Technical Report 1409

Common Cyber Capabilities Assessment

Cory Adis
Personnel Decisions Research Institutes, LLC

Alexander P. Wind
U.S. Army Research Institute

Michelle Wisecarver
Chelsey Byrd
Jaclyn Martin
Lia Engelsted
Personnel Decisions Research Institutes, LLC

Kristophor G. Canali
Colin L. Omori
U.S. Army Research

September 2022

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved:

MICHELLE L. ZBYLUT, Ph.D.
Director

Research accomplished under contract
for the Department of the Army

Personnel Decisions Research Institutes, LLC

Technical review by

Elyssa Johnson, U. S. Army Research Institute
Kimberly Wingert, Consortium of Universities of Washington

DISTRIBUTION

This Technical Report has been submitted to the
Defense Technical Information Center (DTIC).

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
September 2022	Final	January 2018 – January 2019

4. TITLE AND SUBTITLE
Common Cyber Capabilities Assessment

5a. CONTRACT/GRANT NUMBER
W911NF-18-C-0014

5b. PROGRAM ELEMENT NUMBER
633007

6. AUTHOR(S)
Adis, Cory, Wind, Alexander P., Wisecarver, Michelle, Byrd, Chelsy, Martin, Jaclyn, Engelsted, Lia, Canali, Kristophor G., & Omori, Colin L.

5c. PROJECT NUMBER
A792

5d. TASK NUMBER
1008

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Personnel Decisions Research Institutes, LLC
1911 N. Fort Myer Drive, Suite 410
Arlington, VA 22209

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute
for the Behavioral and Social Sciences
6000 6th Street (Bldg. 1464 / Mail Stop: 5610)
Fort Belvoir, Virginia 22060-5610

10. SPONSOR/MONITOR'S ACRONYM(S)
ARI

11. SPONSORING/MONITORING
Technical Report 1409

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

13. SUPPLEMENTARY NOTES
ARI Research POC: Colin Omori, Selection and Assignment Research Unit

14. ABSTRACT

As cyber jobs in the Army grow in importance, it becomes essential to have effective selection and classification processes and tools in place. These can ensure that the Army identifies personnel with the aptitude to succeed and assigns them to appropriate jobs. While tests are available that identify personnel who already have key knowledge needed for cyber jobs, there are currently no tests that can identify service members who have the potential to do well in cyber jobs but do not have existing cyber knowledge. The current research developed an innovative assessment called the Common Cyber Capabilities (C^3) Test to measure seven constructs that were identified in previous job analysis research as important to success in cyber jobs: Active Learning, Complex Problem Solving, Critical Thinking, Deductive Reasoning, Inductive Reasoning, Selective Attention, and Troubleshooting. Preliminary data was collected from a sample of 67 subjects who completed the C^3 Test, a small number of related measures, a demographics questionnaire, and a feedback questionnaire. The operationalization of the constructs, validation results, and recommended next steps are discussed.

15. SUBJECT TERMS
Common cyber capabilities, Cyber assessment, Complex problem solving, Active learning, Critical thinking, Deductive reasoning, Inductive reasoning, Selective attention

SECURITY CLASSIFICATION OF:			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE	Unlimited		Tonia S. Heffner
Unclassified	Unclassified	Unclassified	Unclassified	111	703-545-4408

Standard Form 298 (Rev. 8-98)

Technical Report 1409

Common Cyber Capabilities Assessment

Cory Adis
Personnel Decisions Research Institutes, LLC

Alexander P. Wind
U.S. Army Research Institute

Michelle Wisecarver
Chelsey Byrd
Jaclyn Martin
Lia Engelsted
Personnel Decisions Research Institutes, LLC

Kristophor G. Canali
Colin L. Omori
U.S. Army Research Institute

Selection and Assignment Research Unit
Tonia S. Heffner, Chief

September 2022

EXECUTIVE SUMMARY

Research Requirement:

Given the growing importance of cyber jobs in the Army, as well as the uniqueness of the capabilities required for cyber jobs compared with other warfighter jobs, it becomes increasingly important to have an effective selection and classification system in place to identify and assign recruits who have a high potential to succeed in these jobs. In this project, we developed an innovative assessment called the Common Cyber Capabilities (C^3) Test to measure seven constructs identified as relevant to success across multiple cyber jobs in the Army: Active Learning, Complex Problem Solving, Critical Thinking, Deductive Reasoning, Inductive Reasoning, Selective Attention, and Troubleshooting.

Procedure:

In order to identify recruits with a high potential to succeed in these jobs who do not currently possess cyber knowledge and skills, one key stipulation for the C^3 Test was to create an assessment that did not require cyber or information technology (IT) knowledge. Because constructs such as Complex Problem Solving cannot be completely context-free, test takers needed problems to solve and information they could draw on to solve the problems. In addition, in order to measure Active Learning, test takers would need an opportunity to learn. In order to accommodate these needs, a fictitious context was created for the C^3 Test in which the test taker is starting a new job and is provided with information about the job. In starting their new job, test takers progress through a "learning phase" and an "application phase." This two-phase structure enables test takers to learn relevant information in the first phase of the assessment, then apply that information to solve problems in the second phase of the assessment.

In the C^3 Test, the test taker assumes the role of a newly hired employee at a fictitious futuristic transportation company that uses vacuum tubes for transport. In the assessment, the test taker must first learn job-relevant knowledge through a series of training modules that describe the company, the development of tube travel, and specific components and equipment related to their fictitious job. The test taker then begins the new job and applies this knowledge to problems that emerge in the application phase of the test. Within this overall assessment context, each of the C^3 constructs were operationalized based on the definition of the construct and a review of existing measures. Preliminary data was collected from a sample of 67 subjects who completed the C^3 Test, a small number of related measures, a demographics questionnaire, and a feedback questionnaire.

Findings:

Initial psychometric evidence for the C^3 Test was promising. Most measures had sufficient variance and scores had distributions that were approximately normal. While some distributions were slightly skewed, suggesting measures that were somewhat too easy or too hard for participants, the causes of these characteristics were easily identified and can be fixed in

subsequent versions of the battery. A few measures showed central tendency bias in their distributions and would benefit from procedures to increase the variance.

Many of the correlations among the C^3 constructs and subdimensions showed expected patterns. Though it was not possible to collect construct validity data for all measures in this initial study, the correlations among the C^3 constructs and subdimensions provided some confirmation of construct validity. Most intercorrelations were significant, but small to medium in magnitude. The fact that the correlations were small to medium suggests that no two instruments were measuring the same construct. Since each of the C^3 tests were designed to measure a distinct construct, this provides initial evidence of discriminant validity. Additional research should be conducted to examine the divergence of these measures from other similar measures more closely. Correlations between the C^3 constructs and a general intelligence test were also small to medium, suggesting divergence from general intelligence. Future research should examine the criterion-related validity of the C^3 Test and the extent to which C^3 constructs can predict criteria above and beyond general intelligence.

Although test takers reported finding the test complex and challenging to complete, this is, in part, necessary for a test that focuses primarily on assessing cognitive skills. Test taker feedback and the specific results for each construct are presented and discussed in detail.

Utilization and Dissemination of Findings:

Findings of this research can be used to improve the C^3 Test and prepare for research to evaluate the criterion-related validity of the assessments. Once validated, the C^3 Test can be used to identify recruits and new Soldiers who have abilities that are key to success in cyber-related jobs, even if they do not have pre-existing cyber knowledge and skills. In addition, to the extent that other Army jobs require similar abilities to those on the C^3 Test, the test can be used to identify candidates for other Army jobs, as well.

CONTENTS

PAGE

PAGE

APPENDICES

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

COMMON CYBER CAPABILITIES ASSESSMENT

CHAPTER 1: INTRODUCTION

U.S. Cyber Command attained initial operational capability in May 2010, and in October 2010, the U.S. Army created the U.S. Army Cyber Command (ARCYBER) from the inactive Second U.S. Army. Cyberspace is the fifth domain of warfare, joining land, sea, air, and space. It presents a complex challenge that interconnects and intersects the aforementioned physical domains with information and electronic systems (e.g., Joint Publication 3-12). As the complexity of the hardware and software used for storing and transmitting information increases, personnel in these fields will continue to face growing challenges. Cyber and information professionals need to be able to handle the growing complexity of their jobs, while staying up-to-date with current trends in technology and responding to adaptations in adversaries' techniques and tactics. Given the growing importance and scale of cyber jobs in the Army, as well as the uniqueness of the capabilities required for cyber jobs compared with other warfighter jobs, it is increasingly important to have an effective selection and classification system in place to identify and assign recruits who have a high potential to succeed in these jobs.

Cyber talent assessment is not uniformly conducted across job series, though all use aptitude scores from the Armed Services Vocational Aptitude Battery (ASVAB) and the Armed Forces Classification Test (AFCT). Knowledge tests are another assessment tool. For example, the Cyber Test (Trippe et al., 2014) is currently used by the Air Force and Army to predict perceived job fit (Trippe et al., 2017). Other cyber knowledge tests have been developed for the Army to aid with classification and placement (Trippe et al., 2019). However, the items in technologically-based knowledge tests are susceptible to obsolescence as technology advances. Moreover, cyber knowledge tests can only identify the individuals with training and experience in the cyber knowledge domain, leaving out those with the potential, but not the knowledge. Another strategy is to measure the skills, abilities, and other characteristics that are predictive of success across the multiple cyber positions, without requiring domain knowledge. The Common Cyber Capabilities Test (C^3) developed in this project aims to test skills and abilities that are not covered by extant Army measures and are predictive of trainability for cyber roles. Research suggests that having an assessment that is tailored for the specific skills and abilities required in a job achieves strong validity for personnel selection situations (Schmidt & Hunter, 1998).

Identifying Cyber Selection Constructs

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) conducted initial research to identify the skills, abilities, and other characteristics that were associated with success across multiple cyber military occupational specialties (MOS). Constructs were selected and defined for inclusion in C^3 based on three sources of information: a review of existing cyber job information, results of an Army cyber job analysis (Wind, 2018), and discussions with Army cyber subject matter experts (SMEs).

Review of Existing Job Information

The government-sponsored national database known as Occupational Information Network, or O*NET, provides a tool called the O*NET Content Model which describes the knowledge, skills, and abilities that are required for various occupations (O*NET Resource Center, 2015).

While O*NET does not include specific Army MOS in their list of occupations, it does include civilian cyber jobs. Schmidt et al. (2015) compared civilian private-sector cyber jobs and cyber jobs in the U.S. Air Force and found high similarity between civilian and military cyber professional responsibilities, particularly in information technology (IT) operations and information security. O*NET identified five skills (complex problem solving, critical thinking, judgment and decision making, monitoring, and reading comprehension) and six abilities (deductive reasoning, inductive reasoning, information ordering, oral expression, problem sensitivity, and written expression) as important for civilian cyber jobs.

Another source of information about skills and abilities needed for cyber jobs came from Trippe et al. (2014), who interviewed 72 cyber/IT SMEs from the Air Force (31), Army (3), and Navy (38) in constructing the Information and Communication Technology Literacy test (ICTL). They used two well-known individual differences taxonomies (Carroll, 1993; Fleishman et al., 1999) and interviews with SMEs to identify abilities potentially important for IT and cyber-related occupations. The 12 abilities they identified are: verbal reasoning, nonverbal reasoning, mathematical reasoning, problem sensitivity, originality, information ordering, written communication, oral comprehension, perceptual speed, advanced written comprehension, written expression, and near vision. There is a strong congruency between the O*NET list and the one from Trippe et al. (2014), although that may be due, in part, to the use of the same taxonomies (i.e., Carroll, 1993 and Fleishman et al., 1999).

Army Cyber Job Analysis Results

In order to identify the relevance of various skills and abilities specific to military cyber roles, Wind (2018) had 62 Army cyber SMEs rate the importance of skills and abilities for the following Army jobs: 17C - Cyber Operations Specialists (including offensive operators, defensive operators, and instructors), 25B - Information Technology Specialists, 25D - Cyber Network Defenders, 25N - Nodal Network Systems Operator- Maintainers, 25U - Signal Support Specialists, 35L - Counterintelligence Agents (cyber focus), 35Q - Cryptologic Network Warfare Specialists, 94D - Air Traffic Control Equipment Repairers, and 94F- Computer Detection Systems Repairers. SMEs were asked to use a 10-point scale to indicate how essential each skill or ability was for performing the duties of and excelling in their MOS. SMEs then engaged in a consensus discussion within each MOS to create a final list of the skills and abilities ranked in order from most essential to least essential. Results indicated that there were similarities across Army cyber MOS in the capabilities that were required. Only one capability (Critical Thinking) was ranked within the top 10 important attributes for every MOS. Critical Thinking also had the highest mean ranking value. Based on the results of the rankings, Critical Thinking and six other constructs were selected for inclusion in the C^3 test: Active Learning, Complex Problem Solving, Deductive Reasoning, Inductive Reasoning, Selective Attention, and Troubleshooting. The first two columns of Table 1 list the selected constructs and their O*NET definitions.

Table 1

Common Cyber Capability Constructs and Descriptions

Construct	O*NET Online Definition	Descriptions from SMEs
Active Learning	Understanding the implications of new information for both current and future problem-solving and decision-making	• When you encounter something you don't know, need the drive to learn it for next time
Complex Problem Solving	Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions	• Need to solve problems when things go wrong; use reasoning • Look at a situation, figure out the issue, accomplish the mission • Understand the effects of decisions and explain reasoning • Look deeply at problems to find the root cause
Critical Thinking	Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems	• Judge credibility of information • Checking your assumptions • Assessing risks
Deductive Reasoning	The ability to apply general rules to specific problems to produce answers that make sense	• None described
Inductive Reasoning	The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)	• Need to be able to piece things together
Selective Attention	Concentrating on a task over a period of time without being distracted	• Stay on track while stressed • Concentrate on a task and complete it in timely manner • Get oneself back on track when you get off task • Attention may need to be focused on several things
Troubleshooting	Determining causes of operating errors and deciding what to do about it	• Walking through steps to fix something

Discussions with Army Cyber SMEs

Discussions were held with Army cyber SMEs to gather contextual information about situations in which the seven C^3 constructs were required on the job. Meetings were held with representatives from Signal branch (career military field, or CMF 25) and Cyber Operations Specialists (MOS 17C). Representatives from CMF 25 included a civilian and six senior NCOs, most of which held multiple 25-series jobs during their careers, so they were very familiar with the requirements. The 17C group consisted of five soldiers from a greater range of ranks. Because the 17C MOS was very new at the time of the meetings, most of the more senior participants in the 17C focus group had previously held jobs in CMF 25. Two from the 17C group had less on-the-job experience; however, they were currently in or had recently completed training for 17C and contributed valuable perspective on the requirements for success in cyber training.

These discussions provided a deeper understanding of how the constructs and definitions specifically applied to cyber jobs and how the constructs are expressed on the job by high performers. The SMEs reviewed the seven C^3 constructs, providing additional details about common situations and problems encountered in the cyber field in which the constructs would be applied. These comments can be seen in the third column of Table 1. Selective Attention and Active Learning were described by SMEs as extremely important to success in cyber training and in keeping up with technology developments over the course of one's career. Troubleshooting and Complex Problem Solving were identified as particularly important for success in an environment that was often chaotic and changing. Critical Thinking, Deductive Reasoning, and Inductive Reasoning were described as general capabilities that were less specific to cyber but important as foundational skills. Soldiers who lack these logic skills have difficulty completing cyber training.

Conceptual Design of the Test Battery

The input from SMEs was used to refine our understanding of cyber requirements and create a conceptual design for the C^3 assessments. One key stipulation for the C^3 test was to create an operational assessment that did not require cyber or IT knowledge. Since constructs such as Complex Problem Solving require having some type of information and context in order to solve a problem, we needed to provide test takers with problems to solve and information they could draw on to solve the problems. In addition, in order to measure Active Learning, test takers would need an opportunity to learn, so a knowledge domain would need to be developed and incorporated into the battery.

In order to accommodate these needs, we created a fictitious context for the C^3 test in which the test taker is starting a new job and is provided with information about the job. To mirror the information-rich context associated with cyber performance, we designed the fictitious context such that test takers were required to sort through data and information to find the relevant information for their tasks. Importantly, understanding the fictitious context could not require any cyber knowledge. In starting their new job, test takers would progress through a "learning phase" and then an "application phase." This two-phase structure enables test takers to learn relevant information in the first phase of the assessment, then apply that information to solve problems in the second phase of the assessment.

The test taker assumes the role of a newly-hired employee at a fictitious futuristic transportation company. The assessment setting is approximately 100 years in the future and uses fictional technologies for human transportation through the use of vacuum tubes. The setting was specifically designed to (a) be free of any requirement for cyber or IT knowledge, and (b) use an unknown setting and technology to prevent any test taker from having the advantage of existing knowledge about the context. In the assessment, the test taker must first learn job relevant knowledge through a series of training modules that describe the company, the development of tube travel, and specific components and equipment related to their fictitious job. The test taker then begins their new job and applies this knowledge to problems that emerge in the application phase of the test.

Within this overall assessment context, each C^3 construct was operationalized based on the definition of the construct and a review of existing measures. We next describe the operationalization of each of the seven constructs.

Individual Construct Measures

Active Learning

Active Learning is defined as "understanding the implications of new information for both current and future problem-solving and decision-making" (O*NET Online, n.d.). Active Learning can therefore be a crucial ingredient to problem solving success, especially when problem solving is complex and requires searching for or sharing information. Cyber SMEs emphasized the need to constantly learn new technologies and filter incoming information to determine what might be useful for the future.

Despite the conceptual importance of Active Learning for problem solving, it is a new concept and has not been investigated extensively. As such, there is not a unified framework in the literature for its measurement, and little is known about how to measure it. Related constructs that were identified in the literature include Continuous Learning Orientation (Coetzee, 2014; Kim, Kim, & Bilir, 2014), Career-related Continuous Learning (Rowold & Kauffeld, 2008), Self-regulated Learning (Pintrich, 2000; Schuitema, Peetsma, & van der Veen, 2016), and Intrinsic/Extrinsic Motivation to Learn (Benware & Deci, 1984; Hsia, Huang, & Hwang, 2016; Hwang, Yang, & Wang, 2013). While conceptually related, none of these constructs focuses on measuring the test taker's understanding of how new information applies to Problem Solving. In addition, some constructs, such as Career-related Continuous Learning (Rowold & Kauffeld, 2008), refer specifically to behaviors in a workplace setting, which is not relevant to our population of interest – young adults who have recently graduated from high school.

Other constructs were developed for an appropriate population but focus on different aspects of learning. Self-regulated Learning (Schuitema et al., 2016), for example, focuses on the regulation of the learning process, such as strategies to complete homework, rather than on understanding the implications of new information for Problem-solving and Decision-making. Still other measures focus on depth of learning (e.g., Gibbs & Coffey, 2004), Metacognition for Learning (e.g., Berger & Karabenick, 2016), or other aspects of Metacognition (Flavell, 1979; Schraw & Dennison, 1994; Taasoobshirazi & Farley, 2013). The Student Engagement Questionnaire (Kember & Leung, 2009) captures student perceptions of university teaching and learning

environments, and a number of measures capture students' motivation to learn in specific contexts (e.g., Hwang & Chang, 2011; Hwang, Yang, & Wang, 2013).

Most of the identified constructs were measured using self-report questionnaires (e.g., Coetzee, 2014; Kim et al., 2014; Rowold & Kauffeld, 2008; Schuitema et al., 2016), which are more susceptible to faking and tend to be more trait-based rather than ability-based. The project team decided that measures that focus on behaviorally-based approaches, such as capturing the frequency of engaging in certain learning behaviors (e.g., the frequency of continuous learning behaviors; McAbee, Oswald, & Connelly, 2014; Saunders-Stewart, Gyles, Shore, & Bracewell, 2015) could provide a better model for our measure. Furthermore, it was decided that focusing on specific relevant behaviors would be a useful approach if we could identify a series of behaviors that were associated with AL. Research in educational settings has identified a variety of behaviors that are associated with Active Learning, such as: holding discussions, knowledge exams, participating in study groups, and engaging in higher-order thinking such as forethought, monitoring, control, and reflection (Braxton, Milem, & Sullivan, 2000; Liu, 2017). Many of these behaviors, however, were specifically identified for learning in educational settings.

Given that we were developing a novel assessment context that includes one assessment phase dedicated to learning and another dedicated to applying that knowledge to solve problems, we sought to create an AL measure that would capture test takers' Active Learning actions within the simulation. Two approaches were developed for Active Learning in the C^3 test battery to reflect whether test takers understood the implications of new information for solving current and future problems: one measure focused on whether test takers could identify information that was potentially important to learn ("Identifying Important Information"), and the other focused on the effectiveness of their learning ("Learning Effectiveness").

Identifying Important Information (III) measures behaviors related to recognizing information that is likely to be important for use later while "on the job." Throughout both the learning and application phases of the assessment, the test taker was interrupted during their activities and presented with a page containing sets of facts from earlier in the assessment. Test takers were instructed to select 1-4 fact sets from the page that were the most important for someone in their position to know in order to be successful on the job. Each fact set was weighted by its actual importance, and test takers received scores based on the sum of the importance weights for the selected fact set. Some fact sets were negatively weighted, so selecting more was not always the best choice.

The Learning Effectiveness (LE) measure reflected an ability to recognize and retain important information during the assessment. For this assessment, knowledge tests were administered at the end of the learning and application phases (LE1 and LE2, respectively). These tests assessed knowledge of important material that was encountered while in training (LE1) or while engaged in work tasks such as troubleshooting or problem solving (LE2). All content for LE1 and LE2 was distinct from content used in III.

Complex Problem Solving

Problem Solving is a multistage process that involves a number of elements such as identifying the problem, developing a solution strategy, and monitoring progress toward a goal (e.g., Pretz,

Naples, & Sternberg, 2003). *Complex* Problem Solving refers to engaging in problem solving when the problems are novel, ill-defined problems such as those in complex, real-world settings (O*NET, n.d.). Our goal in creating an assessment for Complex Problem Solving was to identify individuals who are likely to succeed in engaging in Problem Solving on the job when the problems are novel and ill-defined.

Much of the existing research on Complex Problem Solving has used micro-world simulations to measure Complex Problem Solving performance. A micro-world simulation is a useful tool for Complex Problem Solving because it can be used to create and administer a complex problem to a test taker. The test taker's responses and actions in the problem situation are captured and the simulation delivers summary outcomes reflecting performance with respect to the presented problem. Early micro-worlds used business contexts such as a shirt factory (Dorner, 1980; Funke, 2010; Putz-Osterloh, 1981 as cited in Funke, 2010) in which the test taker controls variables such as employee wages, merchandise prices, and maintenance schedules, with the goal of maximizing outcomes such as profits (Danner et al., 2011). Micro-worlds require that problem solvers acquire and apply knowledge in order to build a representation of the problem and solve it.

One drawback of the early Complex Problem Solving micro-world simulations was that they only used one scenario, which prevented calculation of reliability (see Kroner, Plass, & Leutner, 2005; Wustenberg, Greiff, & Funke, 2012), and created scenario-specific effects that made it difficult to compare outcomes across different types of scenarios and simulations (Greiff et al., 2015). More recent micro-world tests of Complex Problem Solving rely on an underlying framework, which allows for easier comparisons of results across situations. Funke (2001) introduced two frameworks for scaffolding Complex Problem Solving micro-worlds that enable the underlying task structures in the micro-worlds to be independent of their larger context (Greiff, Fischer, Wustenberg, Sonnleitner, Brunner, & Martin, 2013). In addition, newer Complex Problem Solving tasks have introduced multiple scenarios in a multiple complex system approach (Fischer, Greiff, & Funke, 2012). This approach shortens the length of each scenario from about 45 minutes in earlier micro-worlds to about five minutes per scenario, with 8-12 different scenarios. This provides multiple assessment items for Complex Problem Solving, enabling a broader assessment as well as calculation of reliabilities (though still requiring 40-60 minutes to measure). While using micro-worlds holds promise as a measurement tool for Complex Problem Solving, research is still needed in areas such as understanding item difficulty and demonstrating criterion validity (e.g., Greiff & Funke, 2009). Another added challenge is that the cost of developing a micro-world simulation may limit its feasibility.

Although there are numerous problem-solving measurement paradigms, the micro-world simulations are the only assessment tools for Complex Problem Solving that have been developed. In order to develop an assessment for Complex Problem Solving, we started with the foundational definition provided by O*NET (see Table 1), which indicates that Complex Problem Solving involves "identifying complex problems and reviewing related information to develop and evaluate options and implement solutions*"* (O*NET Online, n.d.). Several key elements or stages of Complex Problem Solving are apparent from this definition: (1) identifying the problem, (2) reviewing information related to the problem, (3) coming up with and choosing solutions, and (4) putting solutions into action. Each of these stages represents a dimension of Complex Problem Solving that could be measured. In order to limit the assessment time and

scope, we decided to limit our Complex Problem Solving measurement to the two dimensions that would be overtly observable and measurable using automatic scoring: (1) identifying the problem, which we label "Investigating Complex Problems (ICP)," and (2) choosing solutions, which we label "Evaluating Options (EO)."

Both ICP and EO were measured in the application phase of the assessment. The test taker was told that they have completed initial training and they are starting on the job. Information was provided about a possible problem in the tube transport system and the test taker was asked to investigate using the system information tools they are provided. After spending time searching for relevant information in the various screens to which they had access, test takers were given a list of statements that were potentially related to the problem and asked to identify whether the statements are true or false. Test takers who scored highly on ICP had uncovered more information to understand the problem than those who scored poorly on ICP. In addition, some of the test taker's investigative actions were captured, including keywords they used to search for information and the database pages they opened to investigate. Across the application phase, three problem scenarios were presented to the test taker with a set of ICP statements following their investigation of each situation.

After providing responses to the first set of ICP items, test takers were given the correct answers. This feedback was given because it will enable all test takers to start the EO assessment with the same level of accurate information. This served to reduce the dependence of EO scores on the outcomes of ICP.

EO was then measured using a situational judgement test (SJT) approach. After receiving feedback on the ICP items, the test taker was given a list of possible courses of action and asked to rate the effectiveness of each option toward the goal of identifying a solution to the problem. The test takers made these judgements based on the information they learned in their investigations and training (or from ICP feedback). Test taker ratings of each statement were then compared to SME ratings to determine scoring for the EO items.

To summarize, there were three Complex Problem Solving scenarios, and for each Complex Problem Solving scenario test takers went through the following sequence:

Step 1: Receive initial information (i.e., prompt) regarding a problem
Step 2: Investigate using information sources provided (including using search terms to search for information)
Step 3: Answer ICP items
Step 4: Receive feedback on ICP items
Step 5: Rate EO options
Step 6: Receive new information regarding the same situation
Step 7: Repeat steps 2-6 three more times for each scenario

This approach to assessing Complex Problem Solving captured information regarding two critical elements of Complex Problem Solving: the test taker's ability to investigate problems and to evaluate various problem-solving options.

Critical Thinking

Critical Thinking is defined as "using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems" (O*NET Online, n.d.). The most common approach to assess Critical Thinking is the use of SJTs. In SJTs, test takers read scenarios and respond to a series of questions. The questions evaluate the test takers' ability to identify strengths and weaknesses of the information, analyze arguments presented, draw logical conclusions, and recognize errors of logic. For example, in the Psychological Critical Thinking Exam, test takers review a set of conclusions drawn about a scenario and identify problems with the conclusions (Lawson, Jordan-Fleming, & Bodle, 2015). Test takers are assessed on their ability to identify whether there is a problem and to accurately identify the problem. Similarly, in the Brief Assessment of Critical Thinking, respondents read two argument statements from a debate, and then choose from a selection of critiques of the arguments (Jessop & Adams, 2016). Related styles include tests in which participants read vignettes and/or statements and answer multiple questions to identify likely conclusions or follow-up arguments (e.g., Behrens, 1996; Wagner & Harvey, 2006). A somewhat different style of SJT uses "everyday" situations or scenarios to assess individuals' Critical Thinking (e.g., Butler, 2012; Halpern, 2010). For example, a behavioral assessment described by Butler (2012) presents respondents with item sets consisting of everyday negative events that could happen (e.g. "I threw out food because it went bad") and has them identify the decisions that could logically precede them.

One debate in the literature on Critical Thinking surrounded the concept of using open-ended responses to questions (e.g., Newman, Webb, & Cochrane, 1995). For example, the Ennis-Weir Critical Thinking Essay Test requires respondents to write a highly-structured essay evaluating a hypothetical argument test (Davidson & Dunham, 1996; Ennis, 2003), and The Halpern Critical Thinking Assessment (Halpern, 2010) asks test takers to read 25 "everyday" scenarios and assesses the various facets of Critical Thinking using both open-ended responses and multiple choice answers. Taube (1997) suggests that open-ended responses are more likely to capture a respondents' inclination to engage in Critical Thinking behavior (i.e., their disposition towards Critical Thinking) rather than Critical Thinking skills. While both Critical Thinking skill and Critical Thinking disposition are relevant to Critical Thinking (Newman, Webb, & Cochrane, 1995), thought should be given to whether one facet or the other are more relevant in a given situation and consider the implication of the assessment method on the information that is captured. Disposition for Critical Thinking has also been measured with personality-type assessments in which participants rate their level of agreement with a number of statements of beliefs, values, or expectations (e.g., California Critical Thinking Disposition Inventory; Facione, 2000).

For the C^3 assessment of Critical Thinking, we were interested in identifying individuals who are likely to succeed in cyber jobs that require Critical Thinking, so both disposition and skill would potentially be relevant. Given the general success of the SJT approaches and the impracticality of scoring open-ended response options with a large number of potential test takers, we chose to develop an assessment using an SJT approach. While engaged in the learning phase of the C^3 assessment, test takers are presented with written information from newspaper articles and/or discussion among fellow trainees. In order to integrate the measure into the larger tube transport assessment context, the written information describes a situation or incident

involving the tube transport industry. After reading the information provided, the test taker must answer questions about the information by thinking critically about the available information.

The Critical Thinking questions were designed to prompt test takers to use four of the six Critical Thinking skills dimensions identified by Facione (1990): analysis, explanation, evaluation, or interpretation. Analysis questions asked the test taker to indicate the extent to which they agreed with a series of analysis statements, given the content of the information provided. Evaluation questions asked test takers to indicate the extent to which they agreed with a series of evaluation statements, given the information provided. Explanation questions asked how justified a set of statements were, given the information in the reading. Finally, interpretation questions focused on how important various pieces of information were to a given objective. The two dimensions identified by Facione (1990) that were omitted were self-regulation and inference. Self-regulation was not included because of the challenge in identifying objective, observable indicators of this underlying mental behavior and distinguishing it from the other Critical Thinking dimensions. This is not to minimize the importance of self-regulation; future research should look for alternative ways to measure this aspect of the Critical Thinking construct. Inference was omitted because the C^3 battery already includes measures of inductive and deductive reasoning.

Ratings provided by test takers for each statement were compared to ratings made by four industrial/organizational (I/O) psychologist SMEs who were familiar with the Critical Thinking scenarios. The distance of the test takers' ratings from the SMEs' average rating served as the score for each statement, and statement scores were aggregated by dimension. Separate scores were generated for each of the Critical Thinking dimensions by taking the average score across questions of that type.

Inductive Reasoning

Inductive Reasoning is a basic ability that has been measured numerous times in existing research. The O*NET definition of Inductive Reasoning is: "the ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events)" (O*NET Online, n.d.). Inductive Reasoning is typically measured using a serial completion task (e.g., Girelli, Semenza, & Delazer, 2004; Holzman, Pellegrino, & Glaser, 1983; Kotovsky & Simon, 1973; Lefevre & Bisanz, 1986; Simon & Kotovsky, 1963). Serial completion tasks provide the test taker with a series of figures, letters, or numbers that have a pattern, and the test taker must indicate what letter or number occurs next in the series. These measures are commonly used to assess fluid intelligence, and examples include Raven's Advanced Progressive Matrices (Bors & Stokes, 1998; Raven, 1936; 1941; 2000; Raven, Raven, & Court, 1998), Number Series Test (Ekstrom, French, Harman, & Derman, 1976), Primary Mental Ability (PMA) Reasoning Measure (Díaz-Morales & Escribano, 2013), Word Series Test (Schaie, 1985), and Adult Development and Enrichment Project Induction Test (Willis & Schaie, 1986).

While serial completion tasks are the most commonly used measure of Inductive Reasoning, there are other tests that have been developed, such as the reasoning bias task. Reasoning bias tasks are often based on a structure where participants are asked to choose one of two options (e.g., select which job a person is most likely to have) based on various pieces of information

that are given to them (e.g., information about a person). These tasks are assessed on the level of bias in their responses and how quickly participants draw conclusions or make decisions. Examples of these tasks include the Beads Task and Words Task (Jacobson, Freeman, & Salkovakis, 2012), the "Linda" Probability Judgment Task (Tversky & Kahneman, 1983), the Seven Letter Words Task (Moutier & Houde, 2003), and the Probabilistic Reasoning Task (Fraser, Morrison, & Wells, 2006).

Two additional types of Inductive Reasoning tests are causal inference tests, in which participants evaluate arguments as valid-believable, valid-unbelievable, invalid-believable, and invalid-unbelievable based on arguments created using causal sequences (Hayes, Stephens, Ngo, & Dunn, 2018), and statistical syllogism tests, in which participants must read a rule, view a geometrical figure, and determine if the figure is an example of the rule or not (Díaz-Morales & Escribano, 2013).

While there are quite a few options for assessing Inductive Reasoning, given the prevalence, success, and simplicity of the series completion tasks, we chose to use letter series completion tasks for the C^3 Inductive Reasoning measure. We based the measure on the item formalization and development procedures outlined by Simon and Kotovsky (1963). For each Inductive Reasoning item, test takers were presented with a 12-16 letter series containing four sections of a repeating pattern. The test taker was required to identify the pattern and type in the fifth section of letters to complete the series. Within the overarching assessment context, this task was framed as a code-breaking exercise in the application phase of the battery. The test taker received distress signals from some of the transportation equipment and was required to decode the signals in order to understand the distress signal.

Deductive Reasoning

Deductive Reasoning refers to the ability to apply general rules to specific problems to figure out answers or the cause of errors; specifically, O*NET defines Deductive Reasoning as "the ability to apply general rules to specific problems to produce answers that make sense" (O*NET Online, n.d.). Deductive Reasoning is typically measured using some type of syllogistic assessment items that ask test takers to arrive at conclusions based on two or more propositions or rules (e.g., The Syllogistic Test, Kuhn, 1977; or Syllogism Test, Gottesman & Chapman, 1960). As an example, the test taker would be provided with two statements: "All of Tom's ties are red. Some of the things Ada is holding are red." He or she must then determine whether follow on statements are valid or not valid (see Gottesman & Chapman, 1960). Example follow-on statements would be: (a) At least some of the things Ada is holding are Tom's ties. (b) At least some of the things Ada is holding are not Tom's ties. (c) None of the things Ada is holding are Tom's ties. Test takers are then asked to indicate whether each statement is valid or not. In the example provided about Ada and Tom, none of the statements are valid. Simpson and Nestor (2007) provide a detailed framework that can be used to create sets of premises and corresponding valid/invalid statements.

In another variation on syllogistic reasoning, Johnson-Laird's (1995) reasoning task provides test takers with 23 syllogisms and 23 implications, disjunctions, and conjunctions. Test takers determine if a third sentence is appropriate based on the first two sentences. The Syllogistic Reasoning Test (Morsanyi & Handley, 2012) includes information regarding believability in

addition to validity. Individuals review a series of syllogisms in which the validity (i.e. valid or invalid) and the believability (i.e. believable, unbelievable, or abstract) are manipulated, and participants evaluate the validity/believability of the conclusions for each syllogism. A third variation on the syllogism test limits the time the test taker has to respond; the Deductive Reasoning Task (Goel & Dolan, 2004) presents test-takers with three sentences that are presented one at a time in quick succession. They are then given 3.75-7.85 seconds (depending on the condition and series) to determine whether the given conclusion followed logically from the sentences.

To create the C^3 Deductive Reasoning test, we used the framework developed by Simpson and Nestor (2007) and designed a series of syllogism statements using topics related to the test taker's fictitious job. A set of premise statements were given to the test taker, who must determine whether each subsequent statement is valid or invalid based on the premises. No knowledge about the job was needed to respond to the questions, but the items use topics that are related to the fictitious job so that they fit conceptually within the assessment context. The Deductive Reasoning item sets were administered at various points during the learning phase of the battery.

Selective Attention

The O*NET definition of Selective Attention is "the ability to concentrate on a task over a period of time without being distracted" (O*NET Online, n.d.). This was an area that cyber SMEs said was especially important for success in cyber training or in cyber jobs. SMEs emphasized that there are always distractions that must be overcome on the job. According to the Army cyber SMEs, there are many "rabbit holes" that draw Soldiers away from other more important tasks. One key to high performance on these jobs is to not be distracted by these less-important things. This includes recognizing when you are heading down one of these "rabbit holes" and extracting yourself so you can get back on track. Despite its importance to cyber jobs, this construct is relatively undeveloped in the assessment field, with few existing measurement approaches.

Cognition researchers have identified two underlying processes that likely play a role in Selective Attention: attention and inhibition (Dagenbach & Carr, 1994; Hasher & Zacks, 1988; Maher & von Hippel, 2005; Tipper, 1985, 1992). The Stroop Task (Stroop, 1935) assesses both aspects of this joint process by presenting participants with color names printed in non-consistent ink colors. Participants are asked (among other things) to quickly name the color of the ink (attention), while attempting to disregard the color word they are reading (inhibition). This requires them to inhibit the natural tendency to read, in order to selectively attune to the color they are observing. This classic test has been demonstrated to be highly reliable and valid (Spreen & Strauss, 1998) and has been used in numerous studies to assess the construct of Selective Attention (Kane & Engle, 2003; Tams, Thatcher, Grover, & Pak, 2015). Other tests have been developed that assess participants' ability to inhibit distractions and actively focus on the task at hand, associating two types of stimuli such as directions and letters (Eriksen & Eriksen, 1974; Green & Bavelier, 2003; Zelazo et al., 2013) or digits and symbols (Ribas et al., 2010).

A construct similar to Selective Attention is Divided Attention, where individuals must focus on two (or more) tasks simultaneously (Miller, 1982). This construct has been measured using

similar (and often the same) measures as Selective Attention (e.g., the Stroop Task; Maher & von Hippel, 2005). Tasks in which participants are required to simultaneously pay attention to two sources of information are also common. For example, participants may be asked to read a list of words on paper and listen to a recording of words being read out loud, marking when they overlap (McDowd & Craik, 1988), or simultaneously memorize a list of words while solving subtraction problems (Drummond, Gillin, & Brown, 2001). Similar to Divided Attention is the construct Multitasking, which has been measured by asking participants to simultaneously perform various tasks across multiple time intervals. Multitasking has been shown to be predictive of job performance in military samples in which it has been measured (Barron & Rose, 2017; Phillips et al., 2011; Williams, Albert, & Blower, 2000).

Our investigation of Divided Attention and Multitasking revealed that while they have some similarities to Selective Attention, their focus is on completing multiple tasks at once, rather than concentrating on one task without being distracted by another, as is the focus of Selective Attention. Two other constructs were identified, Attention Management and Situational Awareness, that may have more direct definitional overlap with Selective Attention. Assessments developed to measure Attention Management are aimed at measuring "an individual's ability to scan multiple information sources, evaluate alternatives, establish priorities, and select and work on the task that has the highest priority at the moment" (such as the WOMBAT Situational Awareness, O'Hare, 1997; Roscoe, 1997). Similarly, Attention Control has been assessed by measuring an individual's ability to focus attention, shift attention between tasks, and flexibly control thought (Derryberry & Reed, 2002). Although the items in the Derryberry and Reed (2002) Attention Control Scale are self-report, the scale has still been shown to relate to tasks like the Stroop Task, suggesting construct overlap (Derryberry & Reed, 2002).

In order to focus our measurement on the behavior of concentrating on a task without being distracted, we measured Selective Attention by capturing how test takers responded to interruptions during the assessment. At various points throughout both the learning and application phases of the assessment, when the test taker was involved in a specific task (as opposed to navigating between tasks), an interruption pop-up message appears that presented brief information about the message's content, such as a subject line or headline. Test takers were given the option of either opening or closing the interruption, ostensibly to return to it later. All of the interruptions were rated for their importance by I/O psychologist SMEs that were familiar with the assessment to determine whether the appropriate course of action was to open the interruption or to save it for later. Raters determined the appropriateness of each response by considering the subject line of the interruption and using that to judge the interruption's relevance to the job and immediacy.

If the test taker decided to open a message, additional information was provided on the next screen of the popup and the test taker again has the option to learn more about it or save it for later. Each popup interruption had between 4 and 8 screens that can be viewed if the test taker chooses. SMEs rated the appropriate number of screens that should be viewed for each interruption, as well as the appropriate amount of time to spend on each interruption. These estimates were averaged and scaled to obtain z-scores and inherent rank orders for the amount of time that should be spent on each interruption. Test takers' time spent on each interruption was measured and similarly scaled to obtain z-scores relative to each test taker. These z-scores were

compared to SME scores and a deviation index was calculated to assess how far (whether spending more time or less time) each test taker was from the time allocations of the experts, on average.

Troubleshooting

Troubleshooting was defined as "systematically examining possible causes of operating errors in order to identify what needs to be done to return to normal operations" (O*NET Online, n.d.). Past studies have assessed Troubleshooting with exercises that ask subjects to solve Troubleshooting tasks (Kurland et al., 1992; Morris & Rouse, 1985; Swezey et al., 1988). Often these tasks are specific to a given system or type of work, such as electronics (e.g., Gitomer, 1988), information technology (e.g., Ross & Orr, 2009), or automotive or aircraft systems (e.g., Rouse, Rouse, & Pellegrino, 1980). Schaafstal et al. (2000) used engineering installation problems to observe how systematic participants were in their approach to Troubleshooting and how well they followed prescribed steps in diagnosing faults across four engineering problems. The participants were asked to think aloud throughout the Troubleshooting process, while two expert troubleshooters assessed them on various aspects of Troubleshooting.

Saltz and Moore (1953) presented participants a task in which they were tested on the types of checks and tests they ran in relation to a problem. Participants were assessed based on their tendency to avoid difficult checks, make difficult checks when simpler ones suffice, repeat the same checks, make irrelevant checks, or omit relevant checks. Similarly, Rouse et al. (1980) presented an exercise where participants were given a computerized network of nodes and outputs and were asked to test and determine the failed components. They were then assessed on the ratings of the tests they chose to perform. A related method of assessing Troubleshooting that was used by Saupe (1954) asked participants to make hypotheses about a presented problem. Participants were then assessed based on the number of correct or incorrect hypotheses, as well as the amount of time they spent pursuing incorrect hypotheses.

Several tests aimed at measuring similar constructs may also be useful in conceptualizing an assessment of Troubleshooting. For example, Jonassen (2012) developed a rubric for assessing decision making steps such as whether or not premises are stated and/or are relevant, the strength of the evidence for premises, the identification of counter arguments, and whether there is order/organization to arguments. These decision making steps have commonalities with Troubleshooting steps. Functional flow diagrams (Johnson & Satchwell, 1993) have also been shown to relate to performance in Troubleshooting. Ross and Orr (2009) developed a Social Problem-Solving Inventory which examines various aspects of the problem-solving process and includes a section relating to Troubleshooting.

Having knowledge about a task or system facilitates Troubleshooting in that area, so one challenge with assessing Troubleshooting is being able to test Troubleshooting ability or potential without requiring specific knowledge or experience on a given task. Teague and Allen (1997) provided participants with generic Troubleshooting tasks using computer-based diagrams that had 4, 8, or 12 circles connected by line segments. Some of the line segments were "working" and some were "not working." Participants had to test the components by clicking on them to determine where the problem was in the diagram. Three dimensions of Troubleshooting were captured: errors, time, and inefficiency.

Given that we did not want the C^3 Troubleshooting assessment to require knowledge of a particular task or system, we modeled the C^3 task after the generic task developed by Teague and Allen (1997). The C^3 Troubleshooting assessment consisted of a sequence of 9-12 fault diagnosis items. For each item, the test taker was shown a diagram of nodes, links, and outputs. The connections shown on the diagrams told the test taker how the structures are configured, and which components receive inputs from other components. A node that was not working in one location had downstream effects on other nodes causing them to stop working as well. The components of the network could be clicked to show whether each component was working or not working. Using the diagram of connections and information about outputs, the test taker needed to test the components (i.e., click them and receive working/not working information) in order to determine the source of a fault. Once the suspected fault was identified, the test taker double-clicked to replace the faulty component.

In order to incorporate the Troubleshooting assessment within the larger tube transport framework, the test taker was told they were solving problems that had occurred in the tube transportation lines. The test taker was presented with a text-based message identifying a problem in a particular location, as well as some details regarding the problem. This information was not required to complete the task but provided an overarching context and an avenue for the test taker to learn more about their job and role.

Summary

A series of nine interactive web-based assessments were developed to measure seven constructs conceptually related to common cyber capabilities: Active Learning 1 - Identifying Important Information (Active Learning III), Active Learning 2 - Learning Effectiveness (Active Learning LE), Complex Problem Solving 1 - Investigating Complex Problems (Complex Problem Solving ICP), Complex Problem Solving 2 - Evaluating Options (Complex Problem Solving EO), Critical Thinking, Inductive Reasoning, Deductive Reasoning, Selective Attention, and Troubleshooting. All nine assessments were presented within the context of the tube transport system job. Five of the assessments required that test takers apply knowledge or information that they acquired during the learning phase of the assessment (Active Learning III, Active Learning LE, Complex Problem Solving ICP, Complex Problem Solving EO, and Selective Attention); these were specifically dependent on the tube transport system context to capture the construct assessment. The other four assessments (Critical Thinking, Inductive Reasoning, Deductive Reasoning, and Troubleshooting) had the flavor or "wrapper" of the tube transport system, but did not require leveraging tube transport knowledge and information in order to complete the assessment; they could potentially be administered outside of the tube transport context as individual assessments.

One consideration in designing the assessment battery was that there is a high degree of conceptual overlap among the constructs, which can be seen in the O*NET definitions in Table 1. Some C^3 constructs are micro-level constructs (e.g., Inductive and Deductive Reasoning) that serve as building blocks for more macro-level C^3 constructs (e.g., Critical Thinking and Complex Problem Solving). For example, Critical Thinking is defined as "using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems" (O*NET Online, n.d.). This definition emphasizes that logic and reasoning (Deductive and Inductive Reasoning) is a building block of Critical Thinking, and

Critical Thinking is a building block of Problem Solving. Other definitions of Critical Thinking identify several critical thinking steps within Problem Solving: problem identification, problem definition, problem exploration, problem applicability, and problem integration (Garrison, 1992). Given this overlap, an effort was made to integrate the assessments such that the micro-level components were measured first, and more macro-level component measures were administered later in the assessment, allowing us to introduce information for the macro-level components while presenting and assessing the micro-level components.

Once the measures were developed and tested for functionality, research was conducted to collect initial descriptive, psychometric, and construct validity data on the measures. Given the experimental nature of many of the C^3 measures and the lack of suitable comparison measures for many of the constructs, this initial research focused on evaluating three areas: (1) user reactions and comments, (2) scale means and distributions, and (3) convergent and discriminant validity, specifically for Critical Thinking and Deductive Reasoning.

CHAPTER 2: METHOD

Data were collected from a sample of participants who were contacted through the Amazon Mechanical Turk employment system. This platform provides psychological researchers with access to a subject pool and can provide valid and generalizable results (Buhrmester, Kwang, & Gosling, 2011; Cheung, Burns, Sinclair, & Sliter, 2017). Participants completed a series of measures online: C^3 Assessment battery, Feedback Questionnaire, Demographics Questionnaire, and additional assessments of Critical Thinking, Deductive Reasoning, General Cognitive Ability, and Problem Solving.

Participants

Participants in the study included 73 Amazon Mechanical Turk workers who agreed to complete the C^3 battery and other measures. Participants received up to $32 in base compensation. The study was designed to take no more than four hours, so participants received at least $8.00 of compensation per hour. In addition, participants were offered the potential of an added bonus for performing well on the tests. This helped to more closely simulate a selection situation. Any participant that performed in the top 20 percent across all tests was entered into a drawing for one of three bonuses of $40.

Prior to summarizing sample descriptive statistics and conducting analyses, the data were screened for careless or random responding. Data for three individuals were removed because they completed the tasks in such a short amount of time that they could not have been diligently responding (15 minutes or less per test phase). In addition, two respondents were removed because they were not between 18-50 years of age. One person was removed for taking the test 2.5 times and improving his/her scores by applying the feedback that was provided during the assessment. This left a final sample of 67 participants, although analyses that involved PDRI measures had 19 fewer participants for sample size of 48 because several participants did not complete those measures. In addition, some analyses have fewer participants due to missing data on specific items or scales.

Study participants had a mean age of 37.18, (*SD* = 7.39), with a range from 18 to 50. The average level of college education was 3.68 years (*SD* = 2.16), with 76.92% earning a college degree and 10.61% earning an advanced degree. Around one-fourth of the sample (23.88%) indicated that they currently hold or have held a job in information technology, while 4.48% indicated that they currently hold or have held a job in computer security. Nearly half (43.28%) indicated that they have never helped someone fix a computer problem.

Measures

The C^3 Battery

A general description of the C^3 assessments that were developed was provided in the introduction of this report. This section describes additional details in areas such as the number of items, specific scenarios used for each of the C^3 assessments, and details about scoring approaches that relied on subject matter expert (SME) ratings. Internal consistency reliability, where available, is described in the Results.

Active Learning. The Identifying Important Information (III) dimension of Active Learning was measured with 12 items evenly divided among the learning and application phases. Each item had between 10 and 15 statements of fact which could be selected as important or not important.

Each statement was assigned a point value based on SME ratings of the importance of each fact to the test taker's role. SMEs were four I/O psychologists who were involved with developing the assessments. During this process, they rated each statement then discussed disagreements to reach a consensus and final point value for each fact. Consistency among SME ratings was ICC(1) = .92. Each statement ranged from -2 to +2 points. Test takers could select up to four statements per item, and item scores were the sum total of statement values, so each item could potentially range between -8 and +8. In most cases, the range of potential values for an item was different from -8 to +8 due to an uneven balance of high-value and low-value facts (e.g., some items did not include four facts with a -2 value, so they did not have a minimum possible value of -8). Since these ranges were different for each Active Learning III item, the items were standardized before computing averages or other statistics. As a result, individual Active Learning III items had a mean of zero and standard deviation of one, but total scores for III calculated by averaging across items may not have the same characteristics.

For the Learning Effectiveness (LE) dimension of Active Learning, knowledge tests were administered at the end of both the learning and application phases. Both Active Learning LE tests consisted of 15 questions. All questions were multiple-choice with one correct answer and four distractors.

Complex Problem Solving. Two approaches were used to measure Complex Problem Solving. For Investigating Complex Problems (ICP), test takers were presented with scenarios and given problems to investigate. They could explore a set of six databases by using various strategies, such as searching for key words or sorting by variables of interest. Each database had a different purpose and yielded different information about the functioning of the tube network. Test takers were allowed to spend as much time as they wanted searching the databases but were encouraged to work quickly and use their time efficiently.

After test takers were done looking for information in the databases, they were presented with a list of accurate and inaccurate statements and asked to select the accurate statements from the list. Scores were determined using a sensitivity index (d-prime) to account for response bias.

The second measure of Complex Problem Solving, Evaluating Options (EO), was captured after test takers completed the ICP measure. After completing the ICP measure, test takers are given feedback showing which ICP statements were accurate. They are then given a list of potential courses of action to take next and asked to evaluate the effectiveness of each action as a next step. Four I/O psychologists familiar with the scenarios rated these options for effectiveness. If these SMEs differed in their ratings after consensus discussions, ratings were averaged to arrive at the "true" value of each course of action. Test takers' scores were calculated as the distance between test takers' ratings and the average ratings of the SMEs. The scores were rescaled to a 1-5 scale, where 1 equals low agreement with SMEs, and 5 equals high agreement. Consistency of the SME ratings was ICC(1) = .76.

Another experimental approach to assessing Complex Problem Solving, and ICP in particular, is to look at how much database search activity the test takers engage in. Measures of search activity were derived from how test takers used the databases. Search activity was conceptualized as the extent to which 1) the number of times test takers searched in the correct location for information, 2) the number of searches conducted from the correct locations, and 3) the number of correct search terms used when searching the correct locations. The first index is a count of searches. The second index is a value from 0 to 9 representing the number of scenarios (out of 9) in which test takers used the appropriate data source to find information. Finally, the third and most specific index is a measure of the specific search terms that are used to find information. Due to the experimental nature of this index, focus was placed on the search terms that were used in only two Complex Problem Solving scenarios where test takers were directed to the InfoSearch database to find files related to the problem scenario.

Critical Thinking. Critical Thinking (CT) was measured in the learning phase using three scenarios that were related to training material but was distinct from training content. The scenarios were formatted as newspaper articles written by sources outside the organization so that test takers understood the information in the articles was not necessarily factual. Test takers read the articles, then responded to a series of prompts designed to elicit analysis, interpretation, explanation, or evaluation of the information from the articles. Following the scenario, the test taker responded to a series of questions about the article using a 5-point Likert rating scale. For example, an interpretation prompt was, "To what extent was each of these arguments made by the article?" Following each prompt was a set of 5 to 11 statements expressing different possible arguments that could have been made, and the test taker rated the extent to which the author made each argument in the article. Ratings from test takers were compared to SME ratings on the same statements. SMEs were four I/O psychologists familiar with the scenarios. If these SMEs differed in their ratings after consensus discussions, ratings were averaged to arrive at the "true" value of each course of action. Test takers' scores were rescaled so that high scores indicate agreement with SMEs and low scores indicate disagreement with SMEs. SME raters had a consistency of ICC(1) = .82 for the Critical Thinking ratings.

Inductive Reasoning. Inductive Reasoning (IR) was measured at the end of the application phase using a letter series completion task. Test takers were shown a patterned string of letters that was either 12 or 16 letters long and asked to fill in the next six letters in the pattern. Responses were open-ended such that test takers typed the letters of their responses rather than selecting response options. Eight Inductive Reasoning items were used. Participants received full credit for each correct answer and no credit for each incorrect answer, such that their Inductive Reasoning score could range from zero to eight points or zero to 100 percent.

Deductive Reasoning. Seven items administered in the learning phase comprised the Deductive Reasoning (DR) scale. For each item, test takers were presented with a logical premise using information from the learning phase and asked to evaluate if each statement in a subsequent set followed logically from the premise. Statement sets ranged from 8 to 11 statements, each with a distinct logical structure. For each statement, the test taker indicated whether it was logically valid, logically invalid, or "I don't know." Two scoring approaches were tested: dichotomous scoring (correct responses get 1, incorrect or "I don't know" get 0) and formula scoring (correct responses get 1, "I don't know" responses get 0, and incorrect responses get -1). Some of the statements were bias-prone, meaning that people were more likely to answer

them incorrectly due to a reasoning bias. These statements were analyzed separately and aggregated into a separate Deductive Reasoning index.

Selective Attention. Selective Attention (SA) was measured throughout the battery by examining test takers' responses to interruptions. Interruptions appeared in the form of pop-up messages which showed a one-sentence description of the interruption's content. Test takers first decided whether to open the message or save it for later. These decisions were scored as either correct or incorrect based on SME assessments of the importance of the interruption content. SMEs consisted of four I/O psychologists with deep familiarity with the assessment battery. SME interruption importance decisions were reached by first discussing the interruptions and then agreeing on the most appropriate responses. Scores for this Selective Attention decision were calculated using SME ratings and the d-prime sensitivity index. Interruptions that were rated as important by SMEs were considered to be "hits" if the test takers opened and viewed the message or "misses" if the test taker did not open and view the message. Interruptions that were rated as unimportant by SMEs were considered to be "false alarms" if the test takers opened and viewed the message, or "correct rejections" if the test taker did not open and view the message.

In addition to the initial Selective Attention choice, the amount of time that test takers spent on interruptions was used to calculate a Selective Attention duration index. SMEs rated the appropriate amount of time to spend on each interruption and then discussed their ratings to recalibrate their assessments. SME consistency after recalibration was ICC(1) = .90. Estimates across all interruptions were converted to z-scores for each rater and then averaged to come up with an SME Selective Attention duration value for each interruption. Test takers' z-scored duration on each interruption was subtracted from the SME z-scored average, and the average of the absolute value of difference scores was used as the Selective Attention duration metric.

Troubleshooting. For the Troubleshooting (TS) measure, the test taker was required to diagnose faults in nine different Common Systems Networks. The networks were comprised of nodes and outputs that were linked with connections in a specific configuration. The test taker was presented with a diagram of each system as a set of nodes, links, and outputs. The links between the nodes and outputs demonstrated to the test taker how the structures were configured, and which components received inputs from other components. Test takers could click a "Show Status" button to see which readers were indicating output. If a reader did not indicate output, one or more of the components of the network was broken due to a fault in the network that needed to be found and fixed. Using the diagram of connections and information about outputs, the test taker was required to test the components (click them and receive working/not working information) to determine the source of a fault. Once the suspected fault was identified, the test taker could double click to repair the faulty component. Test takers were asked to work systematically to identify the faults with as few tests (single clicks) or fixes (double clicks) as possible. There were three types of Troubleshooting items which differed on the arrangement of the network and the consistency of the faults. Linear items had networks that were arranged in linear sequences. Networked items were arranged in two or more parallel streams, so a fault in one stream could be overcome by information transmission through a neighboring stream. Intermittent items could be linear or networked and had faults that appeared intermittently, every second or third click. Troubleshooting scores were calculated for each item type and overall using two indices: the percent of faults found within the network out of the number present, and

the number of faults identified and fixed in the network, divided by the number of double clicks attempted (plus one to keep scores from being undefined when no double clicks were entered).

C^3 Battery Feedback Questionnaire

The C^3 Battery Feedback Questionnaire was composed of 11 items that were adapted or created specifically for this study. It included four Likert ratings and seven open-ended questions about the participants' experiences with the C^3 Test.

Critical Thinking Dispositions Scale

The Critical Thinking Dispositions Scale was added to assess convergent validity of the C^3 Critical Thinking measure. This scale consisted of 11 items measuring a predisposition to think critically. Responses to these items were made using a 5-point Likert scale ranging from "strongly disagree" to "strongly agree." This scale consisted of two subdimensions: Critical Openness and Reflective Skepticism. Critical Openness had seven items and a reliability of Cronbach's alpha = .78 with the current sample. Reflective Skepticism consisted of four items and a reliability of Cronbach's alpha = .78 with the current sample.

PDRI Deductive Reasoning

PDRI's commercial selection test of Deductive Reasoning consists of multiple-choice items measuring test takers' ability to draw logical conclusions based on evidence. The test is computer adaptive and includes between 16 and 20 items. Test takers had up to three minutes for each question before the next question was presented. This measure was used to gather convergent validity evidence for the C^3 Deductive Reasoning test.

PDRI Cognitive Abilities

PDRI's Cognitive Ability test was used to measure general mental ability. This consists of 45 items incorporating a multiple-choice response format. These items measure a variety of cognitive processes including spatial, analogical, and matrix reasoning. Test takers had three minutes to complete each item.

PDRI Problem Solving

A general problem-solving measure was added to gather convergent validity evidence for the C^3 Complex Problem Solving test. Problem Solving was measured with PDRI's commercially available Problem Solving – Qualitative assessment. This assessment measured test takers' ability to reason through problem situations and suggest potential solutions. The test included 10 items and had a three-minute time limit per item.

Demographics Questionnaire

The Demographics questionnaire consisted of 12 items that were adapted or designed for this study. Demographics information included age, education attainment, and experience questions pertaining to the test takers' past experiences with computer technology and cybersecurity.

Procedure

The study was posted on Amazon's Mechanical Turk website with descriptions of the tasks and objectives, as well as informed consent materials describing participant rights, the benefits and risks of participating, and the right to withdraw without penalty. If workers agreed to participate, they began the task by reading the study instructions and clicking a link to the test website. The demographics questionnaire was administered first, followed by the learning phase of the C^3 assessment battery. Then participants completed a brief survey about their experience with and reactions to the training. Next, participants began the application phase of the assessment followed by a brief post-assessment reactions questionnaire and the comparison measures for construct validity evidence.

CHAPTER 3: RESULTS

Overview

On average, participants in the study spent a total of 170 minutes on the C^3 battery. The average amount of time spent on the learning phase was 91 minutes. The average amount of time spent on the application phase was 79 minutes. These averages include breaks taken during either the learning or application phases, but exclude breaks taken in between phases.

Test takers completed 5-point Likert-type ratings (1 = strongly disagree, 5 = strongly agree) about the assessment battery generally, and about the learning phase training content in particular. Means for each question are ranked from highest to lowest in Table 2 and Table 3. A list of top responses to the open-ended questions is presented in Appendix A.

Table 2

Average Agreement with Statements about the Overall Assessment

	Mean	SD
I found the assessment interface unnecessarily complex.	4.04	1.15
I found the overall system very hard to use.	3.90	1.16
The assessment was interesting.	3.85	1.27
The job tasks and responsibilities of the CRO became apparent as I went through the training.	3.46	0.89
The system works the way I expected it to work.	2.70	1.14
The assessment is easy to use.	2.46	1.23
It was easy to use the control panel to find information about the scenarios.	2.39	1.10
I would imagine that most people would learn to use this system very quickly.	2.31	1.09
The assessment is easy to understand.	2.27	1.10
I learned what I needed to do quickly.	2.21	1.15

Note. 1 = Strongly Disagree to 5 = Strongly Agree

Table 3

Average Agreement with Statements about Training

	Mean	SD
The lesson was well organized.	4.37	0.76
The training has increased my awareness of the CRO position.	4.31	0.84
The goal of the training program was clear.	4.12	0.98
The lesson was interesting.	3.91	1.18

Note. 1 = Strongly Disagree to 5 = Strongly Agree

The C^3 battery of assessments included seven overarching constructs and numerous subdimensions. Distributions of scores from these measures were generally normal, though there was some skewness and kurtosis of the distributions. With larger samples of data, most distributions should approximate normal distributions. Some exceptions will be discussed in the following sections. Means, standard deviations, and intercorrelations for the C^3 constructs are presented in Table 4.

Active Learning

Two dimensions of Active Learning were measured: Identifying Important Information (III) and Learning Effectiveness (LE). Both dimensions were measured separately for the test battery's learning and application phases.

Identifying Important Information

III was measured using a task in which test takers selected relevant information from a menu with 10-15 boxes containing information. Each piece of information was rated by SMEs ahead of time and given an importance value based on the average SME rating. Test takers scored highly by selecting the information rated as highly important and avoiding the information rated as unimportant. For each set of statements, test takers were told to select between one and four pieces of information. Cronbach's alpha for the full set of items was 0.60.

III was measured in the learning and application phases with six items each, and items were standardized prior to additional analyses, so the means for all items were zero and the standard deviations were one. Cronbach's alpha for the learning phase set of items was 0.67. Item minimums and maximums, item-total correlations, and Cronbach's alpha if an item was dropped are presented in Table 5 for learning phase Active Learning III items.

Table 4

C^3 Descriptive Statistics and Intercorrelations with Confidence Intervals

	M	SD	AL III	AL LE	CPS ICP	CPS EO	CT	DR FS	IR	SA
AL III	0.00	0.43								
AL LE	48.27	15.46	.27* [.03, .48]							
CPS ICP	0.29	0.28	.36** [.14, .56]	.38** [.15, .57]						
CPS EO	3.63	0.24	.16 [-.09, .38]	.27* [.03, .48]	.22 [-.02, .44]					
CT	3.84	0.25	.39** [.16, .57]	.32** [.09, .52]	.22 [-.02, .44]	.24 [-.00, .45]				
DR FS	0.66	0.15	.39** [.17, .58]	.21 [-.03, .43]	.39** [.16, .57]	.02 [-.22, .26]	.30* [.06, .50]			
IR	0.35	0.29	.26* [.03, .47]	.36** [.13, .55]	.36** [.14, .56]	.29* [.05, .49]	.28* [.04, .48]	.14 [-.11, .37]		
SA	0.54	0.68	.03 [-.21, .27]	.10 [-.14, .33]	.17 [-.07, .39]	-.05 [-.29, .19]	-.03 [-.27, .21]	-.02 [-.26, .22]	.08 [-.16, .32]	
TS EF	0.35	0.12	.16 [-.08, .39]	.20 [-.04, .42]	.15 [-.09, .38]	.19 [-.05, .41]	.06 [-.18, .30]	.27* [.03, .48]	.40** [.18, .58]	.12 [-.12, .35]

Note. M and SD are used to represent mean and standard deviation, respectively (n = 67). Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).
AL III = Active Learning: Identifying Important Information Overall Score, AL LE = Active Learning: Learning Effectiveness Overall Score, CPS ICP = Complex Problem Solving: Investigating Complex Problems, CPS EO = Complex Problem Solving: Evaluating Options, CT = Critical Thinking Overall Score, DR FS = Deductive Reasoning: Basic Formula Score, IR = Inductive Reasoning, SA = Selective Attention, TS EF = Troubleshooting Efficiency.
*p < .05. **p < .01.

Table 5

Learning Phase Active Learning: Identifying Important Information Item Statistics

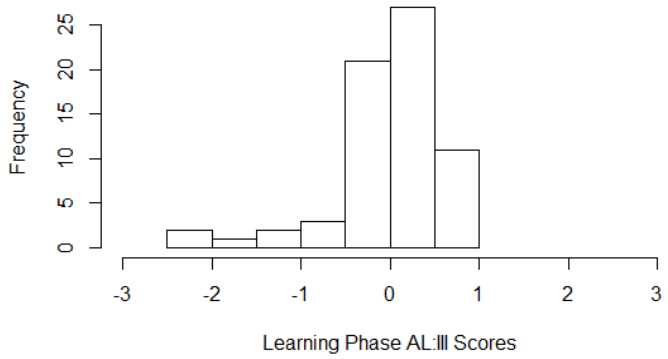
	n	Min	Max	Corrected Item-Total Corr	Alpha if Dropped
AL III Item A	67	-3.69	0.68	0.59	0.55
AL III Item C	67	-3.02	1.01	0.28	0.66
AL III Item D	67	-2.84	1.25	0.33	0.65
AL III Item E	67	-3.99	0.90	0.43	0.61
AL III Item F	67	-2.52	1.32	0.32	0.65
AL III Item G	67	-2.48	0.97	0.43	0.61

Note. Data were not collected for items B and H in an effort to shorten the assessment battery.
AL III = Active Learning: Identifying Important Information Overall Score.

The distribution of Active Learning III scores averaged across items had a skewness of -1.54 and a kurtosis of 5.85. A histogram of the distribution of learning phase III scores is shown in Figure 1.

Figure 1

Histogram of Active Learning: Identifying Important Information Scores (Learning Phase)



Note. AL III = Active Learning: Identifying Important Information Overall Score.

Item statistics for the six application phase Active Learning III items, including minimums, maximums, item total correlations, and Cronbach's alpha if an item was dropped are presented in Table 6. Overall Cronbach's alpha for the application phase was 0.40.

Table 6

Application Phase Active Learning: Identifying Important Information Item Statistics

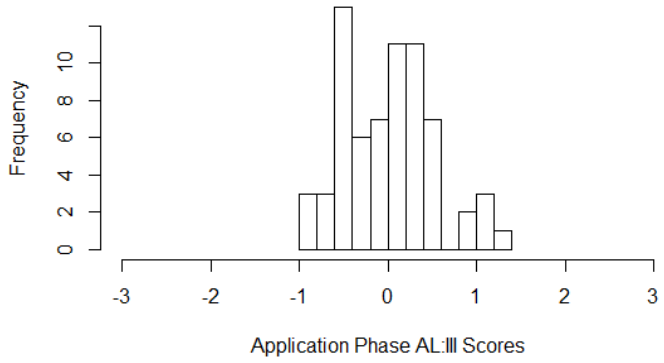
	n	Min	Max	Corrected Item-Total Corr	Alpha if Dropped
AL III Item I	67	-2.15	1.16	0.19	0.37
AL III Item J	67	-1.90	2.64	0.20	0.37
AL III Item K	67	-2.13	1.38	0.20	0.36
AL III Item N	67	-2.58	1.70	0.15	0.40
AL III Item O	67	-2.96	1.38	0.28	0.31
AL III Item P	67	-2.29	1.89	0.15	0.39

Note. Data were not collected for items L and M in an effort to shorten the assessment battery.
AL III = Active Learning: Identifying Important Information.

Overall average scores on III in the application phase had a skewness of 0.41 and a kurtosis of 2.77. A histogram of the distribution of application phase III scores is shown in Figure 2.

Figure 2

Histogram of Active Learning: Identifying Important Information Scores (Application Phase)



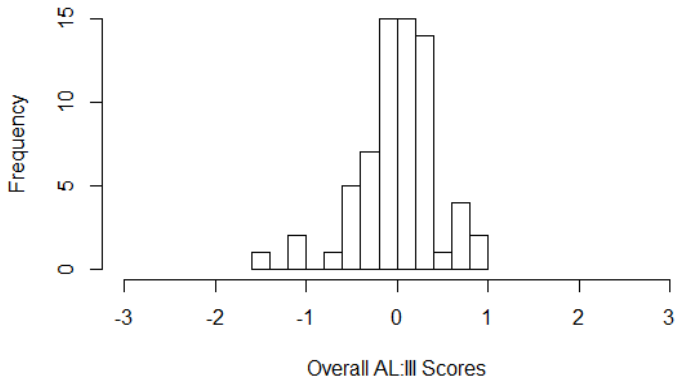
Note. AL III = Active Learning: Identifying Important Information.

The correlation between learning phase and application phase III scores was .18 (p = .14), suggesting a distinction between how test takers recognize what is important in training and how they recognize what is important in a problem-solving situation. Despite these differences, scores on Active Learning III were averaged across both the learning and application phases to reflect overall abilities for identifying important information.

Overall scores on Active Learning III had a mean of 0.00 and standard deviation of 0.43. The distribution of these scores had a skewness of -0.83 and a kurtosis of 4.73. A histogram of the distribution of overall Active Learning III scores is shown in Figure 3.

Figure 3

Histogram of Overall Active Learning: Identifying Important Information Scores



Note. AL III = Active Learning: Identifying Important Information.

Learning Effectiveness

Two tests were used to measure Active Learning LE: one at the end of the learning phase (LE1), and one at the end of the application phase (LE2). A five-option, multiple choice question format with one correct answer was used for the LE tests. Each test consisted of 15 items which asked the test taker about relevant information from its respective phase.

LE1. Item difficulties for the 15 LE1 items ranged from 0.21 to 0.88, with an average of difficulty of 0.5. Item discriminations ranged from 0.14 to 0.73 with a mean of 0.43. Response option discriminations were calculated for correct and incorrect response options. This information can be found in Appendix B, Table B-1.

Corrected item-total correlations ranged from 0.10 to 0.42, with a mean of 0.26. Most items did not detract from the internal consistency of the set. Cronbach's alpha, with individual items removed, ranged from 0.59 to 0.64. The overall internal consistency of the set of items was 0.63.

Table 7 shows the number of times an item was answered, item means (or difficulties in this case) and standard deviations, corrected item-total correlations, item discrimination coefficients, and Cronbach's alpha for the set if an item was dropped for items in the LE1 test.

Table 7

Learning Effectiveness 1 Item Statistics

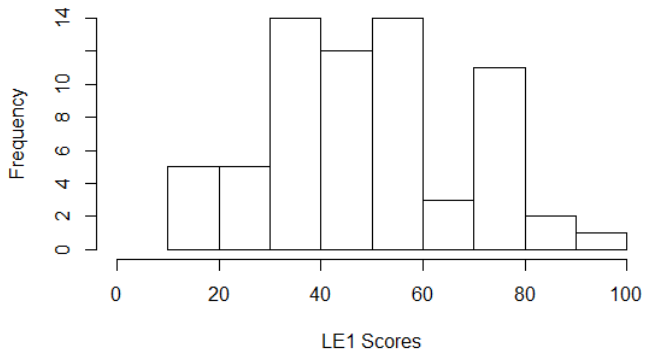
	n	Mean (Difficulty)	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
ALLE1.1	67	0.88	0.33	0.21	0.27	0.62
ALLE1.4	67	0.84	0.37	0.27	0.32	0.62
ALLE1.12	67	0.22	0.42	0.10	0.14	0.64
ALLE1.14	67	0.39	0.49	0.20	0.36	0.63
ALLE1.15	67	0.45	0.50	0.24	0.45	0.62
ALLE1.16	67	0.51	0.50	0.42	0.73	0.59
ALLE1.21	67	0.46	0.50	0.23	0.36	0.62
ALLE1.22	67	0.42	0.50	0.19	0.36	0.63
ALLE1.23	67	0.21	0.41	0.24	0.27	0.62
ALLE1.24	67	0.46	0.50	0.27	0.55	0.62
ALLE1.25	67	0.40	0.49	0.39	0.68	0.60
ALLE1.26	67	0.46	0.50	0.36	0.50	0.60
ALLE1.27	67	0.61	0.49	0.20	0.41	0.63
ALLE1.29	67	0.64	0.48	0.21	0.45	0.63
ALLE1.32	67	0.61	0.49	0.29	0.55	0.62

Note. ALLE = Active Learning Learning Effectiveness.

LE1 scores were calculated as the percent correct out of 15 items. The overall mean and standard deviation of LE1 scores was 50.43 and 18.98, respectively. The distribution of LE1 scores had a skewness of 0.23 and a kurtosis of 2.33. Figure 4 shows a histogram of the LE1 distribution.

Figure 4

Histogram of Active Learning: Learning Effectiveness 1 Scores (Learning Phase)



LE2. Item difficulties for the 15 LE2 items ranged from 0.07 to 0.85, with an average of difficulty of 0.46. Item discriminations ranged from 0.09 to 0.55 with a mean of 0.41. Response option discriminations were calculated for correct and incorrect response options. This information can be found in Appendix B, Table B-2.

Corrected item-total correlations ranged from -0.05 to 0.45, with a mean of 0.22. Most items did not detract from the internal consistency of the set. Cronbach's alpha, with individual items removed, ranged from 0.51 to 0.59. The overall internal consistency of the set of items was 0.57.

Table 8 shows the number of times an item was answered, item means (or difficulties in this case) and standard deviations, corrected item-total correlations, item discrimination coefficients, and Cronbach's alpha for the set if an item was dropped for items in the LE2 test.

Table 8

Active Learning: Learning Effectiveness 2 Item Statistics

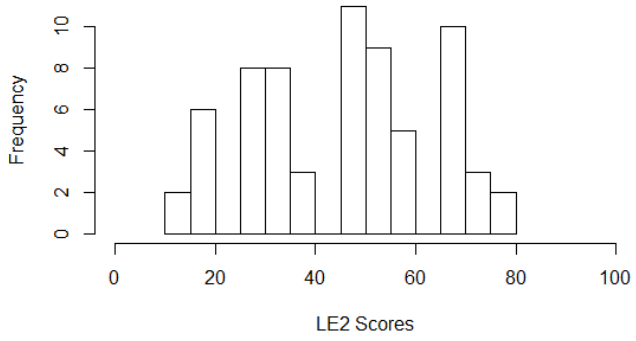
	n	Mean (Difficulty)	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
ALLE2.1	67	0.42	0.50	0.31	0.50	0.53
ALLE2.4	67	0.43	0.50	0.27	0.50	0.54
ALLE2.6	67	0.57	0.50	0.20	0.45	0.56
ALLE2.8	67	0.49	0.50	0.32	0.55	0.53
ALLE2.9	67	0.49	0.50	0.29	0.45	0.54
ALLE2.10	67	0.73	0.45	0.45	0.55	0.51
ALLE2.12	67	0.67	0.47	0.10	0.27	0.57
ALLE2.13	67	0.07	0.26	-0.05	0.09	0.59
ALLE2.14	67	0.30	0.46	0.19	0.45	0.55
ALLE2.15	67	0.49	0.50	0.23	0.45	0.55
ALLE2.16	67	0.43	0.50	0.11	0.32	0.57
ALLE2.17	67	0.27	0.45	0.03	0.27	0.58
ALLE2.18	67	0.39	0.49	0.26	0.45	0.54
ALLE2.19	67	0.85	0.36	0.21	0.27	0.55
ALLE2.21	67	0.30	0.46	0.30	0.50	0.53

Note. ALLE = Active Learning, Learning Effectiveness.

The overall mean and standard deviations of LE2 scores were 46.10 and 17.80, respectively. The distribution of LE2 scores had a skewness of -0.03 and a kurtosis of 1.96. Figure 5 shows a histogram of the LE2 distribution.

Figure 5

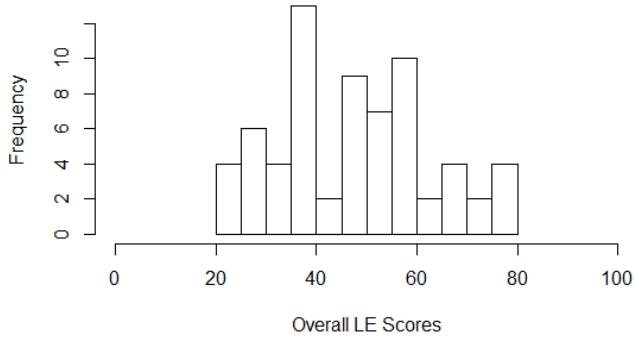
Histogram of Active Learning: Learning Effectiveness 2 Scores (application Phase)



Using a unit-weighting method for both Learning Effectiveness scores (LE1 and LE2), we averaged scores to arrive at an overall LE score. The mean and standard deviations of the overall LE scores were 48.27 and 15.46, respectively. The distribution of overall LE scores had a skewness of 0.20 and a kurtosis of 2.34. Figure 6 shows a histogram of overall LE scores.

Figure 6

Histogram of Overall Active Learning: Learning Effectiveness Scores



The correlation between the Active Learning measures for the two III phases was not significant (r = .18, p = .14), but the correlation between the two LE phases was significant (r = .41, p < .001). The correlation between average III scores and average LE scores was significant (r = .27, p = .03). Active Learning was expected to correlate with Complex Problem Solving (CPS) because of its role as an antecedent to problem-solving performance. Both facets of Active Learning were predicted to be important for Complex Problem Solving, though learning in the

application phase was predicted to be more strongly correlated with Complex Problem Solving scores. Both facets of Active Learning were also expected to correlate with Selective Attention, due to the antecedent role of attention in the learning process. Active Learning was also predicted to correlate with cognitive ability, since there is a strong role for intelligence in learning. Table 9 shows the results of correlation analyses examining these expectations.

Table 9

Active Learning Correlations

	AL LE LP	AL III LP	AL LE AP	AL III AP	AL LE	AL III
CPS ICP	0.34**	0.13	0.29*	0.46**	0.38**	0.36**
CPS EO	0.21	-0.05	0.24*	0.33**	0.27*	0.16
SA	0.00	0.06	0.17	-0.03	0.10	0.03
PDRI CA	0.27	0.08	0.37*	0.36*	0.37*	0.30*

Note. When PDRI measures were involved $n = 48$, otherwise $n = 67$. AL LE LP = Active Learning: Learning Effectiveness Learning Phase, AL III LP = Active Learning: Identifying Important Information Learning Phase, AL LE AP = Active Learning: Learning Effectiveness Application Phase, AL III AP = Active Learning: Identifying Important Information Application Phase, AL III = Active Learning: Identifying Important Information Total Score, AL LE = Active Learning: Learning Effectiveness Total Score, CPS ICP = Complex Problem Solving: Investigating Complex Problems, CPS EO = Complex Problem Solving: Evaluating Options, SA = Selective Attention, PDRI CA = PDRI Cognitive Ability.
* $p < .05$. ** $p < .01$.

Both Active Learning facets were significantly correlated with the ICP facet of Complex Problem Solving. Correlations were significant for the Active Learning III facet ($r = .36$, $p < .001$) and for the LE facet ($r = .38$, $p < .001$). Only the Active Learning LE facet was significantly related to the EO facet of Complex Problem Solving ($r = .27$, $p = .03$). Application-phase Active Learning measures generally resulted in higher-magnitude correlations with both facets of Complex Problem Solving than learning-phase measures. One exception to this was the correlation between Complex Problem Solving ICP and Active Learning LE scores, which decreased slightly from the learning phase to the application phase.

Total scores for both facets of Active Learning were correlated with cognitive abilities with correlations of $r = .37$ ($p = .01$) for Active Learning LE and $r = .30$ ($p = .05$) for Active Learning III. When the two assessment phases were analyzed separately, only the application-phase scores were significantly correlated with cognitive abilities ($r = .36$, $p = .01$ for Active Learning III; $r = .37$, $p = .01$ for Active Learning LE). Selective Attention was not significantly correlated with any Active Learning facet or phase score.

Complex Problem Solving

Complex Problem Solving was measured in the application phase using two dimensions: Investigating Complex Problems (ICP) and Evaluating Options (EO). Three problem-solving scenarios were presented to test takers, each with four parts. After an initial problem description, test takers were given an opportunity to find relevant information in the databases and then they completed ICP and EO items. This cycle repeated 12 times - once for each of the four parts of the three scenarios.

Investigating Complex Problems (ICP)

There were two components to the ICP measurement. First, test takers' search activities were examined to see how they used the databases to find relevant information. Second, test takers received a score for their accuracy in reporting facts about the scenario.

Search activity was conceptualized as (a) the number of times test takers searched in the correct location for information (with searches either conducted in the correct or incorrect control panel locations for each scenario), (b) the number of searches conducted from the correct locations (a count of the number of correct control panel location searches for each scenario), and (c) the number of correct or useful search terms used when searching the correct locations. The first index is a count of searches, which varied from zero to 74 searches in the present sample, with a mean of 14.43. The second index is a value from zero to nine, representing the number of scenarios (out of nine) in which test takers used the appropriate data source to find information. The maximum number of scenarios for this score is nine because we were unable to capture information about the use of the correct data source for the three scenarios that inquire about information in the Alert Monitor portion of the database. The mean of this index was 4.25. Finally, the third and most specific index looks at the specific search terms that were used to find information. Due to the experimental nature of this index, focus was placed on the search terms that were used in only two Complex Problem Solving scenarios where test takers were directed to the InfoSearch database to find relevant files related to the problem scenario.

For each of the 12 scenarios, after test takers searched the databases, they were provided with a set of eight statements about the scenario and asked to select all statements that were true and ignore statements that were false. Sensitivity indices (d-prime) and biases (c) were calculated for these ICP statements using the correctly identified true statements as 'hits' and the false statements identified as true as 'false alarms.' One d-prime score was obtained for each of the 12 scenario and these 12 d-prime scores served as the items for overall ICP. Average d-prime and bias statistics for each item are given in Table 10, along with hit rates and false alarm rates for those items. The mean, standard deviation, and internal consistency reliabilities are given for each problem-solving scenario in Table 11. Correlations between scenarios were .24 (scenarios 1 and 2), .00 (scenarios 1 and 3), and .14 (scenarios 2 and 3), for an average correlation of 0.12.

Table 10

Complex Problem Solving: Investigating Complex Problems Item Statistics

	d-prime	Bias (c)	Hit rate	False alarm rate
ICP item 1	0.07	0.67	0.27	0.24
ICP item 2	0.37	0.39	0.47	0.29
ICP item 3	0.02	0.38	0.40	0.37
ICP item 4	0.01	0.35	0.42	0.39
ICP item 5	0.45	0.46	0.44	0.25
ICP item 6	0.75	0.37	0.59	0.23
ICP item 7	0.51	0.37	0.51	0.28
ICP item 8	0.40	0.12	0.64	0.40
ICP item 9	0.15	0.35	0.44	0.36
ICP item 10	0.35	0.37	0.47	0.30
ICP item 11	0.11	0.40	0.40	0.34
ICP item 12	0.31	0.37	0.48	0.32

Table 11

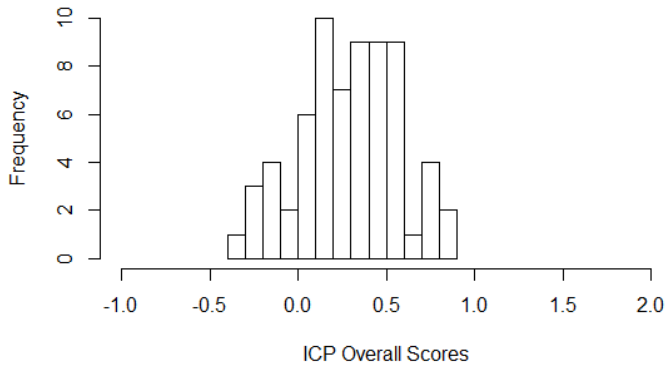
Complex Problem Solving: Investigating Complex Problems Scenario Statistics

	Mean d-prime	SD	Mean c	Cronbach's Alpha
Scenario1	0.12	0.43	0.45	0.20
Scenario2	0.53	0.49	0.33	0.45
Scenario3	0.23	0.36	0.37	-0.10

Cronbach's alpha across all 12 items was 0.37; however, alpha varied by scenario. In the third scenario, items were not reflecting a consistent underlying factor and resulted in a negative alpha value. The average ICP score across scenarios was 0.29 with a standard deviation of 0.28. Overall ICP scores had a skewness of -0.15 and a kurtosis of 2.61. Figure 7 shows a histogram of the ICP score distribution.

Figure 7

Histogram of Complex Problem Solving: Investigating Complex Problems Scores



Evaluating Options (EO)

For the EO measure, test takers were given a problem-solving scenario and an item stem with information about a specific update. Below the stem was a list of several possible courses of action. Test takers were asked to rate each course of action based on its effectiveness. Similar to other SJTs, scores were based on proximity of the test taker's ratings to SME ratings. SMEs rated these options for effectiveness and scores were calculated as the distance between test takers' ratings and the ratings of the SMEs. The scores were rescaled to a 1-5 scale where 1 equals low agreement with SMEs and 5 equals high agreement with SMEs.

There were 12 EO items administered across three scenarios. Item means ranged from 3.47 to 3.78 with an average item mean of 3.62. Corrected item-total correlations ranged from 0.35 to 0.78, with a mean of 0.53. Most items did not detract from the internal consistency of the overall EO item set. Cronbach's alpha for the set if an item was dropped ranged from 0.83 to 0.86. The overall internal consistency of the set of items was 0.85. Table 12 shows the number of responses per item, item means and standard deviations, corrected item-total correlations, and alpha if deleted for the 12 EO items.

Table 12

Complex Problem Solving: Evaluating Options Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
EO1.1	67	3.77	0.31	0.35	0.86
EO1.2	67	3.66	0.35	0.48	0.85
EO1.3	67	3.59	0.46	0.58	0.84
EO1.4	67	3.78	0.37	0.43	0.85
EO2.1	67	3.74	0.27	0.53	0.84
EO2.2	67	3.58	0.36	0.47	0.84
EO2.3	67	3.66	0.32	0.54	0.84
EO2.4	67	3.63	0.41	0.54	0.84
EO3.1	67	3.59	0.36	0.44	0.85
EO3.2	67	3.48	0.44	0.73	0.83
EO3.3	67	3.47	0.46	0.78	0.83
EO3.4	67	3.54	0.54	0.46	0.85

EO results were also calculated by scenario. The mean, standard deviation, and internal consistency reliabilities are given for each problem-solving scenario in Table 13. Correlations between scenarios were 0.39 (scenarios 1 and 2), 0.69 (scenarios 1 and 3), and 0.49 (scenarios 2 and 3), for an average correlation of 0.52.

Table 13

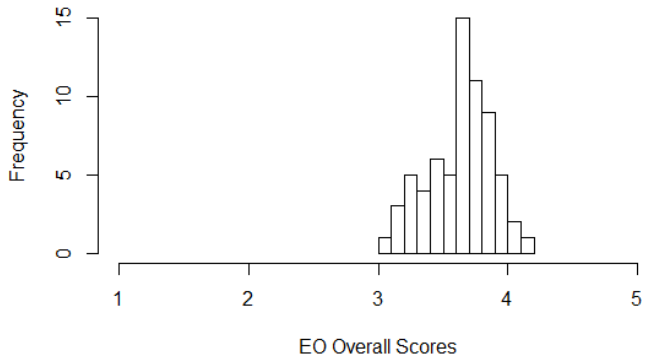
Complex Problem Solving: Evaluating Options Scenario Statistics

	Mean	SD	Cronbach's Alpha
Scenario1	3.70	0.25	0.61
Scenario2	3.65	0.28	0.83
Scenario3	3.52	0.35	0.77

The average EO score across scenarios was 3.63, with a standard deviation of 0.24. Overall ICP scores had a skewness of -0.48 and a kurtosis of 2.75. Figure 8 shows a histogram of the EO score distribution.

Figure 8

Histogram of Complex Problem Solving: Evaluating Options Scores



Intercorrelations between scenario-level scores and total scores for ICP and EO measures are given in Table 14. Correlations between scores for each scenario/dimension and related constructs such as Problem Solving, Critical Thinking, and Cognitive Ability are presented in Table 15.

As shown in the Complex Problem Solving correlation matrix, the correlations between ICP scores by scenario ranged from .00 to .24. The correlations between EO scores by scenario ranged from .39 to .69. Corresponding scenarios of ICP and EO did not significantly correlate (Scenario 1: $r = .19$, $p = .11$; Scenario 2: $r = .11$, $p = .37$; Scenario 3: $r = .13$, $p = .31$), and the correlation between the total scores for the ICP and EO phases of Complex Problem Solving phases was not significant ($r = .22$, $p = .07$).

Complex Problem Solving was expected to correlate with Critical Thinking and outside measures of Problem Solving and Cognitive Ability. Complex Problem Solving facets were also correlated with an IT experience scale in order to determine whether or not elements of the Complex Problem Solving tasks were similar to IT tasks.

Table 14

Complex Problem Solving Means, Standard Deviations, and Correlations with Confidence Intervals

	M	SD	ICP 1	ICP 2	ICP 3	EO 1 Mean	EO 2 Mean	EO 3 Mean	CPS ICP
ICP 1	0.12	0.43							
ICP 2	0.53	0.49	.24 [-.00, .45]						
ICP 3	0.23	0.36	-.00 [-.24, .24]	.14 [-.11, .36]					
EO 1 Mean	3.70	0.25	.19 [-.05, .42]	.19 [-.05, .41]	.02 [-.22, .26]				
EO 2 Mean	3.65	0.28	.09 [-.15, .33]	.11 [-.13, .34]	-.05 [-.29, .19]	.39** [.17, .58]			
EO 3 Mean	3.52	0.35	.12 [-.12, .35]	.25* [.01, .46]	.13 [-.12, .36]	.69** [.54, .80]	.49** [.28, .65]		
CPS ICP	0.29	0.28	.65** [.49, .77]	.76** [.64, .85]	.51** [.31, .67]	.22 [-.02, .44]	.09 [-.15, .33]	.27* [.03, .48]	
CPS EO	3.63	0.24	.16 [-.09, .38]	.22 [-.02, .44]	.04 [-.21, .27]	.78** [.66, .86]	.81** [.71, .88]	.87** [.80, .92]	.22 [-.02, .44]

Note. M and SD are used to represent mean and standard deviation, respectively (n = 67). Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). ICP 1, 2, 3 = Investigating Complex Problem Solving Scenarios 1, 2, and 3, EO 1, 2, 3 = Evaluating Options Scenario s 1, 2, and 3, CPS ICP = Complex Problem Solving: Investigating Complex Problems, CPS EO = Complex Problem Solving: Evaluating Options.
*$p < .05$. **$p < .01$.

Table 15

Complex Problem Solving Correlations

	ICP1	ICP2	ICP3	CPS ICP	EO 1	EO 2	EO 3	CPS EO
CT	0.05	0.28*	0.07	0.22	0.25*	0.12	0.24	0.24
IT XP	-0.05	-0.04	0.01	-0.04	0.22	0.08	0.28*	0.22
PDRI PS	0.09	0.42**	0.03	0.32*	0.33*	0.27	0.33*	0.36*
PDRI CA	0.31*	0.49**	0.10	0.51**	0.36*	0.40**	0.32*	0.43**

Note. When PDRI measures were involved, $n = 48$, otherwise $n = 67$. CT = Critical Thinking Total Score, IT XP = IT Experience, PDRI PS = PDRI Problem Solving, PDRI CA = PDRI Cognitive Ability, ICP 1, 2, 3 = Investigating Complex Problem Solving Scenarios 1, 2, and 3, CPS ICP = Complex Problem Solving: Investigating Complex Problems, EO 1, 2, 3 = Evaluating Options Scenarios 1, 2, and 3, CPS EO = Complex Problem Solving: Evaluating Options.
*$p < .05$. **$p < .01$.

Referring back to the correlation matrix in Table 4, both Complex Problem Solving measures were significantly correlated with Active Learning LE and the Inductive Reasoning test. Correlations with Active Learning LE were ($r = .38$, $p < .001$) for ICP and ($r = .27$, $p = .03$) for EO. Correlations with the Inductive Reasoning were ($r = .36$, $p < .001$) for ICP and ($r = .29$, $p = .02$) for EO. The ICP was also related to the Active Learning III ($r = .36$, $p < .001$), but EO was not significantly related to Active Learning III ($r = .16$, $p = .20$).

Total scores for both facets of Complex Problem Solving were significantly related to cognitive ability scores, with correlations of ($r = .51$, $p < .001$) and ($r = .43$, $p < .001$) for ICP and EO, respectively. ICP and EO were also significantly correlated with Problem Solving, with correlations of ($r = .32$, $p = .03$) and ($r = .36$, $p = .01$), respectively.

Critical Thinking

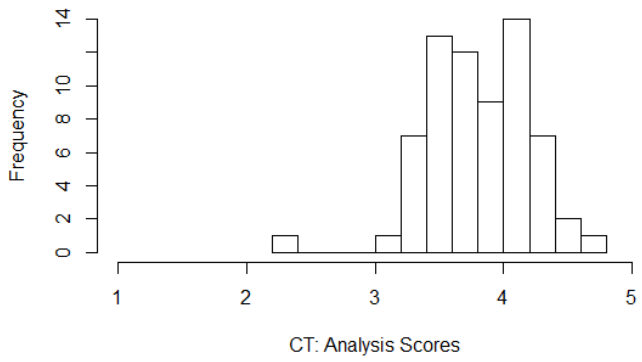
The Critical Thinking (CT) measure consisted of three scenarios that were related to but distinct from the training material (e.g., newspaper articles). Test takers reviewed the material for a scenario and were then presented with prompts asking the extent to which statements analyzed, interpreted, evaluated, or explained the scenario. Two rating scales were used for the questions, depending on the specific dimension, ranging either from 1 = not at all important to 5 = extremely important, or 1 = not at all to 5 = very great extent. SMEs rated these options, and scores were calculated as the distance between test takers' ratings and the ratings of the SMEs. Test takers rated 68 Critical Thinking statements in total, and scores were rescaled to a 1-5 scale where 1 equals low agreement with SMEs and 5 equals high agreement. Each dimension was addressed separately in the analyses, and results are presented by dimension. Statement-level statistics are provided in Appendix C.

Statement means for the Analysis dimension ranged from 2.71 to 4.60, with an average item mean of 3.75. Of the 24 items measuring the Analysis dimension, 10 had a corrected item-total correlation of less than 0.10. These were removed from the scale and the remaining items had corrected item-total correlations ranging from 0.17 to 0.38, with a mean corrected item-total correlation of 0.26.

Critical Thinking Analysis scores were calculated using the remaining 14 items. The distribution of Analysis scores is shown in Figure 9. These scores had a mean of 3.80 with a standard deviation of 0.40. Critical Thinking Analysis scores had a skewness of -0.62 and a kurtosis of 4.49.

Figure 9

Histogram of Critical Thinking: Analysis Scores



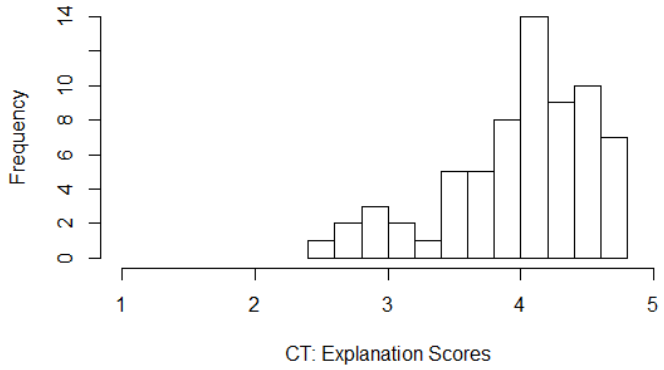
Statement means for the Explanation dimension ranged from 3.70 to 4.32, with an average item mean of 4.06. Of the five items, one had a corrected item-total correlation of less than 0.10. This item was removed from the scale, and remaining items had corrected item-total correlations ranging from 0.22 to 0.60, with a mean corrected item-total correlation of 0.44.

Critical Thinking Explanation scores were calculated using the remaining four items. The distribution of Explanation scores is shown in Figure 10. These scores had a mean of 4.00, with a standard deviation of 0.55. Critical Thinking Explanation scores had a skewness of -0.90 and a kurtosis of 3.07.

For the dimension of Evaluation, statement means ranged from 2.27 to 4.43, with an average item mean of 3.79. Of the 19 items, 14 had a corrected item-total correlation of less than 0.10. When the eight items with negative corrected item-total correlations were removed and item-total correlations recalculated, there were then five items with corrected item-total correlations less than 0.10. The remaining six items had corrected item-total correlations ranging from 0.17 to 0.26, with a mean corrected item-total correlation of 0.23.

Figure 10

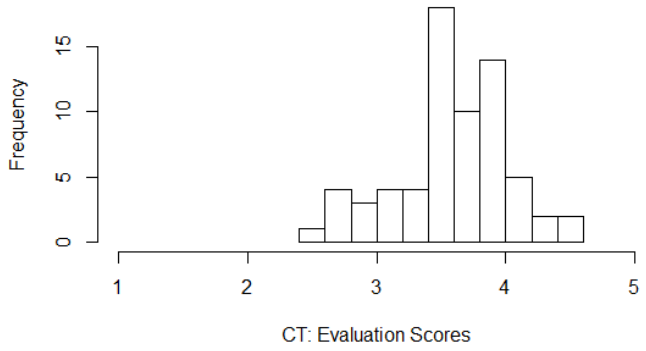
Histogram of Critical Thinking: Explanation Scores



Critical Thinking Evaluation scores were calculated using the remaining six items. The distribution of Evaluation scores is shown in Figure 11. These scores had a mean of 3.59, with a standard deviation of 0.43. Critical Thinking Evaluation scores had a skewness of -0.35 and a kurtosis of 2.95.

Figure 11

Histogram of Critical Thinking: Evaluation Scores

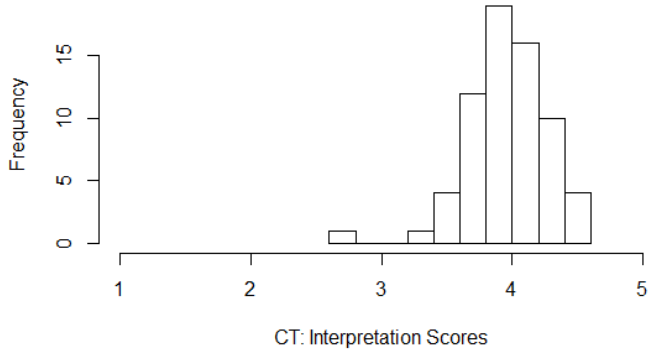


For the dimension of Interpretation, statement means ranged from 2.31 to 4.84, with an average item mean of 3.87. Of the 20 items, nine had a corrected item-total correlation of less than 0.10. When the five items with negative corrected item-total correlations were removed and item-total correlations recalculated, there were two items with corrected item-total correlations less than 0.10. The remaining 13 items had corrected item-total correlations ranging from 0.13 to 0.50, with a mean corrected item-total correlation of 0.23.

Critical Thinking Interpretation scores were calculated using the remaining 13 items. The distribution of Interpretation scores is shown in Figure 12. These scores had a mean of 3.96, with a standard deviation of 0.32. Critical Thinking Analysis scores had a skewness of -0.76 and a kurtosis of 5.71.

Figure 12

Histogram of Critical Thinking: Interpretation Scores



Scale mean, standard deviation, and internal consistency reliabilities for each Critical Thinking dimension are given in Table 16. Intercorrelations between Critical Thinking dimensions ranged from -0.11 to 0.29, with an average of 0.13. The observed correlation matrix is presented in Table 17.

Table 16

Critical Thinking Scale Statistics

	Final # of Items	Cronbach's Alpha
Analysis	14	0.65
Explanation	4	0.62
Evaluation	6	0.47
Interpretation	13	0.58

Table 17

Critical Thinking Means, Standard Deviations, and Correlations with Confidence Intervals

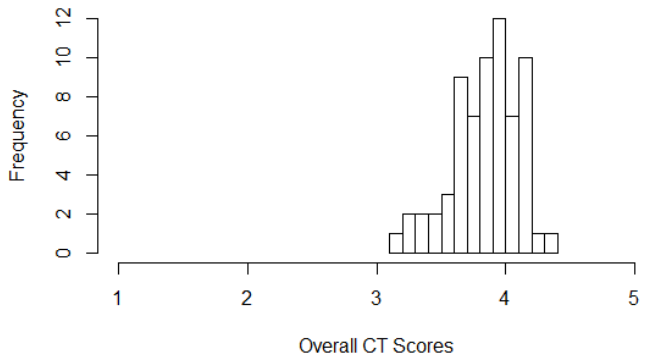
	M	SD	CT A	CT I	CT EV	CT EX
CT A	3.80	0.40				
CT I	3.96	0.32	.12 [-.12, .35]			
CT EV	3.59	0.43	-.11 [-.34, .13]	.29* [.05, .50]		
CT EX	4.00	0.55	.29* [.05, .50]	.09 [-.16, .32]	.10 [-.14, .33]	
CT	3.84	0.25	.54** [.35, .69]	.53** [.33, .68]	.53** [.33, .68]	.73** [.59, .83]

Note. M and SD are used to represent mean and standard deviation, respectively (n = 67). Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). CT A = Critical Thinking: Analysis Dimension, CT I = Critical Thinking: Interpretation Dimension, CT EV = Critical Thinking: Evaluation Dimension, CT EX = Critical Thinking, Explanation Dimension, CT = Critical Thinking Total Score.
*$p < .05$. **$p < .01$.

Scores were averaged across dimensions to get an Overall Critical Thinking score. The distribution of average Critical Thinking dimension scores is shown in Figure 13. These overall Critical Thinking scores had a mean of 3.84, with a standard deviation of 0.25. Overall Critical Thinking scores had a skewness of -0.55 and a kurtosis of 2.98. Internal consistency overall was .65.

Figure 13

Histogram of Overall Critical Thinking Scores



As shown in the correlation matrix for Critical Thinking, there were significant correlations between the Analysis and the Explanation dimensions ($r = .29$, $p = .02$), as well as the

Interpretation and Evaluation dimensions of Critical Thinking (r = .29, p = .02). The other correlations between the dimensions were not significant, with some being quite low (e.g. the Evaluation and Analysis dimension (r = -.11, p = .37).

Critical Thinking was expected to have a positive correlation with two dimensions of an outside measure of Critical Thinking Disposition (Critical Openness and Reflective Skepticism) due to the similarity of these constructs. However, since the Critical Thinking Disposition scale measures general tendencies and the C^3 Critical Thinking assessment measures outcomes of Critical Thinking performance, this correlation is expected to be small-to-moderate in magnitude. Critical Thinking was also predicted to correlate with Cognitive Ability and Problem Solving. Cognitive ability was expected to be an important antecedent of Critical Thinking, and Critical Thinking was expected to be an important antecedent of Problem Solving.

As shown previously in Table 4, Critical Thinking correlated significantly with both Active Learning III (r = .39, p < .001) and Active Learning LE (r = .32, p = .01). Critical Thinking was also significantly correlated with the Deductive Reasoning formula scores (r = .30, p = .01) and Inductive Reasoning (r = .28, p = .02). Table 18 shows the correlations of Critical Thinking with the other related constructs. Critical Thinking was significantly correlated with cognitive ability (r = .30, p = .04). The Critical Thinking Explanation dimension was the only dimension significantly correlated with cognitive ability (r = .32, p = .03). The other three dimensions did not have significant correlations with cognitive ability. None of the Critical Thinking dimensions were significantly positively correlated with problem solving; however, Critical Thinking overall was significantly related to the outside measure Problem Solving (r = .32, p = .03).

Table 18

Critical Thinking Correlations with Related Constructs

	CT A	CT I	CT EV	CT EX	CT
Critical Openness	.07	.01	.03	.03	.06
Reflective Skepticism	.12	.00	-.05	.09	.07
PDRI PS	.25	.21	.14	.21	.32*
PDRI CA	.14	-.07	.26	.32*	.30*

Note. When PDRI measures were involved, n = 48 otherwise n = 67. PDRI PS = PDRI Problem Solving, PDRI CA = PDRI Cognitive Ability, CT A = Critical Thinking: Analysis Dimension, CT I = Critical Thinking: Interpretation Dimension, CT EV = Critical Thinking: Evaluation Dimension, CT EX = Critical Thinking, Explanation Dimension, CT = Critical Thinking Total Score.
*p < .05. **p < .01.

Inductive Reasoning

Inductive Reasoning (IR) was assessed with a letter series completion task. Between 12 and 16 letters in a patterned sequence were shown as stimuli, and test takers had to discern which letters should appear next to continue the pattern. Responses could include any letter entered in a six-character open field, rather than use set response options. Eight Inductive Reasoning items were tested in this study. Item difficulties ranged from 0.16 to 0.81, with an average of 0.35. Item discriminations were calculated by contrasting the top and bottom thirds of the sample. The

difference in number correct from each of these groups, divided by the size of the group, served as the discrimination index. This index ranged from 0.50 to 0.77, with a mean of 0.63. Corrected item-total correlations ranged from 0.35 to 0.68, with a mean of 0.54.

Most items did not detract from the internal consistency of the Inductive Reasoning test. Cronbach's alpha with individual items removed ranged from 0.78 to 0.83. The overall internal consistency of the set of items was 0.82.

Table 19 shows the number of times an item was answered, item means (or difficulties in this case) and standard deviations, corrected item total correlations, item discrimination coefficients, and alpha if deleted for the eight Inductive Reasoning items.

Table 19

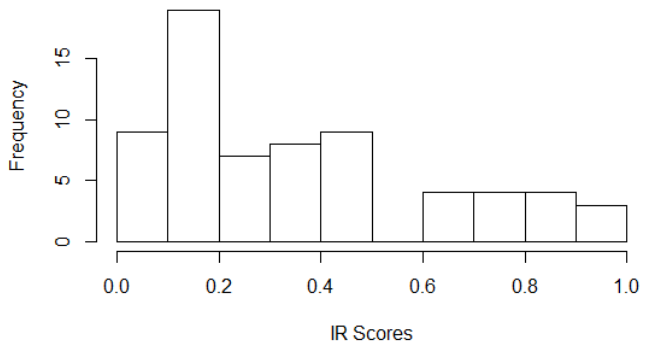
Inductive Reasoning Item Statistics

	n	Mean (Difficulty)	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
IR2	67	0.49	0.50	0.41	0.68	0.82
IR3	67	0.81	0.40	0.35	0.55	0.83
IR4	67	0.31	0.47	0.61	0.77	0.79
IR5	67	0.16	0.37	0.57	0.50	0.80
IR6	67	0.21	0.41	0.56	0.55	0.80
IR8	67	0.27	0.45	0.64	0.68	0.79
IR10	67	0.30	0.46	0.68	0.73	0.78
IR12	67	0.27	0.45	0.52	0.59	0.80

The overall mean and standard deviations of Inductive Reasoning scores were 0.35 and 0.29, respectively. The distribution of Inductive Reasoning scores had a skewness of 0.68 and a kurtosis of 2.39. Figure 14 shows a histogram of the Inductive Reasoning distribution.

Figure 14

Histogram of Inductive Reasoning Scores



Inductive Reasoning was expected to correlate with Cognitive Ability and Problem Solving. A large correlation was expected between Inductive Reasoning and Cognitive Ability, and a moderate correlation was expected between Inductive Reasoning and Problem Solving. Inductive Reasoning was also expected to correlate with Deductive Reasoning at a small to moderate magnitude, since both constructs reflected reasoning. Inductive Reasoning was also predicted to correlate with Troubleshooting. Correlations with overall scores and intermittent fault items were examined and can be found in Table 20.

Table 20

Inductive Reasoning Correlations

	Correlation with IR
PDRI Cognitive Ability	.47**
PDRI Problem Solving	.41**
PDRI Deductive Reasoning	.38**
TS Intermittent-type	.18
TS Efficiency Scores	.40**

Note. When PDRI measures were involved, $n = 48$ otherwise $n = 67$. TS = Troubleshooting.
**$p < .01$.

Inductive Reasoning was significantly correlated with Cognitive Ability ($r = .47$, $p < .001$), Problem Solving ($r = .41$, $p = .01$), Deductive Reasoning ($r = .38$, $p = .01$), and the Troubleshooting Efficiency score across all item types ($r = .40$, $p = .01$). Inductive Reasoning was not significantly correlated with Troubleshooting scores on intermittent items ($r = .18$, $p = .14$).

Deductive Reasoning

Seven Deductive Reasoning (DR) items were tested in this study, each consisting of between eight and 11 statements. Each statement set consisted of between one and two bias-prone statements, which were removed from the sets and analyzed separately. Two scoring approaches were used: (a) dichotomous scoring, in which statements were scored as either correct or incorrect, and (b) formula scoring, in which correct answers were given a point, those marked "I don't know" were neutral, and incorrect answers lost a point. Results for both scoring approaches are described below, beginning with dichotomous scoring.

Across all non-bias-prone statements, statement difficulties ranged from 0.19 to 1.00, with an average difficulty of 0.81. Item discriminations contrasting the top and bottom thirds ranged from -0.09 to 0.50, with a mean of 0.15. These results, along with item-total correlations and Cronbach's alpha for the set if an item was dropped, are presented by item in Appendix D.

Taking the proportion of correct statements per item, the average proportion correct across the seven items was 0.82, with a range from 0.74 to 0.94. The overall alpha was 0.56. Means, standard deviations, corrected item-total correlations, and Cronbach's alpha for the set if an item was dropped are presented in Table 21.

Table 21

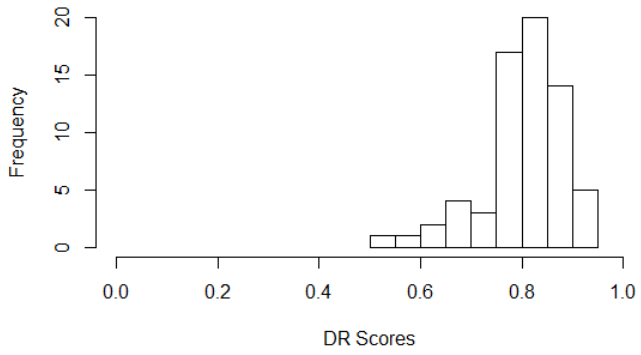
Deductive Reasoning Item Statistics (Dichotomous Scoring)

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
DR item B	67	0.76	0.17	0.21	0.54
DR item F	67	0.94	0.10	0.32	0.51
DR item K	67	0.78	0.18	0.36	0.48
DR item M	67	0.89	0.15	0.25	0.54
DR item N	67	0.74	0.22	0.09	0.59
DR item O	67	0.76	0.12	0.40	0.47
DR item P	67	0.87	0.14	0.28	0.51

The overall mean and standard deviation of the Deductive Reasoning scores was 0.82 and 0.08, respectively. The distribution of Deductive Reasoning scores had a skewness of 0.46 and a kurtosis of 1.58. Figure 15 shows a histogram of the Deductive Reasoning distribution.

Figure 15

Histogram of Dichotomous Deductive Reasoning Scores



The bias-prone Deductive Reasoning statements had item difficulties that ranged from 0.00 to 0.85, and a mean of 0.29 and a standard deviation of 0.15. Overall alpha for bias-prone items was 0.58. Item difficulties, corrected item-total correlations, item discrimination, and Cronbach's alpha for the set if an item was dropped are presented in Table 22. One statement was answered incorrectly by all test takers and was therefore omitted from the table.

Table 22

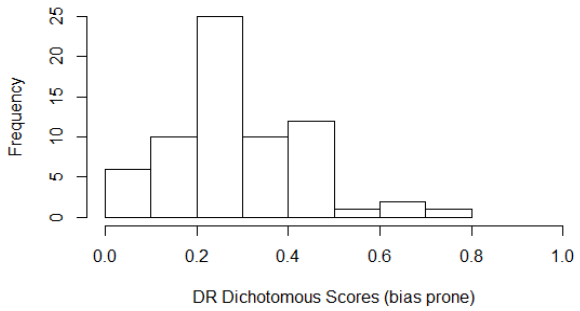
Deductive Reasoning Bias Prone Statement Item Statistics

	n	Mean (Difficulty)	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.B9	67	0.46	0.50	0.18	0.55	0.58
DR.F6	67	0.15	0.36	0.50	0.41	0.49
DR.F5	67	0.16	0.37	0.25	0.32	0.55
DR.K7	67	0.85	0.36	0.17	0.27	0.58
DR.M5	67	0.21	0.41	0.08	0.18	0.58
DR.M6	67	0.10	0.31	0.43	0.27	0.51
DR.N11	67	0.79	0.41	0.28	0.45	0.56
DR.O10	67	0.28	0.45	0.18	0.41	0.56
DR.O6	67	0.01	0.12	0.32	0.05	0.54
DR.P5	67	0.06	0.24	0.12	0.14	0.60
DR.P6	67	0.42	0.50	0.18	0.50	0.57

The distribution of bias-prone scores had a skewness of 0.81 and a kurtosis of 4.05. The histogram of these scores can be found in Figure 16.

Figure 16

Histogram of Bias-Prone Deductive Reasoning Dichotomous Scores



Using the formula scoring approach, Deductive Reasoning statement averages across test takers ranged from -0.99 to 1.00 across all statements. Regular statements averaged 0.66 across test takers, with a range of -0.58 to 1.00. Bias-prone statements averaged -0.36 across all test takers, with a range of -0.99 to 0.72. Test takers scored significantly differently on these two statement types $t(66) = 20.0$, $p < .001$. Table 23 shows means, standard deviations, corrected item-total correlations, and Cronbach's alpha for the set if an item was dropped.

Table 23

Deductive Reasoning Item Statistics (Formula Scoring)

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
DR item B	67	0.59	0.29	0.21	0.57
DR item F	67	0.90	0.19	0.38	0.52
DR item K	67	0.58	0.35	0.41	0.49
DR item M	67	0.80	0.28	0.24	0.57
DR item N	67	0.56	0.37	0.11	0.61
DR item O	67	0.59	0.19	0.43	0.49
DR item P	67	0.75	0.27	0.25	0.55

A histogram of averages of regular statements with formula scoring is presented in Figure 17. This distribution had a mean of 0.66 and a standard deviation of 0.15. The skewness was -1.52, and the kurtosis was 6.35. A histogram of bias-prone statements with formula scoring is presented in Figure 18. This distribution had a mean of -0.36 and a standard deviation of 0.29. The skewness was 0.83, and the kurtosis was 3.93.

Figure 17

Histogram of Average Deductive Reasoning Formula Scores

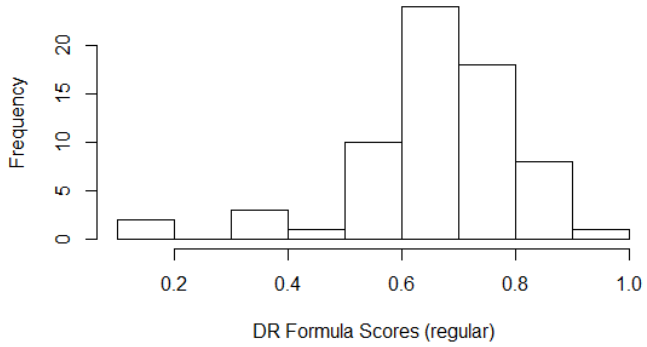
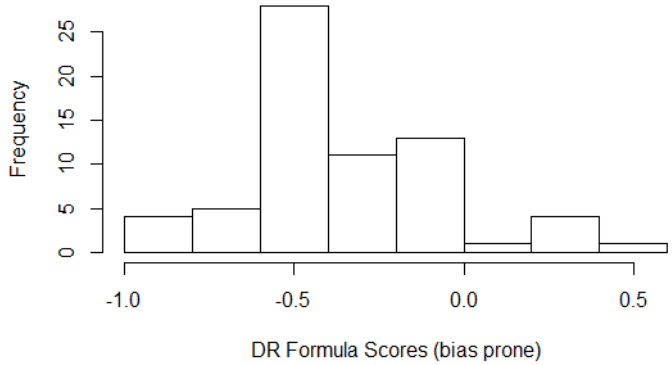


Figure 18

Histogram of Bias-Prone Deductive Reasoning Formula Scores



As shown in Table 24, the correlation between the two Deductive Reasoning scores was not significant (r = -.04, p = .76). The correlations between the dichotomous scores and formula scores were significant for both the regular (r = .91, p < .001) and the bias-prone (r = .94, p < .001) statements.

Table 24

Deductive Reasoning Means, Standard Deviations, and Correlations with Confidence Intervals

	M	SD	DR	DR Prone	DR FS
DR	0.81	0.08			
DR Prone	0.29	0.15	-.04		
			[-.28, .20]		
DR FS	0.66	0.15	.91**	-.13	
			[.86, .95]	[-.36, .12]	
DR Prone FS	-0.36	0.29	-.08	.94**	-.08
			[-.32, .16]	[.91, .97]	[-.31, .16]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). DR = Deductive Reasoning Dichotomous Scoring, DR Prone = Deductive Reasoning Dichotomous Scoring for Bias-prone items, DR FS = Deductive Reasoning Formula Score, DR Prone FS = Deductive Reasoning Formula Scoring for Bias-prone items. *$p < .05$. ** $p < .01$.

It was predicted that Deductive Reasoning should correlate with Troubleshooting because of the requirements for deduction inherent in the Troubleshooting task. Deductive Reasoning was also expected to correlate with an outside measure of deductive reasoning. Cognitive Ability was predicted to correlate with Deductive Reasoning due to the need for high-level cognitive processing during the deductive reasoning process. As shown previously in Table 4, formula-scored Deductive Reasoning Basic scores correlated significantly with Active Learning III ($r = .39$, $p < .001$), Complex Problem Solving ICP ($r = .39$, $p < .001$), and Critical Thinking ($r = .30$, $p = .05$).

With formula scoring, both the regular statement scores and bias-prone statement scores of Deductive Reasoning were significantly correlated with an outside measure of Deductive Reasoning ($r = .40$, $p = .01$ and $r = .36$, $p = .01$, respectively) (see Table 25).

Both were also correlated with Troubleshooting Efficiency ($r = .27$, $p = .03$ and $r = .26$, $p = .03$, respectively). None of the Deductive Reasoning scores correlated significantly with Troubleshooting Accuracy overall nor the accuracy on any of the three Troubleshooting item types. The basic score correlated significantly with Cognitive Ability ($r = .38$, $p = .01$), whereas the biased-prone score did not ($r = .24$, $p = .11$).

As shown in Table 25, the pattern of correlations found when using dichotomous scoring was very similar to pattern of correlations obtained using formula scoring. With dichotomous scoring, both the regular statement scores and bias-prone statement scores of Deductive Reasoning were significantly correlated with an outside measure of Deductive Reasoning ($r = .34$, $p = .02$ and $r = .29$, $p = .05$, respectively). Unlike with formula scoring, only bias-prone items were significantly correlated with Troubleshooting Efficiency ($r = .25$, $p = .04$); basic items were not significantly correlated ($r = .22$, $p = .07$). Neither basic nor bias-prone Deductive Reasoning dichotomous scores correlated significantly with Troubleshooting percent of faults found or the percent of

faults found on any of the three Troubleshooting item types. The basic score correlated significantly with cognitive ability (r = .32, p = .03); whereas the biased-prone score did not (r = .17, p = .25).

Table 25

Deductive Reasoning Correlations

	DR	DR Prone	DR FS	DR Prone FS
TS	0.14	0.23	0.15	0.21
TS EFL	0.14	0.14	0.17	0.16
TS EFN	0.35**	0.24*	0.35**	0.24*
TS EFI	0.03	0.23	0.11	0.24
TS EF	0.22	0.25*	0.27*	0.26*
PDRI DR	0.34*	0.29*	0.40**	0.36*
PDRI CA	0.32*	0.17	0.38*	0.24

Note. When PDRI measures were involved, n = 48 otherwise n = 67. TS = Troubleshooting % Faults Found, TS EFL = Troubleshooting Efficiency Linear Diagram, TS EFN = Troubleshooting Efficiency Network Diagram, TS EFI = Troubleshooting Efficiency Intermittent Diagram, TS EF = Troubleshooting Efficiency, PDRI DR = PDRI Deductive Reasoning, PDRI CA = PDRI Cognitive Ability, DR = Deductive Reasoning Dichotomous Scoring, DR Prone = Deductive Reasoning Dichotomous Scoring for Bias-prone items, DR FS = Deductive Reasoning Formula Score, DR Prone FS = Deductive Reasoning Formula Scoring for Bias-prone items.
*p < .05. **p < .01.

Selective Attention

Selective Attention (SA) was measured by examining test takers' responses to important and unimportant interruptions throughout the assessment. Selective Attention interruptions varied in the rate at which they were opened by test takers, with a minimum of 18% of the time and a maximum of 93% of the time. The average rate at which interruptions were opened by test takers was 58% of the time.

SMEs examined each interruption to determine if the appropriate response was to open and spend time on it (important interruptions) or to not open it (unimportant interruptions). Important interruptions were about 12% more likely to be opened by the participants than unimportant interruptions, on average across interruptions. A paired-sample t-test showed a difference in the rate of opening between important interruptions and unimportant interruptions (t(66) = 4.0, p < .05). A Wilcoxon signed-rank test, corrected for discontinuity in the rate differences, was similarly significant (V = 1042.5, p = .001). If we remove the first interruption, which was an unimportant interruption, then important interruptions were about 24% more likely to be opened than unimportant interruptions. The first interruption was opened by 84% of the sample, making it the second-most frequently opened interruption. Given that this interruption occurred very early in the assessment, and test takers might not have been familiar enough with the Selective Attention task at that point to effectively decide whether or not to open the interruption, this interruption was dropped from the calculation of the scores.

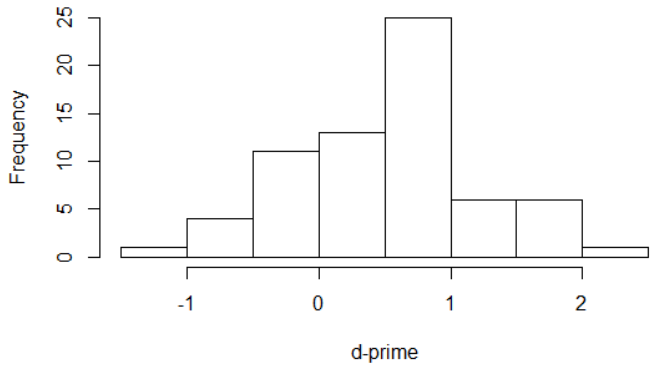
For the remaining interruptions, a d-prime index was calculated for each test taker to indicate sensitivity to appropriate interruptions. SMEs rated interruptions as important or not important to

attend to. Important interruptions that were opened were counted as hits, while unimportant interruptions that were opened were counted as false alarms.

The average Selective Attention d-prime score was 0.54, with a standard deviation of 0.68. Bias across individuals was c = .17. The average hit rate across individuals was HR = .57. The average false alarm rate across individuals was FA = .37. The distribution of d-prime scores had a skewness of -0.07 and a kurtosis of 2.9. Figure 19 shows a histogram of the Selective Attention d-prime scores.

Figure 19

Histogram of Selective Attention d-prime Scores



Due to a technical issue, interruption pages that were opened were cut off after one page of depth. This restricted test takers' ability to delve into the information beyond the first page. Therefore, the amount of time spent on an interruption was not analyzed and only the first decision above - whether to open the interruption or ignore it - was analyzed in the present study.

It was predicted that individuals who did well on the Selective Attention measure should also do well on the C^3 Complex Problem Solving measures, but not necessarily excel on an outside problem-solving measure where problem items were not interactive. Both Selective Attention and Complex Problem Solving require making judgements about what to pay attention to and are both expected to correlate with cognitive ability.

As shown in Table 26, Selective Attention did not correlate significantly with any of the other constructs. Selective Attention had particularly low correlations with the Complex Problem Solving EO (r = -.05, p = .66) and was not significantly correlated with Cognitive Ability (r = .15, p = .31) or Problem Solving (r = .08, p = .62). As shown earlier in Table 4, Selective Attention was not significantly correlated with Critical Thinking (r = -.03, p = .79) or the Deductive Reasoning basic statements (r = -.02, p = .84).

Table 26

Selective Attention Correlations

	Correlation with SA
CPS ICP	.17
CPS EO	-.05
PDRI PS	.08
PDRI CA	.15

Note. All correlations n.s. $n = 48$ when PDRI measures were involved, otherwise $n = 67$. CPS ICP = Complex Problem Solving Investigating Complex Problems, CPS EO = Complex Problem Solving Evaluating Options, PDRI PS = PDRI Problem Solving, PDRI CA = PDRI Cognitive Ability, SA = Selective Attention.

Troubleshooting

Troubleshooting (TS) was measured using a fault diagnosis task, with three different types of fault diagrams: two with different shapes (linear and networked), and one with faults that were intermittently active. The main metrics for Troubleshooting were indices of accuracy and efficiency. The accuracy score was calculated from the percentage of faults found within the network out of the number present. Using this index, item statistics for the nine Troubleshooting items are presented in Table 27. Item means ranged from 34.3% to 97.0%, with an average item mean of 75.5%. Corrected item-total correlations ranged from .10 to .50, with a mean of .37 across all Troubleshooting items. Overall Cronbach's alpha was .73. Figure 20 shows a comparison of Troubleshooting Accuracy means based on item type.

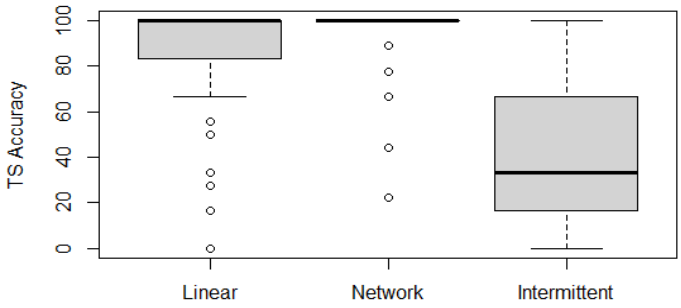
Table 27

Troubleshooting Accuracy Item Statistics

	Item Type	n	Mean Accuracy (%)	SD	Corrected Item-Total Corr	Alpha if Dropped
Item1	Linear	67	82	39	0.31	0.72
Item2	Linear	67	87	26	0.49	0.66
Item3	Linear	67	92	25	0.50	0.66
Item4	Network	67	97	17	0.37	0.69
Item5	Network	67	96	21	0.37	0.69
Item6	Network	67	96	15	0.41	0.69
Item7	Intermittent	67	34	48	0.10	0.76
Item8	Intermittent	67	40	40	0.28	0.73
Item9	Intermittent	67	56	43	0.48	0.69

Figure 20

Troubleshooting Accuracy Scores by Item Type



The efficiency score was calculated using the percent of faults repaired divided by the number of repairs ordered. Using this index, item statistics for the nine Troubleshooting items are presented in Table 28. Item means ranged from 0.10 to 0.57, with an average item mean of 0.35. Corrected item-total correlations ranged from 0.16 to 0.64, with a mean of 0.45 across all Troubleshooting items. Overall Cronbach's alpha was 0.76. Figure 21 shows a comparison of Troubleshooting Efficiency means based on item type.

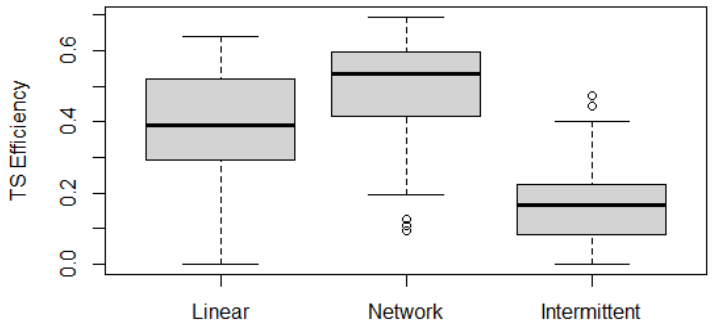
Table 28

Troubleshooting Efficiency Item Statistics

	Item Type	n	Mean Efficiency	SD	Corrected Item-Total Corr	Alpha if Dropped
item1	Linear	67	0.35	0.19	0.39	0.74
item2	Linear	67	0.32	0.22	0.62	0.71
item3	Linear	67	0.51	0.25	0.59	0.71
item4	Network	67	0.36	0.19	0.46	0.73
item5	Network	67	0.57	0.17	0.64	0.70
item6	Network	67	0.56	0.20	0.54	0.72
item7	Intermittent	67	0.10	0.17	0.25	0.77
item8	Intermittent	67	0.10	0.11	0.16	0.78
item9	Intermittent	67	0.31	0.25	0.37	0.74

Figure 21

Troubleshooting Efficiency Scores by Item Type

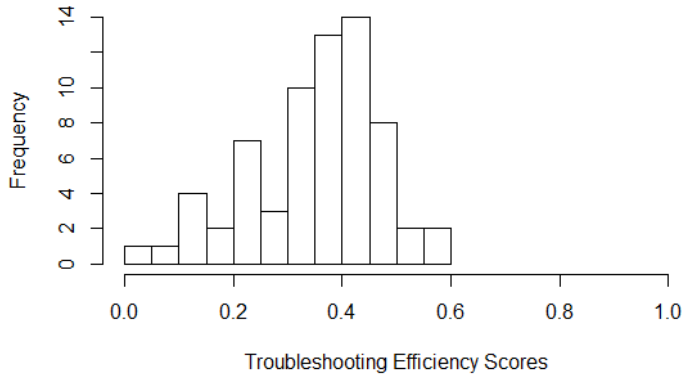


A one-way, within-person analysis of variance was conducted to evaluate differences in Troubleshooting item types. This analysis showed that there were differences between average scores per type across the three network types ($F(2,49) = 158.85$, $p < .001$).

The mean and standard deviation of Troubleshooting Efficiency were 0.35 and 0.12, respectively. Efficiency scores had a skewness of -0.64 and a kurtosis of 2.86. Figure 22 shows a histogram of the Troubleshooting Efficiency score distribution.

Figure 22

Histogram of Troubleshooting Efficiency Scores



Network and linear diagram items were correlated ($r = .66$, $p < .001$). Linear items correlated with intermittent items ($r = .37$, $p < .001$). Network items correlated with intermittent items ($r = .31$, $p = .01$). See Appendix E for intercorrelations among Troubleshooting item types.

The Troubleshooting task required that test takers deduce the location of faults, so it was predicted that Deductive Reasoning would correlate with Troubleshooting. Critical Thinking, Problem Solving, and Cognitive Ability were also predicted to correlate with Troubleshooting.

Troubleshooting scores were correlated with the C^3 Deductive Reasoning ($r = .27$, $p = .03$), as well as an outside Deductive Reasoning measures ($r = .44$, $p < .001$; see Table 29). Troubleshooting was not related to Critical Thinking or an outside measure of Problem Solving ($r = .06$, $p = .63$ and $r = .17$, $p = .25$, respectively). Troubleshooting was significantly correlated with cognitive ability ($r = .46$, $p < .001$).

Table 29

Troubleshooting Correlations with Related Constructs

	TS	TS EF	TS EFL	TS EFN	TS EFI
DR FS	0.15	0.27*	0.17	0.35**	0.11
CT	-0.07	0.06	0.08	0.11	-0.07
PDRI DR	0.34*	0.44**	0.34*	0.54**	0.18
PDRI PS	0.12	0.17	0.16	0.23	0.01
PDRI CA	0.30*	0.46**	0.44**	0.51**	0.14

Note. When PDRI measures were involved, $n = 48$, otherwise $n = 67$. DR FS = Deductive Reasoning Formula Score, CT = Critical Thinking, PDRI DR = PDRI Deductive Reasoning, PDRI PS = PDRI Problem Solving, PDRI CA = PDRI Cognitive Ability, TS = Troubleshooting % Faults Found, TS EF = Troubleshooting Efficiency, TS EFL = Troubleshooting Efficiency Linear Diagram, TS EFN = Troubleshooting Efficiency Network Diagram, TS EFI = Troubleshooting Efficiency Intermittent Diagram.
*$p < .05$. **$p < .01$.

CHAPTER 4: DISCUSSION

In this research, a series of measures were evaluated that assess seven constructs identified as important for cyber-related jobs. Results indicated that initial evidence for the C^3 battery is promising. Most measures had sufficient variance and scores with distributions that were approximately normal. Reliability evidence was mixed, with some construct's scores showing adequate levels of internal consistency and others showing a lack of internal consistency. As will be discussed, internal consistency was not an appropriate form of reliability for a few of the measures and more data are needed to properly assess reliability. The correlations among the C^3 constructs and sub-dimensions showed many expected patterns, providing some confirmation of construct validity, and suggesting a distinction between the C^3 measures and general intelligence. Although test takers reported finding the test complex and challenging to complete, this is, in part, necessary for a test that focuses primarily on assessing cognitive skills. We will discuss in further detail test taker feedback, the intercorrelations among constructs in the C^3 Test as a whole, and the specific results for each construct.

Test Taker Feedback

Test taker feedback was an important component of this study because many non-standard testing approaches were used. Feedback was collected with ratings and open-ended, qualitative responses. Generally, the ratings suggested that test takers found the assessment to be a challenge. Most test takers disagreed with the statements; "the assessment was easy to understand" and "learning the system was quick or would be quick for most people."

We expected that test takers would find the assessment difficult, as the assessment was designed to require learning and incorporate the need to filter though large amounts of information of varying relevance. We also expected that answering the questions would require focused thinking, as many of the tests were cognitive in nature. According to participants, the most challenging aspects of the assessment were: the amount of information and learning the technical terms, determining the relevant information, handling the interruptions, and answering the logic questions. Given that each of these factors is part of a measurement approach for C^3 constructs, the degree of challenge that test takers reported seems appropriate. The test was designed to require that test takers learn jargon and technical terms while sorting through relevant and irrelevant information. If the learning material was not challenging at all, there would be less variance from one person to the next. Similarly, the logic questions and handling of interruptions had to present a challenge sufficient to measure individual differences.

Despite the innate cognitive challenge of the assessment, to the extent possible we are striving to develop an assessment that test takers find easy to use. Most participants, however, disagreed with the statements: "the assessment was easy to use," "the assessment worked as they expected," and "it was easy to use the control panel to find information." In addition, most people also agreed with the statement that "the assessment was unnecessarily complex." These results suggest that improvements to the assessment battery should be considered that would improve the ease of use. For the next version of the C^3 battery we will therefore consider changes that can be made to improve ease of use, but not significantly decrease the challenge of the construct measures themselves. Changes to each measure are proposed below.

C^3 Intercorrelations

Overall, the pattern of relationships observed among the C^3 Test constructs was acceptable. Given the underlying cognitive nature of the constructs, overlap among the constructs should be expected, and was evidenced in significant correlations found among measures. The C^3 Test, however, uses construct measurement approaches that aim to distinguish between constructs by emphasizing the nuances that might allow prediction above and beyond general cognitive ability. As such, finding either relatively low magnitudes of correlations or nonsignificant relationships between the constructs was the goal.

The construct with the greatest number of significant intercorrelations was Inductive Reasoning (IR), which was significantly correlated with all other C^3 constructs except Deductive Reasoning (DR) and Selective Attention (SA). Serial completion tasks such as the C^3 Inductive Reasoning task are commonly used to assess fluid intelligence, so Inductive Reasoning can be viewed as a building block for other cognitive tasks. In addition, both Active Learning dimensions were significantly correlated with five other constructs. Neither Active Learning, Identifying Important Information(III) nor Active Learning, Learning Effectiveness (LE) was correlated with Selective Attention or Troubleshooting (TS), and in addition, Active Learning III was not correlated with Complex Problem Solving, Evaluating Options (EO) and Active Learning LE was not correlated with Deductive Reasoning. Inductive Reasoning and the two Active Learning constructs were likely driven by general intelligence more than any other constructs in the battery, which was reflected in the pervasiveness of their correlations with other constructs.

The constructs that had the least overlap with the other C^3 constructs were Complex Problem Solving EO, Troubleshooting, and Selective Attention. Complex Problem Solving EO was only correlated with Active Learning LE and Inductive Reasoning. Troubleshooting was only correlated with Inductive Reasoning. Selective Attention was not correlated with any other C^3 constructs. This lack of overlap was unexpected, especially the complete lack of correlations of constructs with Selective Attention. The descriptive properties and patterns of correlations for each construct will be discussed in further detail.

Active Learning

Two dimensions of Active Learning were measured: Identifying Important Information (III) and Learning Effectiveness (LE). Both of these dimensions were measured in the learning phase and again in the application phase.

Active Learning III

Active Learning III was conceptualized as a dimension of active learning that allows an individual to recognize that information is important, even if it is not exactly clear how it will be leveraged. It was measured with six items in both phases of the battery. Before standardization, means were higher for the learning phase items than the application phase items, suggesting that the learning phase items were easier. In part, this difference stems from differences in the response options for the two sets. The learning phase set contained fewer negatively-valued distractor boxes for test takers to select, and more positively-valued important information boxes.

As a result, possible scores in the learning phase could only go as low as -2.74. While standardization helps alleviate this problem for analyses, future iterations of the III measure should be modified to address these differences. First, care should be taken to reduce differences in the number and quality of response options in both item sets. Using the data collected in this study, we can identify items that are very frequently or very infrequently selected by test takers and analyze the viability of distractors within an item. Items with few viable distractors or too many high-value selections, should be modified to increase item difficulty.

In addition, differences in role clarity in the learning and application phases might contribute to differences in III scores for the two phases. In the learning phase, the instructions were clear that the test taker's role was to filter the information and learn the most important facts for use later. In the application phase, the information that was given was more specific and situational. Test takers needed deeper knowledge and an understanding of the problem scenarios to begin gathering information for problem-solving activities. This information gathering process was likely impeded by heightened uncertainty or problem-solving ambiguities. If this is the case, scores on these two phases might be telling us slightly different things about recognizing important information capabilities.

The potential differences between learning during the two phases might have been one of the contributors to a suboptimal internal consistency reliability. Using Cronbach's alpha as a guidepost, this measure failed to reach the oft cited benchmark of .70 for research use. Given that validity is limited by reliability, additional work is needed to improve the internal consistency of the scores produced by the Active Learning III measure. An expansion of the item pool may increase internal consistency to research standards, but future research should also examine the dimensionality of the III construct for suggestions on how to further improve the measure.

Suboptimal reliability notwithstanding, Active Learning III average scores had a significant positive correlation with Active Learning LE items. This suggests that the III and LE scores were related to one another, but not to the extent that they would be considered redundant. The magnitude of this correlation was similar to the magnitude of the correlation between III scores and an outside measure of cognitive ability. This indicates that III scores are influenced by cognitive ability, but cognitive ability is not the sole driver of III performance. This is encouraging, as it suggests that III might contribute to prediction above and beyond measures such as the ASVAB or other intelligence tests. Interestingly, cognitive ability was related to III in the application phase, but not in the learning phase. One reason for this may be that the learning phase III items were easier for individuals, restricting the variability and correlation of learning phase items with other measures.

Active Learning III was also related to other C^3 constructs: Critical Thinking (CT), Deductive Reasoning, and Inductive Reasoning. These correlations suggest that the III constructs have a place in the nomological network of cognitively-driven individual capabilities. The overlap between III and the two reasoning constructs was expected, assuming a general mental ability underlies both sets of scores. Individuals who were better able to recognize important information throughout the assessment were also better at evaluating the logical validity of statements and detecting patterns in letter series.

As mentioned, III was most strongly correlated with Critical Thinking and Deductive Reasoning scores. Critical Thinking in the C^3 battery required reading scenarios and item stems, then judging statements for soundness of analysis, interpretation, evaluation, and explanation. In both Active Learning and Critical Thinking, information was interpreted, evaluated, and judged in context. The connection between III and Deductive Reasoning might be explained by the use of deductive selection strategies in the process of selecting statements on the III items. People who are good at Deductive Reasoning are likely good at maximizing points by selecting the most obviously important information first.

The correlation of Active Learning III with Complex Problem Solving, Investigating Complex Problems (ICP) suggests that individuals who were able to discern what information was important or relevant were more likely to be successful at problem investigation. It also suggests that performance on the application-phase problem-solving activities was impacted by learning activities throughout the assessment. The correlations between III from the learning phase and Complex Problem Solving ICP or EO scores were not significant, but III scores from the application phase were significantly related to both Complex Problem Solving dimensions. As with the comparison of III and cognitive ability, these results may have been caused by the lack of difficulty and range restriction in the learning-phase III scores. However, the correlations between application-phase III and the Complex Problem Solving dimensions should have also been higher due to the similarity of information that was used in these measures. The Complex Problem Solving tasks were administered in the application phase only, and ability to learn application-phase information should have had an impact on a test taker's ability to apply that information for problem solving.

Though Active Learning III scores from the application phase were related to Complex Problem Solving EO, the Overall scores across the two assessment phases were not significantly correlated with EO. Overall scores were also not significantly correlated with Selective Attention, or Troubleshooting. The small but nonsignificant correlations with Complex Problem Solving EO and Troubleshooting are likely reflections of real relationships but are not significant due to a smaller than expected effect size. If that is the case, there may be an underlying similarity between these measures based on general mental ability, but some distinction in how they require mental ability to be applied. This distinction should be explored further to help differentiate Active Learning III from the other measures as much as possible. The correlation between Active Learning III and Selective Attention was close to zero, suggesting the greatest degree of distinction between these measures. Selective Attention, as a skill that is required for maintaining and focusing attention in the face of distractions, might not be as important when responding to the distributed Active Learning items throughout the assessment. This is because when the Active Learning items appeared, the test taker could not do anything else until the item was answered, and there were no competing tasks to draw away the test taker's attention.

In addition to the modifications described already that are aimed at improving the descriptive statistics and distributions of Active Learning III scores, there are improvements to the Active Learning III measure that were suggested by test taker feedback and the correlations between constructs. One example is to further integrate the Active Learning III interface with the training and on-the-job simulation; by adding a Notepad tool that captures and displays all the statements that are selected in the Active Learning III items, we can allow the test takers to come back and use the information they identified as potentially useful. Moreover, this Notepad tool could allow

test takers to type their own notes about the training content or the application-phase problem-solving activities and refer to their notes later as appropriate. This may open additional avenues for measuring Active Learning, such as a stealth-assessment approach that examines the extent and types of information that are captured by test takers' notetaking.

Active Learning LE

The second Active Learning dimension, Active Learning LE, was also measured in the learning and application phases, but unlike III, items for LE were given in blocks at the end of the phase, as opposed to throughout the phase. Both learning and application phase blocks contained 15 multiple-choice items. Items in both sets had a range of difficulties, suggesting that the items measured LE across the range of values. However, the overall percent correct for each LE test was low (50.4% and 46.9% in the learning and application phases, respectively).

These scores suggest that the LE tests were difficult overall. Scores from the two phases were significantly correlated with each other, but the overlap was not extreme. The magnitude of this correlation can be explained by differences in the required content knowledge for the two phases, the learning setting (training versus "on-the-job"), and ambiguity about learning requirements. Since both learning- and application-phase LE tests had similar item difficulty ranges, it is not likely that one test had harder content than the other. Instead, these scores likely reflect the somewhat different learning processes required for the two phases.

Similar to III, LE showed suboptimal internal consistency. Unlike III, LE measures knowledge breadth and is better thought of as a formative construct than a reflective construct. As such, internal consistency is less of a consideration. More research is needed to demonstrate stability of scores over repeated administrations of the measure. Until such data are collected, the retest reliability of LE is unknown. This fact must be considered when interpreting LE correlations.

As mentioned above, the correlation between total scores for LE (averaged across phases) was significantly correlated with total scores for III. The learning-phase LE scores were not correlated with either learning or application phase III scores; however, application phase LE was significantly correlated with application phase III scores. Individuals who performed better on the knowledge test (LE) at the end of the application phase also performed better on the information selection task (III) throughout the application phase. A positive correlation was expected, as both tasks required learning the application phase material, and the lack of correlation between the LE and III scores for the learning phase was surprising. Given the conceptual similarity between these constructs, it is likely that this correlation would be significant in a larger sample, however, a small correlation is preferred to reduce measurement overlap. Both III and LE were significantly correlated with cognitive ability.

Average LE scores across the two phases were significantly correlated with Complex Problem Solving ICP and EO. This indicates that the test takers who did a better job learning the important information throughout the assessment, did better on the complex problem-solving portion of the assessment. This result may be due to the relationship of Complex Problem Solving with ability to learn, or it could be due to specific knowledge that was learned, which facilitated the Complex Problem-Solving process. While LE tests from both phases were correlated with Complex Problem Solving ICP scores, only the application phase LE scores were

correlated with EO scores. This may be due to an increased judgment requirement for the application-phase LE test above the learning-phase LE test, since the information in the application phase had to be learned "on-the-job" as opposed to being delivered through training.

Average LE scores across the two phases were also significantly related to Critical Thinking scores and Inductive Reasoning scores. These correlations are likely partially driven by the correlation of these constructs with cognitive ability; however, Critical Thinking was more strongly related to LE than it was related to cognitive ability, suggesting a deeper relationship between these two constructs. The correlation between LE and Inductive Reasoning might stem, in part, from the need to induce what information was going to be important to remember, making Inductive Reasoning a potentially important part of the Active Learning process. Since the knowledge reflected in the LE tests was not used in the Critical Thinking or Inductive Reasoning tests, the correlations between these constructs and LE is likely a reflection of ability to learn, as opposed to the possession of specific knowledge.

Average LE scores were not significantly related to C^3 construct measures of Deductive Reasoning, Selective Attention, or Troubleshooting. This provides some discriminant validity, since there is an expected distinction between the deductive logic processes required for Deductive Reasoning and performance learning and recognizing information. The Troubleshooting measure, similarly, had a deductive logic requirement, stemming from the need to evaluate information from the fault diagrams and deduce the source of faults. Therefore, the divergence between Troubleshooting and LE also conforms to expectations.

Selective Attention was also expected to diverge from LE. In the learning phase especially, there was little requirement for test takers to selectively deploy their attention beyond what would be required to focus on the task. Test takers did need to concentrate on the training and filter through the less important information to focus on the relevant information, but all the training was linear and self-paced, making it somewhat sheltered from Selective Attention requirements. In the application phase, information was presented in a less linear fashion, so Selective Attention may be more of a requirement for LE performance. This is reflected in the data – learning-phase LE had no correlation with Selective Attention, while application phase LE has a small correlation with Selective Attention – but neither correlation was significant in the current sample.

Given the results of the LE test, modifications to the measure can be identified to address descriptive statistic results, as well as potentially improve the validity of the measure. With respect to the descriptive statistics, item difficulties and discriminations can be improved. While all items should be able to discriminate between individuals, a variety of difficulties would be advantageous, providing a range of information for identifying ability levels among test takers. Given the data, items can be selected with the best discriminations and a range of difficulties. A few items that were low in difficulty can be modified by improving the distractors and retesting their qualities.

In order to improve the validity of the LE measure, steps could be taken to hone the item pool such that the questions most closely adhere to only the most important facts in training and application. This feature was part of the initial design of the LE items but may need to be reevaluated as the assessment battery has evolved. More development and testing could also be

used to further distinguish LE in learning situations versus application situations, where information requirements may differ. If distinct measures are warranted, items would be modified to reflect either learning or application, and scoring would specifically address these two facets. Finally, the LE items could be better integrated into the overall assessment battery so that they feel more like part of the game and less like a separate test. One suggestion for how to accomplish this is to make LE questions sound more like they could come from a pedantic colleague or supervisor quizzing the "new employee" (i.e., test taker). LE tests could be broken into smaller subsets and spread through the assessment as though they are being asked by multiple different colleagues.

Complex Problem Solving

Complex Problem Solving was measured in two ways, reflecting two stages of the Complex Problem-Solving process.

Complex Problem Solving: ICP

The first measurement approach was Investigating Complex Problems (ICP), which examined test takers' abilities to interpret a problem scenario, conduct searches for relevant information pertaining to the problem, and extract useful information from among large amounts of distractor information. Examining the extent of search activity first, we saw a lot of variability from person to person in how much searching was done. On average, test takers used correct locations to find information 14.43 times across the nine scenarios for which data were available. This equated to about one and a half searches per scenario, which was insufficient to find all the information pertaining to the scenario and recognize accurate information in the Complex Problem Solving ICP items. The distribution of this index was heavily skewed to the right, due to many test takers (44%) conducting less than 10 searches in correct locations. These individuals were neglecting the search activity entirely, a phenomenon which is reflected in the low Complex Problem Solving ICP d-prime scores described below.

Search activity can also be scored by the proportion of times that test takers searched the correct locations, regardless of the number of searches done in those locations. Here again, there was variability from person to person, but overall, the average was very low (mean percent of correct searches = 1.9). Twenty-two participants did not search in the correct location in the nine scenarios for which data are available. Combined with the search activity results above, these results suggest that a large proportion of the participants were not engaged in the search task. While this may have been influenced by test takers' understanding of the task objectives, it is most likely due to a lack of motivation on the part of many participants. Since this was not being used to select someone for a job and the test outcomes had no impact on test takers' careers or job placement, some study participants may have been unwilling to put in the effort required to find information in the databases.

Modifications to the search activity interface should yield improvements in the information that is gained from this portion of the test. When analyzing the search activities done by test takers in the Complex Problem Solving databases, we identified two blind spots where we did not have insights into the materials at which test takers were looking. In areas of the control panel called the Network Map and Alerts Monitor, the test takers could go in and find relevant information,

but they did not need to open any dialog boxes or expand any collapsed sections in order to see that information. Because of this, we have no record of the information that test takers accessed during their searches of these control panel areas. A layer of interaction capabilities is needed for both the Network Map and the Alerts Monitor control panel locations so that additional information can be captured about how the test takers are using these data sources. For example, in the Alerts Monitor, one approach is to have the alerts expand when clicked to show test takers more information about the problems that are occurring. In addition to these modifications, numerous changes to the search interfaces, designed to improve the ease of searching and increase searching activity, will be discussed after reviewing the ICP results.

The lack of search activity should be apparent in the outcomes of the Complex Problem Solving task. Aside from search activities, the approach to measuring Complex Problem Solving ICP used a d-prime sensitivity index reflecting test taker ability to recognize accurate information and ignore inaccurate information. The means on the d-prime index suggested that test takers had difficulty identifying accurate information. The overall mean across Complex Problem Solving ICP scenarios was only slightly above chance performance. Moreover, some test takers had out-of-range values in the form of negative d-prime scores. These negative scores were likely due to sampling error and chance deviations from a mean of 0.0 for those individuals. If numerous negative d-prime scores were present in a larger sample, it would suggest that test takers were confused regarding appropriate responses to the task (e.g., selecting the inaccurate information rather than the accurate information).

Like Active Learning LE, Complex Problem Solving ICP may be a formative construct given that the problems that needed investigation varied from scenario to scenario and required searching distinct locations for information about the problem. As a formative construct, internal consistency reliability is less of a concern; however, the present research was not able to address any other forms of reliability, so reliability information is missing for this measure. Additional inquiry and future data collection focused on the search activities that test takers engage in while solving the ICP items may reveal more insights into the construct and aid in a more complete specification of the construct's components. As with LE, the lack of reliability data should temper interpretations of other ICP results.

Complex Problem Solving ICP scenarios varied in difficulty and the correlations between scenarios were low. The low correlation between scenarios suggests that more research is needed to assess the reliability of scenarios and to determine the aspects of Complex Problem Solving that they are most closely measuring. If Complex Problem Solving ICP is a reflective construct, the internal consistency across scenarios is well below acceptable levels, which may be partially due to varying scenario difficulties. If Complex Problem Solving ICP is a formative construct, composed of performance on many different scenarios, additional consideration of types of scenarios would yield measurement improvements and customization.

In the other scenarios, means and internal consistencies were showing a similar picture. The mean for Scenario 1 was only slightly above zero or chance performance, suggesting that test takers had a lot of difficulty discriminating between the accurate and inaccurate statements for this scenario. For Scenario 2, test takers had slightly better means, reflecting an easier time finding and reporting the relevant problem-solving information and ignoring the irrelevant information. Both of these scenarios had Complex Problem Solving ICP scores that correlated

with cognitive ability, suggesting that test takers with higher cognitive ability were better able to find the relevant information, recognize the accurate information in the statements, or both. Scenario 2 scores also had correlations with Critical Thinking and an outside measure of problem solving, which was the expected result for all the scenarios. This suggests that Scenario 2 was working better and getting better responses from test takers than the other two scenarios.

Unlike the scenario subscores, the overall Complex Problem Solving ICP scores had significant correlations with other C^3 constructs. While the overlap with other C^3 constructs was moderate, it might have been more substantial with more variability in the Complex Problem Solving ICP scores. Complex Problem Solving ICP scores were correlated with both measures of Active Learning, which suggests that learning abilities are important for problem solving outcomes. These results conform to the literature on Complex Problem Solving, where a knowledge acquisition phase is recognized as an important component of success in complex situations (Fischer, Greiff, & Funke, 2012).

Complex Problem Solving ICP scores were also correlated with both C^3 reasoning measures of Inductive Reasoning and Deductive Reasoning. This may be due to the role of these processes in problem solving; for example, test takers had to use Inductive Reasoning to navigate the databases and were most successful when they were able to recognize patterns in data and make inferences about the trends they were seeing. This facet of Complex Problem Solving revolves around an investigation into an ambiguous, uncertain situation and test takers likely used Deductive Reasoning to aid in this detective work. Using what they knew about the context, they had to reason from one database clue to the next and keep in pursuit of key causes.

On the other hand, the relationships between Complex Problem Solving ICP and the two reasoning constructs might have been driven by a common relationship between these measures and cognitive ability. Cognitive ability was the strongest correlate of Complex Problem Solving ICP, overlapping with Complex Problem Solving ICP more strongly than any other C^3 construct or outside measure. More research is needed to examine these effects and determine the degree of overlap with cognitive ability.

A number of changes can be made to improve the Complex Problem Solving ICP measure, including reexamining the scenarios and databases to ensure that the information presented in the scenarios is clear and that the test taker can make meaningful progress exploring and uncovering useful information. Given that the Complex Problem Solving ICP scores were very low, it is possible that the problem-solving scenarios were too difficult in either getting an understanding of the situation, or in finding additional clues to improve one's understanding. Improvements to the scenarios and databases could make the items easier. Another way to make the items easier would be to modify the statements in the ICP items such that the accurate statements are easier to distinguish from the inaccurate statements. Identifying the appropriate level of difficulty is challenging; the Complex Problem Solving ICP items should not be too hard but should also not be so easy as to allow the test takers to guess the correct responses without performing some research into the scenario.

It is also possible that the Complex Problem Solving ICP items were difficult because the task itself was unclear. One way to improve this is to include practice problems, so that the test takers can get some practice searching the databases and answering questions without it counting

against their score. These practice problems could be used to give the test taker some early "wins" on the task, providing some scaffolding for how the search process should unfold, and giving them a preview of the kinds of information for which they should be searching. In addition to including practice problems, numerous suggestions were provided by test takers to improve the search interface so that it is easier to find information. These improvements range from simply increasing the amount and quality of instructions to improving the way information is indexed and how it is returned from the databases. After incorporating some of these modifications in subsequent versions, the Complex Problem Solving ICP measure should be greatly improved.

Complex Problem Solving: EO

The second dimension of Complex Problem Solving was EO. This was measured with a situational judgement test approach in which test takers rated the effectiveness of different courses of action in response to problem solving scenarios. In contrast to ICP, EO had higher means both overall and by scenario. Variability between individuals was lower for EO than for other constructs, suggesting a somewhat lessened ability of this measure to differentiate between individuals. Closer examination revealed that participants showed a central tendency bias and avoided the lower end of the scale in particular. This resulted in many scores that were near the middle of the distribution. EO, however, was correlated with other constructs, and therefore did seem to produce meaningful variability between individuals.

Overall EO scores had an acceptable level of internal consistency reliability for 12 items. For two of the three broad problem scenarios, internal consistency was sufficient, but scores from the first problem scenario were slightly below the benchmark of .70. At only four items for each of these scenarios, all three item sets showed relatively high cohesion among the items. Additional data establishing retest stability would further bolster the evidence for EO's reliability.

EO was related to fewer C^3 constructs than ICP, with significant relationships with only Active Learning LE and Inductive Reasoning. These results suggested that test takers that learned the training content and application phase information more effectively were better able to recognize effective courses of action in the problem-solving scenarios. Similarly, those with better ability to reason inductively were better able to recognize the appropriate courses of action. As with ICP, these relationships may have been influenced by the role of cognitive abilities since cognitive abilities played a role in both Inductive Reasoning and EO scores.

In order to address the central tendency bias, the next version of the EO test will include instructions to the test takers to try to use all points of the 5-point scale when making their ratings. In addition, statements can be developed and selected so that there are equal numbers of statements across the scale of effectiveness. Moreover, these statements should be selected such that the effectiveness of each course of action is not overly obvious but challenges the test taker to evaluate the action. Additional research is needed to demonstrate the construct and criterion-related validity of this measure. Subsequent iterations of the EO measure should investigate EO with different types of problems to examine the extent to which EO is a general ability, as opposed to an ability that is tied to specific domains or depends on specific knowledge.

Critical Thinking

Critical Thinking (CT) was measured using Likert-type ratings on four dimensions: Analysis, Interpretation, Evaluation, and Explanation. These were measured on three different occasions during the learning phase. Each measurement utilized an article (newspaper or magazine) from an outside, non-training source to deliver information to the test taker. Since this information did not come from official sources, the contents had to be critically evaluated and the situation had to be analyzed for the existence of opinion and biases. Individual Critical Thinking scores were calculated for each dimension, and an overall average was calculated across the dimensions.

Scores for the Critical Thinking dimensions were similar to each other, with generally low variance and distributions that were approximately normal, with slight negative skew and moderate kurtosis. Overall, these characteristics are favorable for the Critical Thinking measurement approach, though there was some central tendency bias in Critical Thinking responses, similar to results for Complex Problem Solving EO. The central tendency bias issue can be addressed with instructions as described for the Complex Problem Solving EO measure. In addition, future iterations of the measure should include additional statements to increase the pool of viable items.

Internal consistency of the dimension scales was generally low, with Cronbach's alphas below the acceptable range for applied uses. With the exception of the Explanation dimension, the statements for each dimension pertained to two or three different articles. Therefore, dimension scales are formative in the sense that they ask about different instances of Critical Thinking or different ways it can be expressed, rather than asking reflective questions about underlying factors. As such, internal consistency is a limited reflection of reliability. Additional research is needed to establish test-retest or another form of reliability as opposed to using Cronbach's alpha.

Bivariate relationships between the different Critical Thinking dimensions were small. Only two reached statistical significance in the current sample: the correlation between Evaluation and Interpretation, and between Analysis and Explanation. All other bivariate relationships were around the .10 level, suggesting little overlap between the dimensions, as was expected based on previous research on the dimensionality of Critical Thinking (e.g., Facione, 1990). Only the correlation between Analysis and Evaluation was negative, though the confidence intervals were wide and included zero. If the true correlation is indeed negative, this would suggest that people who are good at Analysis are not good at Evaluation, and vice versa.

Of the four dimensions, only Explanation was significantly related to cognitive ability. Analysis and Evaluation had small, positive correlations with cognitive ability, and Interpretation had a small, negative correlation with cognitive ability, but these did not reach statistical significance. It is surprising that the other correlations with cognitive ability are not larger, given the apparent cognitive nature of Critical Thinking. The overall Critical Thinking score, an average of scores from the four dimensions, was significantly correlated with cognitive ability. The magnitude was moderate, but again, it is surprising that it was not larger given the interrelationships between these two constructs. More research is needed to further examine the relationship between Critical Thinking and cognitive ability. If the divergence persists, this measure might contribute to prediction of important job-related outcomes beyond general mental ability.

None of the dimensions or the overall Critical Thinking scores were significantly correlated with two dimensions of the Critical Thinking Disposition Scale - Critical Openness and Reflective Skepticism. This is also surprising. Though these constructs are distinct, one might expect moderate correlations. This may be partially due to a power issue, but also to the critical thinking disposition dimensions that were used in this study, and the lack of overlap between personality for critical thinking and skills at critical thinking. Also, the format of these assessments could be driving the lack of overlap. The Critical Thinking Disposition Scale is self-report measure, and therefore subject to self-serving biases; whereas the C^3 assessment is performance-based. Moreover, the C^3 Critical Thinking assessment requires other capabilities that are not present in disposition scales, for example, reading comprehension and working memory capacity.

On average, the correlations were small for the four Critical Thinking dimensions and an existing measure of problem-solving abilities; however, all four dimensions failed to reach the .05 critical alpha level. The correlation between problem solving and the overall Critical Thinking average, on the other hand, was significant. This may be due to the role of Critical Thinking in problem solving, and probably also reflects the role of cognitive ability in both of these constructs. Further research is needed to examine these relationships. There is likely to be complex nested relationships among these constructs, such that some constructs are at a higher level and partially composed of the other constructs. For example, Critical Thinking may be an important part of Problem Solving, and cognitive abilities may be an important underlying part of both Critical Thinking and Problem Solving.

Inductive Reasoning

Inductive Reasoning (IR) was measured with a pattern completion task in which test takers were presented with a series of 12 to 16 letters and then discerned and entered the next six letters. The measure was administered at the end of the application phase using eight letter series completion items. Most items were difficult for test takers, with about a third of the sample answering each question correctly, on average. Difficulties ranged, however, showing a variety of difficulty levels. Overall internal consistency of the items was high. These properties together suggest that the Inductive Reasoning measure could be shortened from its eight-item length, especially if a cutoff threshold for Inductive Reasoning could be identified.

Means on the Inductive Reasoning scale were low, and the distribution showed a positive skew with some range restriction at the lower end due to the zero percent lower bound for scores. Larger sample sizes are needed to determine if scores would conform more closely to a normal distribution or if transformation would be beneficial.

Inductive Reasoning scores had an acceptable level of internal consistency reliability. Additional data establishing retest stability would further bolster the evidence for EO's reliability, but the current data suggest potential in this measurement approach. Moreover, the impact of the Inductive Reasoning distribution on correlation analyses seemed to be small, since many large correlations were found with the other measures. Inductive Reasoning was significantly correlated with all other C^3 constructs except Deductive Reasoning and Selective Attention. While some convergence between Deductive and Inductive Reasoning is feasible, these are distinct forms of logic, and should not be highly correlated except as they reflect more general underlying mental abilities. Similarly, Selective Attention is not conceptually related to Inductive

Reasoning, except through their relationships with mental abilities like fluid intelligence and attention control (e.g., Unsworth, Fukuda, Awh, & Vogel, 2014) (which should be smallest for Selective Attention of all the C^3 constructs). Inductive Reasoning was correlated with all the other C^3 measures.

In addition to pervasive correlations with other C^3 constructs, Inductive Reasoning was also related to outside measures of cognitive abilities, problem solving, and deductive reasoning. The significant positive correlations with Inductive Reasoning are expected, due to the central role of general fluid intelligence in Inductive Reasoning. The largest correlation among other C^3 constructs was with Active Learning LE scores. As mentioned above, there may be an element of inductive logic required for Active Learning, making these measures correlate slightly.

The pervasive correlations between Inductive Reasoning and other cognitively-loaded constructs and the moderately-sized correlations with cognitive ability and problem solving are consistent with the assumption that the Inductive Reasoning measure is measuring what it was intended to measure. No other measures of Inductive Reasoning were suitable for comparison in this study, so more direct evidence of convergent validity was not captured. However, the letter series completion task that was developed for the C^3 battery followed existing procedures for series completion task construction. These sound development procedures provide some evidence of the validity of the Inductive Reasoning measure.

One option identified for modifying the Inductive Reasoning assessment is to include a "no pattern" response option and items, where no discernable pattern would complete the series. There is a natural tendency to look for patterns, and patterns can be found even when they do not exist. Including the "no pattern" option would force test takers to evaluate whether there is a pattern in the letters and only look to complete the pattern when a pattern exists.

Another interesting addition to this type of test would be to include confidence ratings with each inductive reasoning question. Since inductive reasoning is not certain and requires a "leap of faith," it would be interesting to know whether the individual differences in one's confidence for inductively-derived solutions are important for cyber performance. Outside of the assessment situations, the human brain is typically ready to recognize patterns based on a small sample of data, even when those patterns do not exist. Difficult problems in complex environments require probabilistic reasoning and weighing of evidence. Unlike the typical serial completion task, in real-world situations it is often uncertain whether a derived solution is correct. Successful inductive reasoning in this context should reflect appropriate amounts of uncertainty in confidence ratings.

Deductive Reasoning

Deductive Reasoning (DR) was measured in the learning phase using seven items, each consisting of a one- to two-sentence premise and a list of statements derived from that premise. Test takers were asked to determine which statements were logically valid. Some of the statements within each item were more difficult and were analyzed separately as an index of bias-prone Deductive Reasoning.

Excluding bias-prone statements, scores on Deductive Reasoning were high, suggesting that it was not difficult for test takers to identify most logically valid statements and reject most logically invalid statements. Within each item, there were some statements that were answered correctly more infrequently, but most statements were frequently answered correctly. This phenomenon was the same for both dichotomous scoring and formula scoring approaches.

The bias-prone statements were considerably more difficult for test takers, conforming to the expectation that these statements are impacted by reasoning biases and more difficult to answer correctly. Though means on the bias-prone statements were low on average, there were some bias-prone statements that were answered correctly more often and therefore had higher means. Overall, a range of difficulties were observed, suggesting that this metric had the ability to measure people at different ability levels.

The two scoring methods produced very similar descriptive statistics and distributions of scores. The formula scores had somewhat lower means, and therefore, slightly better distributional characteristics. In particular, there was less possibility for range restriction on the upper end of the scale. Because of this advantage, formula scores were used for the correlation analyses.

Scores obtained from the Deductive Reasoning measure had an internal consistency reliability that was well below the threshold of acceptability. A partial remedy would be to add more items to the existing seven. Additionally, retest reliability data is needed to further bolster the reliability evidence for the measure. Due to the low internal consistency and lacking any other form of reliability data, the Deductive Reasoning results should be confirmed with additional research.

Among other C^3 constructs, basic Deductive Reasoning correlated with Active Learning III, Complex Problem Solving ICP, Critical Thinking, and Troubleshooting. Most of these correlations were small to moderate in magnitude, suggesting some degree of overlap, but not redundancy. The convergence of Deductive Reasoning and Troubleshooting is of particular interest because the Troubleshooting task requires reasoning about where the faults might exist in the network. In the Troubleshooting task, test takers searched for faults by taking action, then examining the outcome of that action and deducing what that outcome indicated about fault locations. Given this, it was somewhat surprising that Deductive Reasoning scores for both basic and bias-prone statements were not correlated with Troubleshooting scores on linear or intermittent-type items for basic and bias-prone items, respectively. Deductive Reasoning scores were significantly related to network-type items and overall Troubleshooting Efficiency scores. Linear-type items occurred first, and relative to network-type items, would not have benefited from practice effects. Unlike intermittent-type items, test takers performed well on network-type items, suggesting that characteristics of these two item types drove difficulty differences and different relationships with Deductive Reasoning.

Selective Attention

As described in the Results section, a technical issue with the Selective Attention (SA) measure caused the opened interruption pages to be cut off after one page of depth, restricting test takers' ability to delve into the information beyond the first page. Because of this restriction, only the initial decision – whether to open the interruption or ignore it – was analyzed in the study.

There were 16 interruptions administered in the C^3 battery: eight in the learning phase and eight in the application phase. In the initial instructions before the learning phase, and subsequent instructions at the onset of the application phase, test takers were told what kinds of information to focus on and how to spend their time to achieve high performance. The initial pop-ups announcing the interruptions were designed to reveal the main subject of the interruption so the test taker could evaluate whether an appropriate response was to view the information or ignore it.

Of the 16, four were rated as unimportant by SMEs and the other 12 were rated as important. Generally, test takers ignored the unimportant interruptions at a higher rate than the important interruptions. This suggests that the interruptions were having the intended effect, with test takers evaluating whether to attend to each distraction based on the instructions and what they knew about the objectives in each phase of the battery. However, there were also individual differences in the ability to discriminate between important and unimportant interruptions, as reflected in the distribution of d-prime scores. These qualities have a favorable impact on the likely success of the Selective Attention decision task as a measure of attention capabilities.

The present research was unable to provide reliability data, so reliability evidence for Selective Attention is currently lacking. One sensitivity score is obtained from all the attention decisions throughout the assessment, and with only one item, internal consistency cannot be calculated. Subsequent research should expand the measure with more interruptions in order to create items and calculate internal consistency. Additionally, test-retest reliability evidence should be gathered in future research to bolster validation of the measurement approach.

Selective Attention did not correlate significantly with any other C^3 constructs, and there were no available comparison measures to evaluate the construct validity of the Selective Attention measure. Selective Attention did not correlate with other outside measures like cognitive ability, deductive reasoning, or problem solving either. The matter of construct validity is therefore left unanswered, and this must be addressed by future research. Selective Attention's lack of demonstrable relationships with other constructs, however, is potentially a valuable asset which can facilitate the C^3 battery's ability to predict performance above and beyond general cognitive ability.

In the next iteration of the Selective Attention measure, the interruptions must be corrected so that they are delivered appropriately. Once these are functioning correctly, additional data are needed to examine how the measure is working. One outstanding question is whether the interruption lengths will be sufficient to provide the necessary distraction, or if the test taker will be able to click through and scan them without taking too much time. If they can be reviewed too quickly, the interruptions could be improved by adding more to do or including different branches of information to explore. In addition to this potential change, improvements for the next iteration of the assessment are to equalize the number of important and unimportant interruptions and to more clearly distinguish between the important and the unimportant interruptions by revising the content of the initial prompts. As with other changes to affect the difficulty, care must be taken to make sure that revisions do not make the decisions too obvious. Beyond these changes, additional research is needed to place this Selective Attention measure in the nomological network of related constructs. This measure is unique from other attention-based

measures, and little is currently known about how this measure relates to other attention constructs or similar constructs.

Troubleshooting

The final C^3 measure, Troubleshooting (TS), was measured with a fault diagnosis task. In this task, test takers examined a set of interconnected components and traced a malfunction in the set to one or a few faults. Scores varied by item type, with the intermittent-type items being the hardest for test takers to complete efficiently. This finding conforms to prior research and expectations that the items with intermittent faults would be the most difficult to complete. All three types of Troubleshooting items correlated with the overall Troubleshooting score, contributing to the internal consistency of the overall index. All item types correlated with each other, but linear and network type items had the most overlap.

The reliability of scores produced by the nine Troubleshooting items was low for operational use. The three item types are likely detracting from cohesion and additional network or linear items would likely improve internal consistency reliability. Nevertheless, the internal consistency that was obtained exceeded the commonly accepted threshold for research use of .70, lending credence to validity evidence.

Among the other C^3 constructs, Troubleshooting was only correlated with the reasoning measures. The relationship between Deductive Reasoning and Troubleshooting was discussed in the Deductive Reasoning section and attributed to the role of deduction in the process of finding faulty nodes in the fault diagrams. The relationship between Inductive Reasoning and Troubleshooting suggests that an induction process was also involved in the fault-finding task. The magnitude of the correlation between Troubleshooting and Inductive Reasoning was greater than the correlation between Troubleshooting and Deductive Reasoning, suggesting that Inductive Reasoning might have a more important role than Deductive Reasoning. Perhaps rather than use deductive processes like the process of elimination, an educated guess approach is more useful for finding the faults in an efficient manner. Additional research could help to clarify the relationship between Troubleshooting and reasoning skills.

The relationship between Troubleshooting and Inductive Reasoning is likely partially driven by the role of cognitive ability in both of these measures as well. Overall Troubleshooting scores had a moderate correlation with cognitive ability and a moderate correlation with an outside measure of deductive reasoning. These were of similar magnitude and stronger than the correlation between Troubleshooting and the C^3 index of Deductive Reasoning.

Looking at the item types, intermittent items were not correlated with cognitive ability or deductive reasoning, whereas the other two item types were correlated with these outside measures. The benefit of general cognitive abilities for completing the fault diagnosis task appeared to be reduced when solving intermittent items. Perhaps this was due to overconfidence on the part of the higher cognitive ability test takers, such that when they thought they identified the location of a fault, they did no further testing. In intermittent-type items, overconfidence would lead test takers to conclude their reasoning about fault location was accurate and complete, but in reality, they may have checked a faulty node only at one point in time when it

appeared to be working. A more cautious approach would be to double-check the conclusions, which, for the intermittent items, would often reveal that faults persisted.

Based on the results of these analyses, there are a few things that could be done to improve the Troubleshooting measure. First, items can be selected with the greatest range of difficulties, so that the test can distinguish between ability levels equally well across the scale from low ability to high ability. A determination must be reached on whether to keep both linear items and network items. Though network items appeared to be the easiest, this was likely due in part to practice effects. In order to more clearly determine if these item types differ, more research should be conducted in which administration of these items is done in a counterbalanced way.

In addition to improving the items included in the measure, the Troubleshooting measure could be improved by adjusting and adding to the metrics that are collected. Currently, the efficiency score takes into consideration the numbers of faults identified relative to the number of resources used to find those faults. There may be ways to improve these calculations; for example, by including the number of individual node checks, or the number of checks of the entire set of nodes, in addition to the number of nodes repaired. Test takers were given information about the costs of each action in terms of resources used, but the optimal approach to troubleshooting might depend on how those costs are assigned and whether costs or errors are more important to minimize. In addition, the Troubleshooting measure might be improved in future iterations of the assessment by including information about the troubleshooting process, rather than just the troubleshooting outcomes. This information might include the sequences of actions taken and whether those sequences are most efficient, yielding information at every stage and not repeating information that was already obtained. This would conform to other troubleshooting metrics in the literature (e.g., Henneman & Rouse, 1984).

Conclusions

A detailed evaluation of the descriptive and correlational results for each of the C^3 constructs identified a number of ways that each measure could potentially be improved; however, initial psychometric evidence for the C^3 battery is promising. Most of the measures demonstrated an appropriate level of variance and had scores with approximately normal distributions. While some distributions were slightly skewed, reflecting measures that were somewhat too easy or hard for participants, the causes of these characteristics were easily identified and can be fixed in subsequent versions of the battery. A few measures showed central tendency bias in their distributions and would benefit from procedures to increase the variance; for example, Critical Thinking and Complex Problem Solving EO would both benefit from encouraging test takers to use the entire scale when making their responses.

Though it was not possible to collect construct validity data for all measures in this initial study, the correlations among the C^3 constructs and subdimensions provided some confirmation of construct validity. Most intercorrelations were significant, but small to medium in magnitude. This is likely due to the underlying influence of a general mental ability factor affecting scores across constructs. As this relationship was expected, this provides some evidence of convergent validity, though future research is needed to show stronger relationships between each of the constructs and outside measures of those constructs. The fact that correlations were small to medium suggests that no two measures were measuring the same construct. Since each of the

C^3 tests were designed to measure a distinct construct, this provides initial evidence of discriminant validity. Nevertheless, additional research should be conducted to more closely examine the divergence of these measures from other similar measures. Although additional research is needed to demonstrate discriminant validity evidence in comparing C^3 constructs with general intelligence, the fact that correlations between the C^3 constructs and general intelligence tests were small to medium suggests divergence from general intelligence. Future research should examine the criterion-related validity of the C^3 Test, and the extent to which C^3 constructs can predict criteria above and beyond general intelligence.

REFERENCES

Barron, L. G., & Rose, M. R. (2017). Multitasking as a predictor of pilot performance: Validity beyond serial single-task assessments. *Military Psychology, 29*(4), 316.

Behrens, P. J. (1996). The Watson-Glaser critical thinking appraisal and academic performance of diploma school students. *Journal of Nursing Education, 35*(1), 34-36.

Benware, C. A., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal, 21*(4)*,* 755-765.

Berger, J., & Karabenick, S. A. (2016). Construct validity of self-reported metacognitive learning strategies. *Educational Assessment, 21*(1)*,* 19-33.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3)*,* 382-398.

Braxton, J. M., Milem, J. F., & Sullivan, A. S. (2000). The influence of active learning on the college student departure process: Toward a revision of Tinto's theory. *The Journal of Higher Education, 71*(5)*,* 569-590.

Burhmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspective on Psychological Science, 6(1),* 3-5.

Butler, H. A. (2012). Halpern critical thinking assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology, 26*(5)*,* 721-729.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Cheung, J. H., Burns, D. K., Sinclair, R. R., Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology, 32(4),* 347-361.

Choi, I., Nisbett, R. E., & Smith, E. E. (1997). Culture, category salience, and inductive reasoning. *Cognition, 65*(1)*,* 15-32.

Coetzee, M. (2014). Measuring student graduateness: Reliability and construct validity of the Graduate Skills and Attributes Scale. *Higher Education Research & Development*, *33*(5), 887-902.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29.

Dagenbach, D., & Carr, T. H. (Eds.). (1994). *Inhibitory processes in attention, memory, and language.* Academic Press.

Danner, D., Hagemann, D., Holt, D. V., Hager, M., Schankin, A., Wüstenberg, S., & Funke, J. (2011). Measuring performance in dynamic decision making: Reliability and validity of the Tailorshop simulation. *Journal of Individual Differences, 32*(4), 225–233.

Davidson, B. W., & Dunham, R. L. (1996). Assessing EFL student progress in critical thinking with the Ennis-Weir Critical Thinking Essay Test. *Japan Association for Language Teaching Journal, 19*(1), 43-57.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*(2)*,* 225.

Díaz-Morales, J. F., & Escribano, C. (2013). Predicting school achievement: The role of inductive reasoning, sleep length and morningness–eveningness. *Personality and Individual Differences, 55*(2)*,* 106-111.

Dörner, D. (1980). On the difficulty people have in dealing with complexity. *Simulation & Games 11*(1)*,* 87–106.

Drummond, S., Gillin, J. C., & Brown, G. G. (2001). Increased cerebral response during a divided attention task following sleep deprivation. *Journal of Sleep Research, 10*(2)*,* 85-92.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests (rev. ed.)*. Educational Testing Services.

Ennis, R. H. (2003). *Critical thinking assessment.* D. Fasko (Ed.), Critical thinking and reasoning (pp. 293–310). Hampton Press.

Eriksen, B. A. & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non-search task. *Perception and Psychophysics, 16*(1), 143–149.

Facione, P. A. (1990). *The California critical thinking skills test – college level. Technical report #1: Experimental validation and content validity.* California Academic Press.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20*(1), 61-84.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving*, *4*(1), 19–42.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906-911.

Fleishman, E. A., Costanza, D. P., & Marshall-Mies, J. (1999). Abilities. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 49–69). American Psychological Association.

Fraser, J., Morrison, A., & Wells, A. (2006). Cognitive processes, reasoning biases and persecutory delusions: A comparative study. *Behavioural and Cognitive Psychotherapy, 34*(4)*,* 421-435.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, *7*(1), 69-89.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing, 11*(2)*,* 133-142.

Gibbs, G., & Coffey, M. (2004). The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active Learning in Higher Education*, *5*(1), 87-100.

Girelli, L., Semenza, C., & Delazer, M., (2004). Inductive reasoning and implicit memory: evidence from intact and impaired memory systems. *Neuropsychologia, 42*(7), 926–938.

Gitomer, D. H. (1988). Individual differences in technical troubleshooting. *Human Performance, 1*(2), 111-131.

Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition, 93*(3), B109-B121.

Gottesman, L., & Chapman, L. J. (1960). Syllogistic reasoning errors in schizophrenia. *Journal of Consulting Psychology, 24*(3)*,* 250-255.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*(6939), 534-537.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: the MicroDYN approach. In: Scheuermann F (ed.), *The Transition to computer-based assessment—lessons learned from large-scale surveys and implications for testing.* Office for Official Publications of the European Communities, Luxembourg.

Greiff, S., Fischer, A., Stadler, M., & Wustenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*(3), 356-382.

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*(5), 579-596.

Halpern, D. F. (2010). Halpern Critical Thinking Assessment. Schuhfried (Vienna Test System). Available at http://www.schuhfried.com/vienna-test-system-vts/alltests-from-a-z/test/hcta-halpern-critical-thinking-assessment-1/

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193-225). Academic Press.

Hayes, B. K., Stephens, R. G., Ngo, J., & Dunn, J. C. (2018). The Dimensionality of Reasoning: Inductive and Deductive Inference can be Explained by a Single Process. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 44*(9), 1333-1351.

Henneman, R. L., & Rouse, W. B. (1984). Measures of human problem solving performance in fault diagnosis tasks. *IEEE Transactions on Systems, Man, and Cybernetics, 14(1), 99-112.*

Holzman, T., Pellegrino, J., & Glaser, R., (1983). Cognitive variables in series completion. *Journal of Educational Psychology, 75*(4)*,* 603–618.

Hsia, L. H., Huang, I., & Hwang, G. J. (2016). A web-based peer-assessment approach to improving junior high school students' performance, self-efficacy and motivation in performing arts courses. *British Journal of Educational Technology*, *47*(4), 618-632.

Hwang, G. J., Yang, L. H., & Wang, S. Y. (2013). A concept map-embedded educational computer game for improving students' learning performance in natural science courses. *Computers & Education*, *69*, 121-130.

Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education, 56*(4), 1023-1031.

Jacobsen, P., Freeman, D., & Salkovskis, P. (2012). Reasoning bias and belief conviction in obsessive-compulsive disorder and delusions: Jumping to conclusions across disorders? *British Journal of Clinical Psychology, 51*(1), 84-99.

Jessop, N., & Adams, G. (2016). Internationalising the psychology curriculum: Preliminary notes on conception and assessment of anticipated benefits. *Psychology Teaching Review, 22*(2), 41-52.

Johnson, S. D., & Satchwell, S. E. (1993). The effect of functional flow diagrams on apprentice aircraft mechanics' technical system understanding. *Performance Improvement Quarterly, 6*(4), 73-91.

Johnson-Laird, P. N. (1995). Mental models, deductive reasoning, and the brain*.* In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 999–1008). MIT Press.

Jonassen, D. H. (2012). Problem typology. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (Vol. 6, pp. 2683–2686). Springer.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology, 132*(1), 47-70.

Kember, D., & Leung, D. Y. P. (2009). Development of a questionnaire for assessing students' perceptions of the teaching and learning environment and its use in quality assurance. *Learning Environments Research, 12*(1), 15-29.

Kim, M. K., Kim, S. M., & Bilir, M. K. (2014). Investigation of the dimensions of Workplace Learning Environments (WLEs): Development of the WLE measure. *Performance Improvement Quarterly, 27*(2), 35-57.

Kotovsky, K., & Simon, H. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4*(3), 399-424.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*(4), 347-368.

Kuhn, D. (1977). Conditional reasoning in children. *Developmental Psychology, 13*(4)*,* 342-353.

Kurland, L. C., Granville, R. A. & MacLaughlin, D. M. (1992). Design, development, and implementation of an intelligent tutoring system for training radar mechanics to troubleshoot. In M. Farr & J. Psotka, (Eds.), *Intelligent Instruction by Computer* (pp. 205-237). Taylor & Francis.

Lawson, T. J., Jordan-Fleming, M. K., & Bodle, J. H. (2015). Measuring psychological critical thinking: An update. *Teaching of Psychology, 42*(3), 248-253.

Lefevre, J., & Bisanz, J., (1986). A cognitive analysis of number series problems: sources of individual differences in performance. *Memory & Cognition, 14*(4), 287–298.

Liu, S. H. (2017). Relationship between the factors influencing online help-seeking and self-regulated learning among Taiwanese preservice teachers. *Computers in Human Behavior, 72,* 38-45.

Maher, A., & von Hippel, C. (2005). Individual differences in employee reactions to open-plan offices. *Journal of Environmental Psychology, 25*(2), 219-229.

McAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality, 28*(6), 604-619.

McDowd, J. M., & Craik, F. I. (1988). Effects of aging and task difficulty on divided attention performance. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 267-280.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247-279.

Morris, N. M. & Rouse, W. B. (1985). Review and evaluation of empirical research in troubleshooting. *Human Factors, 27*(5), 503-530.

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 596-616.

Moutier, S., & Houdé, O. (2003). Judgement under uncertainty and conjunction fallacy inhibition training. *Thinking & Reasoning, 9*(3), 185-201.

Newman, D. R., Johnson, C., Webb, B., & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co-operative learning. *Journal of the American Society for Information Science, 48(6),* 484-495.

O*NET Online (n.d.). Retrieved from http://www.onetonline.org/find/descriptor/browse/Skills/.

O'Hare, D. (1997). Cognitive ability determinants of elite pilot performance. *Human Factors, 39*(4)*,* 540-552.

Phillips, J. B., Chernyshenko, O. S., Stark, S., Drasgow, F., & Phillips, I. V. (2011). Development of scoring procedures for the Performance Based Measurement (PBM) test: Psychometric and criterion validity investigation (No. NAMRU-D-12-10). Naval Medical Research Unit.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press.

Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. In J. E. Davidson & R. J. Sternberg (Eds.), *The Psychology of Problem Solving* (pp. 3-30). Cambridge University Press.

Ribas, V. R., Martins, H. A. D. L., Amorim, G. G., Ribas, R. D. M. G., Almeida, C. Â. V. D., Ribas, V. R., & Castro, R. M. D. (2010). Air traffic control activity increases attention capacity in air traffic controllers. *Dementia & Neuropsychologia, 4*(3), 250-255.

Roscoe, S. N. (1997). Predicting and enhancing flightdeck performance. In R. Telfer and P. Moore (Eds.), *Aviation Training: Pilot, Instructor, and Organization* (pp. 195-502). Ashgate.

Ross, C., & Orr, R. R. (2009). Teaching structured troubleshooting: integrating a standard methodology into an information technology program. *Education Technology Research & Development, 57*(2)*,* 251-265.

Rouse, W. B., Rouse, S. H., & Pellegrino, S. J. (1980). A rule-based model of human problem solving performance in fault diagnosis tasks. IEEE Transactions on Systems, Man, & Cybernetics, *10*(7), 366-376.

Rowold, J., & Kauffeld, S. (2008). Effects of career-related continuous learning on competencies. *Personnel Review, 38*(1), 90-101.

Saltz, E. and Moore, S. V. (1953). *A preliminary investigation of trouble-shooting* (Technical Report 53-2). Air Force Personnel and Training Research Center.

Saunders-Stewart, K. S., Gyles, P. D. T., Shore, B. M., & Bracewell, R. J. (2015). Student outcomes in inquiry: Students' perspectives. *Learning Environments Research, 18*(2), 289-311.

Saupe, J. L. (1954). Troubleshooting electronic equipment: An empirical approach to the identification of certain requirements of a maintenance occupation. [Doctoral Dissertation, University of Illinois].

Schaafstal, A., Schraagen, J. M., & van Berlo, M. (2000). Cognitive task analysis and innovation of training: The case of structured troubleshooting. *Human Factors, 42*(1), 75-86.

Schaie, K. W. (1985). *Manual for the Schaie-Thurstone adult mental abilities test (STAMAT).* Consulting Psychologists Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274.

Schmidt, L., O'Connell, C., Miyake, H., Shah, A. R., Baron, J., Nieboer, G., et al. (2015). *Cyber Practices: What Can the U.S. Air Force Learn from the Commercial Sector?* RAND Corporation.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460-475.

Schuitema, J., Peetsma, T., & van der Veen, I. (2016). Longitudinal relations between perceived autonomy and social support from teachers and students' self-regulated learning and achievement. *Learning and Individual Differences*, *49*, 32-45.

Simon, H., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review, 70*(6), 534–546.

Simpson, R. W. & Nester, M. A. (2007, June 10-13). Taxonomy for Reasoning Questions Using Logic-Based Measurement. [Paper presentation]. IPMAAC Conference on Personnel Assessment 31st Annual Meeting, St. Louis, MO, United States. Available at: http://www.ipacweb.org/conf07/simpson4.pdf

Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological tests*. Oxford University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643-662.

Swezey, R. W., Perez, R., & Allen, J. (1988). Effects of instructional delivery system and training parameter manipulations on electromechanical performance. *Human Factors, 30*(6), 751-762.

Taasoobshirazi, G., & Farley, J. (2013). Construct validation of the physics metacognition inventory. *International Journal of Science Education*, *35*(3), 447-459.

Tams, S., Thatcher, J., Grover, V., & Pak, R. (2015). Selective attention as a protagonist in contemporary workplace stress: implications for the interruption age. *Anxiety, Stress, & Coping, 28*(6), 663-686.

Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *Journal of General Education, 46*(2), 129-164.

Teague, R. C., & Allen, J. A. (1997). The reduction of uncertainty and troubleshooting performance. *Human Factors, 39*(2), 254-267.

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *The Quarterly Journal of Experimental Psychology*, *37A*(4), 571-590.

Tipper, S. P. (1992). Selection for action: The role of inhibitory mechanisms. *Current Directions in Psychological Science*, *1*(3), 105-109.

Trippe, D. M., Canali, K. G., Wind, A. P., & Koch, A. J. (Eds.). (2019). *Expanded Development of Cyber Selection Tests* (Technical Report 1376). U.S. Army Research Institute for the Behavioral and Social Sciences.

Trippe, D. M., José, I. J., Reeder, M. C., Brown, D., Heffner, T. S., Wind, A. P., Thomas, K. I., & Canali, K. (2017). *Validation of the Information/Communications Technology Literacy Test* (Technical Report 1360). U.S. Army Research Institute for the Behavioral and Social Sciences.

Trippe, D. M., Moriarty, K. O., Russell, T. L., Carretta, T. R., & Beatty, A. S. (2014). Development of a cyber/information technology knowledge test for military enlisted technical training qualification. *Military Psychology, 26*(3), 182.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy and probability judgment. *Psychological Review, 90*(4)*,* 293-315.

Unsworth, N., Fukuda, K., Awh, E., Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71,* 1-26.

U.S. Department of Defense. (February 2013). *Cyberspace operations* (Joint Publication 3-12). Joint Chiefs of Staff.

Wagner, T. A., & Harvey, R. J. (2006). Development of a new critical thinking test using item response theory. *Psychological Assessment, 18*(1), 100-105.

Williams, H. P., Albert, A. O., & Blower, D. J. (2000). *Selection of officers for U.S. naval aviation training.* Naval Aerospace Medical Research Laboratory.

Willis, S. L., & Schaie, K. W. (1986). Training the elderly on the ability factors of spatial orientation and inductive reasoning. *Psychology and Aging, 1*(3), 239-247.

Wind, A. (2018). *C^3 Capabilities Selection*. [Unpublished manuscript]. U.S. Army Research Institute for the Behavioral and Social Sciences.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence, 40*(1), 1-14.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development, 78*(4), 16-33.

APPENDIX A

TEST TAKER COMMENTS TO OPEN-ENDED QUESTIONS

Table A1

Test Taker Challenges during Training

Comment	n	Endorsements
Amount of information/it was too long	91	44
Determining which pieces of information are important	91	17
Technical terms/jargon	91	15
Interruptions	91	10
Logic questions	91	10
Duties/roles/functions for different positions/teams	91	8
Amount of reading	91	7
Amount of unnecessary information	91	6
Staying engaged/information was boring	91	6
Limited amount of time	91	5
Mechanics of the system/parts of the system	91	5
Differences between squads/which drone does what	91	5
Organization of the material (e.g., labeling of sections, jumping back and forth with content, information flow)	91	4
Lack of visual data to compliment text/too much long text	91	4
Complexity of the material	91	4
Level of detail of the information	91	3
Acronyms	91	3
No hands-on work/practical application	91	3
Lack of examples (e.g., no examples of documented incidents)	91	2
Emergency response section – when to use different responses	91	2
Keeping organization levels straight	91	1
Remembering how to access reports	91	1
Small text	91	1
Making decisions	91	1
Not being able to go back	91	1
Monitoring	91	1
Pods	91	1
Questions at the end	91	1
No breaks	91	1
Remembering data	91	1
The information about the gas and magnets – there was a lot of information that seemed to be added in that wasn't so important	91	1
Knowing at which stage I am supposed to intervene	91	1

Table A2

Test Taker Suggested Revisions – Removing Content or Sections of Training

Comment	n	Endorsements
Generally shorten the training/include less information	91	21
Include only critical information needed to perform the job of the CRO	91	3
Remove the articles/news stories	91	3
Spend less time on the history of the tubes	91	3
Remove redundant information	91	2
Remove employee conversations	91	2
Remove logic questions	91	2
Reduce the material on the Crisis Response Team	91	1
Reduce the amount of jargon	91	1
Remove pop-ups/interruptions	91	1
Reduce the contents and avoid numerical factors	91	1

Table A3

Test Taker Suggested Revisions – Changes to the Training Section

Comment	n	Endorsements
Allow people to go back and review past material/allow for revisitation with a Wiki-style format or persistent links	91	5
Give more emphasis to the necessary information and put the extraneous information in its own section so an employee can access it as they have free time or are interested/change the order of the information so that there would be more relevant information shown first	91	5
Change the interruptions	91	2
Add more examples	91	2
Add more explanation for the logic questions	91	2
Add more details about the security protocol	91	1
Structure the training to take place over several days	91	1

Table A4

Test Taker Suggested Revisions – Additions to the Training Section

Comment	n	Endorsements
Add fact check quiz at the end of (or within) each section to test knowledge	91	17
Add live examples/make it interactive/more hands on	91	11
Add more diagrams/charts/visualizations/pictures	91	5
Add videos	91	5
Include breaks/slow it down/more time	91	5
Add summaries at the end of the sections to highlight important information	91	4
Add a place for note-taking/highlighting	91	2
Add real examples about actual incidents that occurred/case study	91	2
Add a glossary for definitions where information can be searched	91	1
Add FAQ page for questions or common misunderstandings where trainees can go to clear up common issues	91	1
Add more information about the tasks involved in the job	91	1

Table A5

Test Taker Suggested Revisions – No Change

Comment	n	Endorsements
Nothing/not much/no changes	91	6

APPENDIX B

ACTIVE LEARNING: DISTRACTOR ANALYSES FOR LEARNING EFFECTIVENESS ITEMS

Table B1

Learning Phase Learning Effectiveness Items

correct	key	n	rspP	pBis	discrim	lower	mid	upper
ALLE1.12.x								
	1	25	0.37	-0.21	-0.05	0.46	0.24	0.41
	2	23	0.34	-0.22	0.07	0.25	0.48	0.32
*	3	15	0.22	0.10	0.15	0.12	0.29	0.27
	4	3	0.04	-0.29	-0.12	0.12	0.00	0.00
	5	1	0.01	-0.11	-0.04	0.04	0.00	0.00
ALLE1.4.x								
*	1	56	0.84	0.27	0.33	0.62	0.95	0.95
	2	6	0.09	-0.25	-0.12	0.17	0.05	0.05
	3	1	0.01	-0.24	-0.04	0.04	0.00	0.00
	4	1	0.01	-0.20	-0.04	0.04	0.00	0.00
	5	3	0.04	-0.34	-0.12	0.12	0.00	0.00
ALLE1.14.x								
	1	1	0.01	0.11	0.05	0.00	0.00	0.05
	2	6	0.09	-0.27	-0.17	0.17	0.10	0.00
	3	12	0.18	-0.22	-0.03	0.21	0.14	0.18
*	4	26	0.39	0.20	0.38	0.21	0.38	0.59
	5	22	0.33	-0.38	-0.23	0.42	0.38	0.18
ALLE1.25.x								
*	1	27	0.40	0.39	0.65	0.12	0.33	0.77
	2	9	0.13	-0.24	-0.12	0.17	0.19	0.05
	3	8	0.12	-0.32	-0.17	0.17	0.19	0.00
	4	16	0.24	-0.35	-0.24	0.38	0.19	0.14
	5	7	0.10	-0.29	-0.12	0.17	0.10	0.05
ALLE1.24.x								
	1	8	0.12	-0.23	-0.12	0.17	0.14	0.05
*	2	31	0.46	0.27	0.48	0.25	0.43	0.73
	3	27	0.40	-0.46	-0.31	0.54	0.43	0.23
	4	1	0.01	-0.20	-0.04	0.04	0.00	0.00
	5	0	0.00	NA	0.00	0.00	0.00	0.00
ALLE1.32.x								
*	1	41	0.61	0.29	0.54	0.42	0.48	0.95
	2	15	0.22	-0.53	-0.42	0.42	0.24	0.00

correct	key	n	rspP	pBis	discrim	lower	mid	upper
	3	2	0.03	-0.22	-0.04	0.04	0.05	0.00
	4	4	0.06	-0.11	-0.04	0.08	0.05	0.05
	5	5	0.07	-0.11	-0.04	0.04	0.19	0.00
ALLE1.26.x								
	1	16	0.24	-0.30	-0.11	0.25	0.33	0.14
	2	10	0.15	-0.40	-0.20	0.29	0.05	0.09
*	3	31	0.46	0.36	0.48	0.25	0.43	0.73
	4	4	0.06	-0.35	-0.12	0.12	0.05	0.00
	5	6	0.09	-0.16	-0.04	0.08	0.14	0.05
ALLE1.21.x								
	1	10	0.15	-0.23	-0.08	0.12	0.29	0.05
*	2	31	0.46	0.23	0.35	0.38	0.29	0.73
	3	4	0.06	-0.09	0.00	0.04	0.10	0.05
	4	13	0.19	-0.30	-0.16	0.25	0.24	0.09
	5	9	0.13	-0.37	-0.12	0.21	0.10	0.09
ALLE1.29.x								
*	1	43	0.64	0.21	0.45	0.42	0.67	0.86
	2	19	0.28	-0.46	-0.41	0.46	0.33	0.05
	3	1	0.01	-0.24	-0.04	0.04	0.00	0.00
	4	3	0.04	0.01	0.05	0.04	0.00	0.09
	5	1	0.01	-0.24	-0.04	0.04	0.00	0.00
ALLE1.27.x								
	1	13	0.19	-0.28	-0.16	0.25	0.24	0.09
*	2	41	0.61	0.20	0.41	0.46	0.52	0.86
	3	2	0.03	-0.22	-0.08	0.08	0.00	0.00
	4	2	0.03	-0.03	0.00	0.00	0.10	0.00
	5	9	0.13	-0.39	-0.16	0.21	0.14	0.05
ALLE1.16.x								
	1	23	0.34	-0.54	-0.41	0.54	0.33	0.14
	2	7	0.10	-0.42	-0.29	0.29	0.00	0.00
	3	2	0.03	-0.13	-0.04	0.04	0.05	0.00
*	4	34	0.51	0.42	0.69	0.12	0.62	0.82
	5	1	0.01	0.20	0.05	0.00	0.00	0.05
ALLE1.15.x								
*	1	30	0.45	0.24	0.47	0.21	0.48	0.68
	2	21	0.31	-0.34	-0.19	0.38	0.38	0.18
	3	7	0.10	-0.16	-0.03	0.12	0.10	0.09
	4	7	0.10	-0.39	-0.20	0.25	0.00	0.05
	5	2	0.03	-0.09	-0.04	0.04	0.05	0.00
ALLE1.22.x								
	1	8	0.12	-0.11	0.01	0.08	0.19	0.09
	2	1	0.01	-0.20	-0.04	0.04	0.00	0.00

correct	key	n	rspP	pBis	discrim	lower	mid	upper
*	3	28	0.42	0.19	0.34	0.29	0.33	0.64
	4	17	0.25	-0.31	-0.16	0.29	0.33	0.14
	5	13	0.19	-0.34	-0.16	0.29	0.14	0.14
ALLE1.1.x								
	1	1	0.01	-0.07	0.00	0.00	0.05	0.00
*	2	59	0.88	0.21	0.29	0.71	0.95	1.00
	3	3	0.04	-0.29	-0.12	0.12	0.00	0.00
	4	4	0.06	-0.31	-0.17	0.17	0.00	0.00
	5	0	0.00	NA	0.00	0.00	0.00	0.00
ALLE1.23.x								
	1	27	0.40	-0.27	-0.05	0.42	0.43	0.36
	2	12	0.18	-0.44	-0.29	0.33	0.14	0.05
	3	4	0.06	-0.05	0.01	0.08	0.00	0.09
*	4	14	0.21	0.24	0.28	0.04	0.29	0.32
	5	10	0.15	-0.08	0.06	0.12	0.14	0.18

Table B2

Application Phase Learning Effectiveness Items

correct	key	n	rspP	pBis	discrim	lower	mid	upper
ALLE2.1.x								
*	1	28	0.42	0.31	0.47	0.08	0.65	0.55
	2	8	0.12	-0.53	-0.29	0.29	0.04	0.00
	3	21	0.31	-0.39	-0.26	0.46	0.26	0.20
	4	3	0.04	-0.02	0.06	0.04	0.00	0.10
	5	7	0.10	-0.10	0.02	0.12	0.04	0.15
ALLE2.4.x								
	1	3	0.04	-0.18	-0.08	0.08	0.04	0.00
*	2	29	0.43	0.27	0.49	0.21	0.43	0.70
	3	18	0.27	-0.40	-0.28	0.33	0.39	0.05
	4	17	0.25	-0.34	-0.12	0.38	0.13	0.25
	5	0	0.00	NA	0.00	0.00	0.00	0.00
ALLE2.6.x								
*	1	38	0.57	0.20	0.47	0.33	0.61	0.80
	2	13	0.19	-0.29	-0.15	0.25	0.22	0.10
	3	6	0.09	-0.29	-0.17	0.17	0.09	0.00
	4	7	0.10	-0.30	-0.12	0.17	0.09	0.05
	5	3	0.04	-0.15	-0.03	0.08	0.00	0.05
ALLE2.8.x								
	1	10	0.15	-0.43	-0.25	0.25	0.17	0.00
	2	3	0.04	-0.18	-0.08	0.08	0.04	0.00

correct	key	n	rspP	pBis	discrim	lower	mid	upper
	3	16	0.24	-0.47	-0.36	0.46	0.13	0.10
	4	5	0.07	0.04	0.10	0.00	0.13	0.10
*	5	33	0.49	0.32	0.59	0.21	0.52	0.80
ALLE2.9.x								
*	1	33	0.49	0.29	0.54	0.21	0.57	0.75
	2	8	0.12	-0.21	-0.07	0.17	0.09	0.10
	3	3	0.04	-0.15	-0.04	0.04	0.09	0.00
	4	16	0.24	-0.43	-0.36	0.46	0.13	0.10
	5	7	0.10	-0.30	-0.08	0.12	0.13	0.05
ALLE2.10.x								
	1	3	0.04	-0.31	-0.08	0.08	0.04	0.00
	2	6	0.09	-0.34	-0.21	0.21	0.04	0.00
*	3	49	0.73	0.45	0.62	0.38	0.87	1.00
	4	3	0.04	-0.26	-0.12	0.12	0.00	0.00
	5	6	0.09	-0.43	-0.21	0.21	0.04	0.00
ALLE2.12.x								
	1	8	0.12	-0.21	-0.08	0.12	0.17	0.05
*	2	45	0.67	0.10	0.31	0.54	0.65	0.85
	3	2	0.03	-0.25	-0.08	0.08	0.00	0.00
	4	4	0.06	-0.17	-0.03	0.08	0.04	0.05
	5	8	0.12	-0.26	-0.12	0.17	0.13	0.05
ALLE2.13.x								
	1	5	0.07	-0.13	-0.03	0.08	0.09	0.05
	2	24	0.36	-0.22	-0.08	0.33	0.48	0.25
	3	7	0.10	-0.23	-0.08	0.12	0.13	0.05
*	4	5	0.07	-0.05	0.02	0.08	0.04	0.10
	5	26	0.39	-0.08	0.18	0.38	0.26	0.55
ALLE2.14.x								
	1	0	0.00	NA	0.00	0.00	0.00	0.00
	2	44	0.66	-0.44	-0.30	0.75	0.74	0.45
*	3	20	0.30	0.19	0.38	0.17	0.22	0.55
	4	2	0.03	-0.16	-0.04	0.04	0.04	0.00
	5	1	0.01	-0.13	-0.04	0.04	0.00	0.00
ALLE2.15.x								
	1	7	0.10	-0.18	-0.03	0.08	0.17	0.05
	2	6	0.09	-0.14	-0.08	0.12	0.09	0.05
	3	9	0.13	-0.54	-0.29	0.29	0.09	0.00
*	4	33	0.49	0.23	0.46	0.29	0.48	0.75
	5	12	0.18	-0.20	-0.06	0.21	0.17	0.15
ALLE2.16.x								
*	1	29	0.43	0.11	0.36	0.29	0.39	0.65
	2	27	0.40	-0.37	-0.26	0.46	0.52	0.20

correct	key	n	rspP	pBis	discrim	lower	mid	upper
	3	3	0.04	-0.02	0.01	0.04	0.04	0.05
	4	1	0.01	-0.04	0.00	0.00	0.04	0.00
	5	7	0.10	-0.30	-0.11	0.21	0.00	0.10
ALLE2.17.x								
	1	25	0.37	-0.25	-0.07	0.42	0.35	0.35
	2	17	0.25	-0.24	-0.14	0.29	0.30	0.15
*	3	18	0.27	0.03	0.24	0.21	0.17	0.45
	4	7	0.10	-0.16	-0.03	0.08	0.17	0.05
	5	0	0.00	NA	0.00	0.00	0.00	0.00
ALLE2.18.x								
	1	11	0.16	-0.23	-0.06	0.21	0.13	0.15
	2	7	0.10	-0.47	-0.25	0.25	0.04	0.00
	3	21	0.31	-0.25	-0.08	0.33	0.35	0.25
	4	2	0.03	-0.19	-0.04	0.04	0.04	0.00
*	5	26	0.39	0.26	0.43	0.17	0.43	0.60
ALLE2.19.x								
	1	2	0.03	-0.06	0.01	0.04	0.00	0.05
	2	3	0.04	-0.34	-0.12	0.12	0.00	0.00
	3	2	0.03	-0.16	-0.04	0.04	0.04	0.00
	4	3	0.04	-0.31	-0.12	0.12	0.00	0.00
*	5	57	0.85	0.21	0.28	0.67	0.96	0.95
ALLE2.21.x								
	1	19	0.28	-0.31	-0.18	0.38	0.26	0.20
	2	1	0.01	0.00	0.00	0.00	0.04	0.00
	3	16	0.24	-0.41	-0.28	0.38	0.22	0.10
*	4	20	0.30	0.30	0.52	0.08	0.26	0.60
	5	11	0.16	-0.21	-0.07	0.17	0.22	0.10

APPENDIX C

CRITICAL THINKING STATEMENT STATISTICS

Table C1

Critical Thinking: Analysis Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
CT.13	67	3.24	0.89	-0.05	0.47
CT.14	67	4.13	0.74	0.22	0.42
CT.15	67	3.56	0.97	0.22	0.42
CT.16	67	3.69	1.02	0.23	0.41
CT.17	67	3.49	1.11	0.26	0.41
CT.18	67	4.00	0.82	-0.06	0.47
CT.19	67	3.57	0.97	-0.22	0.49
CT.28	67	3.13	1.23	0.17	0.44
CT.31	67	3.47	1.39	0.19	0.43
CT.32	67	4.46	1.09	0.34	0.40
CT.34	67	4.60	0.35	0.39	0.38
CT.35	67	2.83	1.09	0.09	0.44
CT.36	67	3.44	1.17	0.14	0.44
CT.37	67	4.13	1.00	0.11	0.44
CT.39	67	3.64	1.31	0.02	0.46
CT.40	67	4.09	1.19	-0.16	0.48
CT.41	67	2.71	1.05	0.12	0.43
CT.42	67	3.40	1.16	-0.15	0.48
CT.70	67	4.05	0.76	-0.03	0.46
CT.71	67	4.05	0.58	0.02	0.45
CT.72	67	3.77	0.83	0.20	0.43
CT.73	67	4.43	0.26	0.31	0.39
CT.74	67	4.19	0.86	0.29	0.39
CT.75	67	4.01	0.98	0.01	0.45

Table C2

Critical Thinking: Explanation Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
CT.49	67	4.13	0.52	0.20	0.56
CT.50	67	4.32	0.40	0.05	0.62
CT.51	67	3.98	0.75	0.59	0.34
CT.52	67	3.70	1.15	0.56	0.35
CT.53	67	4.18	0.66	0.37	0.46

Table C3

Critical Thinking: Evaluation Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
CT.20	67	4.38	0.65	-0.04	0.12
CT.21	67	3.91	0.72	0.07	0.12
CT.22	67	2.27	0.96	0.10	0.11
CT.23	67	3.59	0.95	-0.09	0.20
CT.24	67	3.82	0.90	0.21	0.07
CT.25	67	3.43	0.77	-0.03	0.15
CT.26	67	4.40	0.50	-0.03	0.12
CT.27	67	2.92	0.89	0.13	0.10
CT.44	67	3.75	1.22	0.08	0.12
CT.45	67	4.38	0.41	0.27	0.04
CT.46	67	3.72	0.98	0.06	0.12
CT.47	67	4.01	0.66	-0.02	0.15
CT.48	67	3.00	0.73	0.08	0.13
CT.76	67	3.35	0.76	0.07	0.12
CT.77	67	4.30	0.29	-0.26	0.23
CT.78	67	4.36	0.77	0.02	0.16
CT.79	67	4.18	0.54	-0.08	0.14
CT.80	67	3.87	0.80	-0.01	0.13
CT.81	67	4.43	0.49	0.18	0.11

Table C4

Critical Thinking: Interpretation Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Alpha if Dropped
CT.1	67	3.69	0.74	0.17	0.34
CT.2	67	4.23	0.71	0.17	0.31
CT.3	67	4.09	0.73	0.09	0.36
CT.5	67	4.17	0.73	0.12	0.35
CT.6	67	3.13	0.98	-0.08	0.41
CT.8	67	4.04	0.81	0.31	0.30
CT.9	67	3.72	0.71	0.07	0.37
CT.54	67	3.40	1.00	-0.03	0.39
CT.55	67	4.84	0.48	0.05	0.35
CT.56	67	4.24	0.54	-0.22	0.45
CT.57	67	3.67	0.74	0.26	0.34
CT.58	67	3.32	1.10	-0.03	0.39
CT.59	67	3.75	0.64	0.14	0.35
CT.60	67	3.97	0.59	0.16	0.33
CT.61	67	3.43	0.96	0.33	0.32
CT.62	67	2.31	1.22	0.25	0.33
CT.64	67	4.28	0.85	0.28	0.28
CT.65	67	4.26	0.50	0.09	0.35
CT.66	67	4.54	0.78	0.12	0.35
CT.67	67	4.32	0.32	-0.12	0.41

APPENDIX D

DEDUCTIVE REASONING ITEM STATISTICS

Table D1

Deductive Reasoning Item B Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.B1	67	0.81	0.40	0.24	0.18	0.63
DR.B2	67	0.64	0.48	0.31	0.68	0.63
DR.B3	67	0.70	0.46	0.27	0.50	0.63
DR.B4	67	0.51	0.50	0.17	0.68	0.66
DR.B5	67	0.90	0.31	0.53	0.27	0.56
DR.B6	67	0.93	0.26	0.29	0.18	0.62
DR.B7	67	0.91	0.29	0.33	0.23	0.61
DR.B8	67	0.36	0.48	-0.10	0.36	0.70
DR.B10	67	0.96	0.21	0.21	0.05	0.62
DR.B11	67	0.90	0.31	0.68	0.32	0.52

Table D2

Deductive Reasoning Item F Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.F2	67	0.82	0.39	0.17	0.55	0.16
DR.F3	67	0.94	0.24	0.22	0.18	0.00
DR.F4	67	0.93	0.26	-0.04	0.23	0.35
DR.F7	67	0.99	0.12	0.14	0.05	0.13
DR.F8	67	0.97	0.17	0.05	0.09	0.29

Table D3

Deductive Reasoning Item K Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.K1	67	0.88	0.33	0.65	0.36	0.45
DR.K2	67	0.73	0.45	0.52	0.73	0.52
DR.K3	67	0.79	0.41	0.40	0.64	0.58
DR.K4	67	0.90	0.31	0.70	0.32	0.43
DR.K5	67	0.99	0.12	0.25	0.05	0.60
DR.K8	67	0.99	0.12	-0.05	0.05	0.69
DR.K9	67	0.19	0.40	0.02	0.23	0.66

Table D4

Deductive Reasoning Item M Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.M1	67	0.97	0.17	-0.03	0.05	0.39
DR.M2	67	0.70	0.46	0.24	0.64	0.25
DR.M4	67	0.82	0.39	0.51	0.55	-0.09
DR.M7	67	0.88	0.33	0.44	0.32	0.06
DR.M8	67	0.97	0.17	-0.12	0.09	0.46

Table D5

Deductive Reasoning Item N Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.N1	67	0.75	0.44	0.61	0.59	0.69
DR.N2	67	0.72	0.45	0.71	0.73	0.67
DR.N3	67	0.67	0.47	0.62	0.77	0.69
DR.N4	67	0.64	0.48	-0.09	0.23	0.79
DR.N5	67	0.99	0.12	0.25	0.05	0.74
DR.N6	67	0.99	0.12	0.31	0.05	0.74
DR.N7	67	0.75	0.44	0.69	0.68	0.68
DR.N8	67	0.75	0.44	0.59	0.59	0.69
DR.N9	67	0.70	0.46	0.21	0.41	0.75
DR.N10	67	0.48	0.50	0.25	0.64	0.74

Table D6

Deductive Reasoning Item O Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.O1	67	0.97	0.17	-0.07	0.05	0.17
DR.O2	67	0.36	0.48	0.17	0.77	0.07
DR.O3	67	0.94	0.24	0.03	0.09	0.14
DR.O4	67	0.96	0.21	0.27	0.14	-0.13
DR.O7	67	0.48	0.50	0.13	0.77	0.15
DR.O8	67	0.93	0.26	0.10	0.18	0.15
DR.O9	67	0.43	0.50	-0.31	0.09	0.31

Table D7

Deductive Reasoning Item P Item Statistics

	n	Mean	SD	Corrected Item-Total Corr	Item Discrimination	Alpha if Dropped
DR.P1	67	0.79	0.41	0.22	0.50	0.22
DR.P2	67	0.93	0.26	0.27	0.23	0.21
DR.P3	67	0.94	0.24	0.26	0.18	0.04
DR.P4	67	0.97	0.17	0.05	0.09	0.33
DR.P7	67	0.97	0.17	0.05	0.09	0.34
DR.P8	67	0.61	0.49	-0.14	0.64	0.48

APPENDIX E

TROUBLESHOOTING INTERCORRELATIONS

Table E1

Troubleshooting Intercorrelations

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. TS	75.46	16.46							
2. TS-lin	86.90	23.02	.76** [.63, .84]						
3. TS-net	96.19	13.20	.64** [.48, .77]	.58** [.40, .72]					
4. TS-int	43.28	30.58	.77** [.65, .85]	.22 [-.02, .44]	.17 [-.07, .40]				
5. TS EF	0.35	0.12	.73** [.59, .82]	.56** [.37, .71]	.57** [.38, .71]	.51** [.31, .67]			
6. TS EFL	0.39	0.17	.58** [.39, .72]	.62** [.44, .75]	.46** [.25, .63]	.27* [.03, .48]	.88** [.81, .93]		
7. TS EFN	0.50	0.15	.52** [.32, .68]	.44** [.22, .61]	.60** [.43, .74]	.25* [.01, .47]	.84** [.76, .90]	.66** [.50, .78]	
8. TS EFI	0.17	0.12	.67** [.52, .79]	.24* [.00, .46]	.27* [.03, .48]	.79** [.68, .87]	.65** [.49, .77]	.37** [.14, .56]	.31* [.08, .51]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. TS = Troubleshooting % faults found, TS-lin = % faults found in linear-type, TS-net = % faults found in network-type, TS-int = % faults found in intermittent-type items, TS EF = Troubleshooting efficiency, TS EFL = Troubleshooting efficiency on linear-type items, TS EFN = Troubleshooting efficiency on network-type items, TS EFI = Troubleshooting efficiency on intermittent-type items.
*$p < .05$. **$p < .01$.