

Evaluating Cybersecurity Education Interventions: Three Case Studies

Jelena Mirkovic | University of Southern California

Melissa Dark | Purdue University

Wenliang Du | Syracuse University

Giovanni Vigna | University of California at Santa Barbara

Tamara Denning | University of Utah

The March/April 2015 installment of this department outlined a five-step process for designing an evaluation for an education intervention:¹

- Determine the purpose of the evaluation.
- Frame the evaluation.
- Determine the evaluation questions.
- Determine the information needed to answer the evaluation questions.
- Establish a systematic method for collecting the information, including timing, target population, and instruments.

We applied this five-step approach in designing an evaluation for three education interventions. The goals of this exercise were to show how to design an evaluation for a real intervention from beginning to end; to highlight the common intervention goals and propose suitable evaluation instruments; and to discuss the expected investment

of time and effort in preparing and performing the education evaluations, which can be significant.

Education Interventions

We selected three successful and widely used, yet very different, education interventions.

The University of California at Santa Barbara's (UCSB) International Capture the Flag (iCTF) exercise is a distributed, wide-area security exercise in which teams compete against each other to perform security-related tasks. iCTF is the world's largest and longest-running educational hacking competition that integrates both attack and defense aspects in a live setting. Organized by coauthor Vigna, the competition is held once a year in different locations, and participants come from all over the world (<https://ictf.cs.ucsb.edu>). For this intervention, we consider preparation for the competition as part of the intervention.

SEED labs are a collection of more than 30 hands-on exercises that cover a variety of cybersecurity topics (www.cis.syr.edu/~wedu/seed). These exercises are distributed as prebuilt virtual machine images that participants download and run on their computers, making the exercises highly portable and easily adopted. SEED labs are developed under the leadership of coauthor Du, and have been adopted by hundreds of educators worldwide.

Control-Alt-Hack is a tabletop card game about ethical hacking, designed by coauthor Denning, Tadayoshi Kohno of the University of Washington Computer Security and Privacy Research Lab, and Adam Shostack, an honorary member of the lab. The game introduces many cybersecurity concepts in a fun setting and is widely adopted by educators (www.controlalthack.com).

Determine the Purpose of the Evaluation

We designed summative evaluations for all three interventions that ask whether the intervention achieved the education goals. Summative evaluation can also determine why an intervention succeeded or failed to achieve its desired goals, but this is beyond the scope of this article.

Frame the Evaluation

For each evaluation, we attempted to identify its antecedents (the state of education that the intervention aims to change), transactions (activities that change the state), and desired outcomes (educational goals).² We also identified the underlying assumptions, beliefs,

Table 1. Antecedents, transactions, beliefs, and outcomes for each intervention.

Intervention	Antecedents	Transactions
International Capture the Flag (iCTF) exercise	Students have only superficial knowledge of cybersecurity.	Force participants to work in teams, plan ahead and prepare before the competition, strategize and think like an adversary, and gain practical skills.
	Students lack practical skills in cybersecurity.	Popularize iCTF broadly.
	Students have low interest in cybersecurity.	
	Students engage in only passive learning.	
SEED labs	Students and workforce members lack practical skills.	Allow students to gain practical skills needed for their career and increase student and knowledge retention.
	Teachers who lack practical skills can't teach these skills to students.	Produce materials for teachers that let them assign practical exercises to students even if they aren't experts in a given area of cybersecurity.
	Schools with underrepresented populations can't teach practical skills due to lack of resources.	Exercises are portable and easily adoptable (not tied to infrastructure).
	There are low student interest and retention rates in the cybersecurity field.	Advertise and disseminate materials broadly.
	There is low student retention of cybersecurity knowledge.	
Control-Alt-Hack	People understand little about the importance of cybersecurity and their risk of being attacked.	Introduce diverse fictional characters that players can identify with and present diverse threats in diverse environments.
	People don't understand the diversity of cybersecurity threats.	Advertise and disseminate materials broadly.
	Students have low interest in cybersecurity.	
	Underrepresented populations have even lower interest in cybersecurity.	
	Cybersecurity is perceived as a complex field for a narrow technical population.	

constraints, and theories that relate to each intervention. Table 1 summarizes this data.

Each intervention has numerous educational goals, and some overlap. All three interventions aim to increase student interest in cybersecurity, and iCTF and SEED labs both aim to teach practical

skills. This overlap suggests that evaluation approaches for these common outcomes can be reused by many cybersecurity educators. Other outcomes were specific to each intervention. iCTF emphasizes defensive skills, how to think like an adversary, and teamwork—all of which are needed to prepare

students for a career in cybersecurity. SEED labs aim to help underrepresented teachers and institutions and to increase the retention of students and knowledge in cybersecurity—goals that look at improving the current quality of classroom education. Control-Alt-Hack's objectives are

Underlying assumptions, beliefs, constraints, and theories	Desired outcomes
Competition-based learning generates interest among a variety of students because it appeals to students' desire to compete.	Motivate students to go beyond the call of duty.
Competitions are a form of active learning, which is effective for developing deeper knowledge and practical skills because it requires that students engage in problem-solving activities that promote analysis, synthesis, and evaluation of content.	Improve students' practical skills.
Cooperative learning, problem-based learning, and the use of case methods and simulations are often coupled with active learning and have been shown to improve students' ability to work in teams.	Increase the number of students interested in cybersecurity.
	Teach students how to apply the adversarial mindset under substantial pressure.
Practical exercises increase knowledge retention.	Improve students' practical skills.
Exercises not tied to infrastructure will be more readily adopted and infused into the curriculum of teachers who lack practical skills, especially at schools with underrepresented populations.	Enable teachers lacking practical skills to expose students to practical work in security.
Practical learning exercises will increase student interest and help retain them in cybersecurity programs of study.	Improve retention of material taught in class by reinforcing it in exercises.
	Improve retention of students in the cybersecurity field. Increase students' interest in cybersecurity.
	Enable schools to provide students with practical security skills without spending too much money.
Diverse fictional characters help a broader cross-section of people identify with the field of cybersecurity, showing that the field and its professions welcome people with diverse backgrounds.	Increase understanding of the importance of cybersecurity and the potential risks posed by inadequate security safeguards.
When many people play a computer security game in different settings, it helps increase their understanding of the importance of and interest in cybersecurity. A game as a nonstandard awareness tool might reach different audiences than organized lectures and encourage casual and voluntary play.	Convey the breadth of technologies for which cybersecurity is relevant, including conventional computing platforms and emerging platforms like pervasive technologies and cyber-physical systems.
	Improve understanding of the diversity of potential threats that security designers must consider and the creativity of attackers.
	Increase interest and enthusiasm in cybersecurity.
	Show that the information technology community and its professions welcome people with diverse backgrounds.

to improve the general population's understanding of cybersecurity and to promote diversity in the field. These goals are specific to raising awareness rather than improving skills or acquiring knowledge.

We can group these interventions' outcomes into the following broad categories:

- *skills*—acquiring or improving a skill or set of skills,
- *interest and awareness*—raising awareness of and interest in cybersecurity,
- *learning*—acquiring or retaining knowledge, and
- *impact*—helping underrepresented populations.

Determine the Evaluation Questions

We developed evaluation questions that ask whether the desired outcomes were met. Table 2 summarizes the questions for all three interventions by category. We believe that many other cybersecurity education interventions will

Table 2. Evaluation questions by category.

Category	Questions
Skills	After the intervention, are participants better able to devise strategic, in-depth offensive and defensive tactics?
	Do participants have more practical skills in cybersecurity?
	Are participants better able to use the adversarial mindset under pressure?
	Does the intervention teach participants to plan and work in teams?
Interest and awareness	Are participants more interested in cybersecurity after the intervention?
	Does the intervention increase student retention in the cybersecurity field?
Learning	Do participants better retain cybersecurity knowledge?
	Do participants have a better understanding of the importance of cybersecurity and the risks posed by attacks?
	Do participants have a better understanding of the breadth of technologies that can be affected by cyberattacks?
	Do participants have a better understanding of the diversity of cyberthreats?
Impact	Can institutions that serve underrepresented populations easily adopt the intervention?
	Can teachers who lack practical skills easily adopt the intervention?
	Do participants see themselves studying or practicing cybersecurity after engaging in the intervention?
	Were diverse populations impacted by the intervention?

share desired outcomes with these three interventions, so we focus on designing evaluation instruments for each question, independent of the intervention.

Determine the Information Needed to Answer the Evaluation Questions

This step includes identifying and defining the quality or attribute to be measured, determining a set of operations by which the attribute might be perceived, and establishing a set of procedures or definitions for translating observations into quantitative statements of degree or amount. Here, we discuss ways to evaluate whether the desired outcomes were met using the categories shown in Table 2. We considered a plethora of possible evaluation instruments that could be used to measure given learning goals, regardless of the complexity or cost needed to implement each instrument. In practice, researchers must choose appropriate instruments based on their suitability to the evaluation task at hand as well as the time and effort needed to implement them.

Skills

To determine whether the intervention helps participants acquire or improve a skill, evaluators could apply the following evaluation instruments.

Quiz on hypothetical scenarios.

Describe hypothetical scenarios to participants before and after the intervention. Ask them to reply to questions that reveal their mastery of targeted skills. Compare these measures to see if the mastery of skills improved after the intervention. For example, ask students to provide essay answers to a hypothetical security scenario (in either the defensive or offensive role) and have someone evaluate the depth and quality of their solutions before and after participating in iCTF. This would measure whether they learned to go beyond the call of duty (see the first iCTF outcome in Table 1).

Simulate scenarios. Place students into simulated scenarios that force them to use targeted skills and either automate measurements or pair someone with each participant

to evaluate their skills. For example, tabulate the practical skills students learn in each SEED lab and create computer-based scenarios in which students must employ these skills to achieve some goal. Measure the mastery of targeted skills by recording students' command-line input and then analyzing the input's sophistication and speed of achieving the goal.

Measure as students learn. Apply the same automated measurement approach from the simulated scenarios to the intervention. This would enable analysis of not only whether students learn with the intervention but also how they learn. For example, track command-line inputs for each participant and system state during an iCTF exercise and have someone analyze the sophistication of each participant's action, the evolution of this sophistication during the exercise, and the level of teamwork. For many interventions, however, learning happens in the preparatory phase as well as the execution phase, so this evaluation approach would have to span both phases.

Self-assessment. Survey students about their perception of what they know before and after the intervention using multiple-choice questions (for instance, rate your level of proficiency in network administration) or essay questions (for instance, list the tasks you know how to perform in network administration). Quiz and self-assessment approaches require the evaluator to precisely identify skills learned in an intervention. This might be challenging for loosely specified interventions like iCTF, in which students learn independently in an informal setting, and in which students might employ many skills to achieve the iCTF goal (such as hacking into the system). Furthermore, self-assessment might not align well with actual skill level because it's subjective. Simulated scenarios and measuring while learning offer more objective evidence of learning. But, if automated measurement is deployed, it requires instrumentation of the intervention and aggregation of low-level measurements into high-level goals. Pairing an evaluator with participants to measure skill attainment requires a large time investment and limits the evaluation scale.

Interest and Awareness

Measuring interest and awareness is well-understood and often applied in the education community. These qualities can be assessed using the following evaluation instruments.

Self-assessment. Participants are surveyed about their interest in the field or their intent to continue to be involved in cybersecurity.

Follow-up. Participants are tracked for some time after the intervention, and data is gathered to evaluate their engagement with the field—did they take a follow-up course, enroll in graduate school,

join a professional group, or continue self-learning?

Of these two options, self-assessment is easier to design and conduct but might overestimate the intervention's impact, as it's likely that some students will say they intend to stay engaged but won't follow up on those intentions with actions.

Control group. A control group—students who weren't affected by the intervention—is needed to properly measure an increase in interest and awareness. One way to achieve this would be to draw the control group from the same population as the intervention group. This works in the case of voluntary interventions such as iCTF and Control-Alt-Hack, in which a student can choose to participate or not. It doesn't work well in the case of class-based interventions, as a teacher usually wants all students to benefit from an intervention.

Another way to evaluate the impact of class-based interventions would be to draw a control group from a similar population—for example, students in cybersecurity classes at comparable institutions with regard to student demographics and institution strength and resources. In this case, care must be taken to control for variables such as different teachers, class syllabi, and so on.

We'd like to draw attention to a well-known fact in survey-based evaluations: how you ask questions matters.³ Care must be taken to not lead participants to any specific answer. For example, asking “did your interest increase?” biases the participant to answer “yes.” Instead, ask for a rating of the level of interest before and after the intervention and compare these numerical values to see if there was an increase. It's also possible that students learning about security and privacy will become more sensitive to sharing personal information. This makes

research consent even more important to protect participants' well-being while ensuring that the results are ecologically valid.

Knowledge

The following evaluation instruments could be applied to measure knowledge acquisition and retention.

Quiz on concepts. Students take a quiz with either multiple-choice or essay questions that evaluates their understanding of the concepts.

Simulated scenarios. Students are placed in simulated scenarios that require them to apply the knowledge they gained in the intervention. Automatically collect low-level measurements and convert them into higher-level concepts that students learned, or pair an evaluator with each student to determine which concepts they understand based on their actions and verbal output.

Of these two approaches, a quiz is much easier to design and conduct. To evaluate learning based on the intervention, the same evaluation would be conducted before and after the intervention. To evaluate knowledge retention, the same evaluation would be repeated after some time has passed since the intervention. This requires evaluators to collect student contact information and follow up with students after a specific interval.

Impact

Evaluate the impact of an intervention on specific populations by surveying the participants. A survey can consist of multiple-choice questions (for instance, rate any difficulties you had in adopting the materials on a scale of 0 to 5, where 0 is none and 5 is many difficulties) or essay questions (for instance, list any difficulties you had in adopting the material). Along with the collection of the level of impact, the demographics of participants or

institutional characteristics must be collected often. This is necessary so the impact can be reported separately for each group of interest (such as Hispanic Americans).

Establish a Systematic Method for Collecting the Information

For many of the above-listed evaluation instruments, evaluators could collect the required information from each participant in the intervention or from a randomly sampled subset.

The biggest challenge we faced in evaluating the three interventions was due to their wide adoption by other educators. This removed the participant pool from the intervention authors and made data collection difficult. Although it's possible to ask teachers who adopt an intervention to collect some data for its evaluation, the logistics are complex.

As a community of cybersecurity educators, we want successful and far-reaching interventions—those that are widely adopted and affect broad, international populations. Our efforts, then, should be directed to designing standard evaluation instruments that can be adopted by educators who use the intervention and whose results could be easily reported back to the intervention authors. Ideally, these evaluation instruments and result reports would be automated. For example, quizzes could be created by intervention authors and conducted using an online service like SurveyMonkey. Teachers who use an intervention would announce these quizzes in their classes at a specific time. As another example, student email addresses could be collected by requiring students to sign up before participating in an intervention. The intervention author would use the email

addresses after the intervention, for example, to evaluate retention in cybersecurity.

Reality Check

So, what was actually done to evaluate our chosen interventions? The intervention authors considered many of our approaches but implemented only some of them, owing to time and cost constraints, some ill-defined goal metrics, or a disagreement. In addition, it should

A viable path forward could be to increase interaction and collaboration between the cybersecurity and education communities.

be noted that most faculty positions are judged by research publications much more than teaching innovations as criteria for promotions and tenure. When innovative teaching approaches are designed and implemented—often as a sideline from primary research directions—the time and effort needed to properly evaluate their impact on learning can be prohibitively expensive (in terms of time or monetary investment), and venues likely to publish the work might not be considered top-tier in the faculty member's area.

The iCTF authors considered many of our evaluation approaches and implemented a postexercise survey in 2009 to measure team composition, skill development, and learning.⁴ Of 56 teams, 35 responded. The survey asked simple yes or no questions and thus provided only coarse evidence of iCTF's positive effects. iCTF was further studied by education researchers to evaluate the impact of teamwork on cyberdefense effectiveness.⁵ Researchers followed two teams from the 2011 iCTF and found that the

better-organized team fared much better in the competition.

SEED employed pre- and postlab questionnaires to evaluate student learning and gathered a great deal of valuable data. Evaluation adoption information and teacher feedback were collected voluntarily and were thus incomplete—if a teacher didn't report back with information, the author had no way to collect it. This highlights the necessity of designing automated information gathering and possibly integrating the measurements with the materials. For example, users who want to download materials could be required to first input their affiliation and email address. The system could then periodically

email teachers with a survey link asking for feedback. Although teachers might not reply to these emails, an automated system such as this has a higher chance of gathering useful information than an approach that relies on teachers volunteering feedback.

The Control-Alt-Hack creators opted to send open-ended surveys to educators who adopted the intervention and to conduct small user studies with pre- and postintervention questionnaires, the results of which were published.⁶ Only 22 out of 150 polled educators returned the surveys. Those who replied represented more than 450 students at the high school, undergraduate, and graduate level. The authors elected to use open-ended questions to avoid leading questions. Although this method requires more manual analysis, it helps provide answers that could be missed in a more structured survey. Authors reported the number of educators who described positive or negative results for engagement and awareness and recounted specific critiques from the surveys.

Overall, we see the following trends in these evaluations: simple evaluation instruments are favored because of low time and effort cost, even though they might not be best-suited to the evaluation task; evaluation is often done as an afterthought and in an ad hoc manner, resulting in a small sample size and impacting validity; evaluation is often done by cybersecurity experts and not education experts; and intervention authors understand how to design better evaluations but lack the time and resources for this task. Furthermore, some intervention goals might be difficult to measure quantitatively, for example, measuring whether a student has acquired an adversarial mindset.

A Viable Path Forward

The complexity of designing and delivering an evaluation with strong validity is understated. Many confounding factors can make it very difficult to attribute outcomes to specific interventions. As the cybersecurity field evolves, we need to investigate and report on the effects of educational interventions on cybersecurity awareness, learning, retention, motivation, interest, and so on. We then need to be circumspect in how advances in these areas contribute to cybersecurity in knowing which interventions to stop, start, and continue. Despite the numerous challenges in doing this well, this is important work.

Both the complexity and cost of sound evaluation approaches indicate the need for us to address these issues as a community. As we discussed, many interventions share the same goals of increasing awareness, teaching practical skills, improving learning and knowledge retention, and so on. This makes it possible for a community of researchers to each invest a fraction of their time and develop sound evaluation instruments that quantitatively measure the attainment

of the chosen goal and can be coupled with automated data gathering. Others could then easily locate the appropriate instruments, integrate them with their interventions, and invest time only in processing and reporting on the automatically gathered data.

A viable path forward could be to increase interaction and collaboration between the cybersecurity and education communities. Education experts understand how to design sound evaluation instruments and how to gather data—this is what they do for living. Cybersecurity experts are often enthusiastic about teaching and designing new interventions but don't have evaluation cycles. Teaming up might just be the winning combination to producing sound evaluations of cybersecurity education interventions without jeopardizing careers.

We welcome feedback from readers on educational evaluation needs to help us discern priority areas. ■

References

1. M. Dark and J. Mirkovic, "Evaluation Theory and Practice Applied to Cybersecurity Education," *IEEE Security & Privacy*, vol. 13, no. 2, 2015, pp. 75–80.
2. P. Rossi, M. Lipsey, and H.E. Freeman, *Evaluation: A Systematic Approach*, Sage Publications, 7th edition, 2003.
3. D. Dillman and P. Salant, *How to Conduct Your Own Survey*, Wiley, 1994.
4. N. Childers et al., "Organizing Large Scale Hacking Competitions," *Proc. 7th Int'l. Conf. Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA 10)*, 2010, pp. 132–152.
5. S. Jariwala et al., "Influence of Team Communication and Coordination on the Performance of Teams at the iCTF Competition," *Proc. Human Factors and Ergonomics*

Society Ann. Meeting, vol. 56, no. 1, 2014, pp. 458–462.

6. T. Denning et al., "Control-Alt-Hack: The Design and Evaluation of a Card Game for Computer Security Awareness and Education," *Proc. 2013 ACM SIGSAC Conf. Computer & Communications Security (CCS 13)*, 2013, pp. 915–928.


Jelena Mirkovic is a research faculty member and project leader at the University of Southern California's Information Sciences Institute. Contact her at mirkovic@isi.edu.

Melissa Dark is the W.C. Furnas Professor of Technology in the College of Technology at Purdue University and the assistant director of Educational Programs at CERIAS (the Center for Education and Research in Information Assurance and Security). Contact her at dark@purdue.edu.

Wenliang Du is a professor in the Department of Electrical Engineering and Computer Science at Syracuse University. Contact him at wedu@syr.edu.

Giovanni Vigna is a professor in the Department of Computer Science at the University of California at Santa Barbara and the CTO at Lastline, Inc. Contact him at vigna@cs.ucsb.edu.

Tamara Denning is an assistant professor in the School of Computing at University of Utah. Contact her at tdenning@cs.utah.edu.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.

Got an idea for a future article?
Email editor Melissa Dark (dark@purdue.edu).