014.

[19] N. Jorion, B. Gane, K. James, L. Schroeder, L. V. DiBello, and J. Pellegrino, "An analytic framework for evaluating the validity of concept inventory claims," *Journal of Engineering Education*, vol. 104, pp. 454–496, 10 2015.

[20] J. Ryan and F. Brockmann, *A Practitioners Introduction to Equating with Primers on Classical Test Theory and Item Response Theory*. Distributed by ERIC Clearinghouse, 06 2009.

[21] J. Cappelleri, J. Lundy, and R. Hays, "Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures," *Clinical Therapeutics*, vol. 36, pp. 648–662, 05 2014.

[22] L. J. Cronbach, "Coefficient alpha and internal structure of tests," *Psychometrika*, vol. 16, pp. 297–334, 09 1951.

[23] P. Panayides, "Coefficient alpha: Interpret with caution," *Europes Journal of Psychology*, vol. 9, no. 4, 11 2013.

[24] S. Testa, A. Toscano, and R. Rosato, "Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to blooms taxonomy," *Frontiers in Psychology*, vol. 9, 08 2018.

[25] D. DiBattista, J.-A. Sinnige-Egger, and G. Fortuna, "The none of the above option in multiple-choice testing: An experimental study," *The Journal of Experimental Education*, vol. 82, no. 2, pp. 168–183, 2014.

[26] T. Scheponik, E. Golaszewski, G. Herman, S. Offenberger, L. Oliva, P. A. H. Peterson, and A. T. Sherman, "Investigating crowdsourcing to generate distractors for multiple-choice assessments," *Cyber Summit*, pp. 1–15, 06 2019.