# Challenges Arising from Prerequisite Testing in Cybersecurity Games

Valdemar Švábenský
Masaryk University
Faculty of Informatics
Institute of Computer Science
Brno, Czech Republic
svabensky@ics.muni.cz

Jan Vykopal
Masaryk University
Institute of Computer Science
Brno, Czech Republic
vykopal@ics.muni.cz

## ABSTRACT

Cybersecurity games are an attractive and popular method of active learning. However, the majority of current games are created for advanced players, which often leads to frustration in less experienced learners. Therefore, we decided to focus on a diagnostic assessment of participants entering the games. We assume that information about the players' knowledge, skills, and experience enables tutors or learning environments to suitably assist participants with game challenges and maximize learning in their virtual adventure. In this paper, we present a pioneering experiment examining the predictive value of a short quiz and self-assessment for identifying learners' readiness before playing a cybersecurity game. We hypothesized that these predictors would model players' performance. A linear regression analysis showed that the game performance can be accurately predicted by well-designed prerequisite testing, but not by self-assessment. At the same time, we identified major challenges related to the design of pretests for cybersecurity games: calibrating test questions with respect to the skills relevant for the game, minimizing the quiz's length while maximizing its informative value, and embedding the pretest in the game. Our results are relevant for educational researchers and cybersecurity instructors of students at all learning levels.

## KEYWORDS

active learning, cybersecurity games, diagnostic assessment, prerequisite testing, self-assessment, linear regression modeling

**ACM Reference format:**
Valdemar Švábenský and Jan Vykopal. 2018. Challenges Arising from Prerequisite Testing in Cybersecurity Games. In *Proceedings of The 49th ACM Technical Symposium on Computer Science Education, Baltimore, MD, USA, February 21–24, 2018 (SIGCSE '18),* 6 pages.
https://doi.org/10.1145/3159450.3159454

## 1 INTRODUCTION

Cybersecurity games allow participants to test their knowledge and exercise their skills in different areas of computer security. Although carried out in a closed and controlled environment, the games simulate practical, real-world situations. The players can attack and defend computer systems, analyze network traffic, or disassemble binaries without any negative consequences in reality.

Studies confirm multiple benefits of cybersecurity games [16, 17, 21]. They can inspire interest in computer security and motivate participants to explore the field further. Games designed specifically for education enrich the curriculum and test the learners' competence in an authentic setting, enabling them to discover their strengths and weaknesses. Ranking well in competitive games often leads to peer recognition, (monetary) prizes, or job opportunities.

Competitions and games of various difficulty levels and focus are spreading widely, from informal online hacking communities to universities and professional security conferences. The number of participants in cybersecurity games is growing exponentially [17]. At the same time, several authors argue that although high-quality games are available, they offer little educational value to learners [16, 20]. The games often require substantial knowledge of the problem domain, as well as practical expertise, in advance. As a result, the majority of computer science students are unable to participate. Even worse, some students' interest and motivation may diminish after an unsuccessful attempt [16]. Research suggests that games and contests are effective only for already skilled players, whose skills "closely match those required by the competition" [17].

Achieving *game balance* (assigning tasks that are suitable for the player's skill, neither trivial nor impossible to solve [14, 16]) is vital in educational games. One approach to achieving game balance is introducing methods of adaptive learning [4], which change the difficulty of the tasks during the game based on the player's success rate. Another solution is a diagnostic assessment by prerequisite testing, which is the topic of this paper. This approach, suggested in pedagogical theory [9, 13], refers to testing the player before or during a game to determine whether the player's skills are sufficient to finish the tasks, thus providing game balance [16].

This work's main motivation is the demand for timely identification of students who may require help while playing, so that their individual needs can be appropriately addressed. This can be done by providing learners with more precise instructions, hints, or relevant study materials.

To explore the predictive value of both self-assessment and prerequisite tests, we conducted an experiment involving 67 learners and two games. They completed a short questionnaire, played a game, and then reflected on their experience in another survey. Tracking the players' actions during the game allows us to compare the assessment results with the in-game performance. Based on this experiment, we seek to answer two research questions:

*SIGCSE '18, February 21–24, 2018, Baltimore, MD, USA*

(1) Can a prerequisite test accurately model learners' performance in cybersecurity games?
(2) Is self-assessment a reliable indicator of ability in the context of cybersecurity?

## 2 RELATED WORK

Research into diagnostic assessment of cybersecurity skills is sparse and not yet mature. Therefore, this section also mentions works dealing with assessment and testing in other educational domains.

While Nagarajan et al. [14] stress that measuring skills before and after playing is vital to determine the game's effectiveness, they report that security training programs do not implement this measurement. To the best of our knowledge, we can confirm this observation.

Mirkovic et al. [11, 12] emphasize considering individual skills in team cybersecurity games to balance the teams and give everyone an equal chance to succeed. Before using cybersecurity games in classrooms, Mirkovic and Peterson [11] surveyed the students about their skills to create balanced teams. Unfortunately, the paper does not provide details about the process. In another study [12], the participants reported their knowledge of programming, security, and tools that were to be used later. Again, the survey results were used to balance the teams. However, the authors concluded that this led to inequality among the teams, as the self-assessment was often inaccurate. They recommend "conducting a short quiz-type assessment prior to the event", but do not specify how to do this.

Next, Bolívar-Cruz et al. [3] examined self-assessment of university students in oral communication. Their literature review shows that self-assessment's accuracy is generally low or questionable, but also warns readers about methodological errors in some of the previous studies. Allen and Van Der Velden [1] advocate using self-assessment complemented by independent, objective tests to increase the reliability of results. They argue that people know the level of their skills best but also warn readers of its issues, including misunderstood skill items, an ambiguous rating scale, and the risk of an unreliable answer (either intentional or not).

Finally, Govindasamy [5] suggests applying pretests in e-learning courses to test both minimum requirements and proficiency. Based on the results, the learner can be directed to a simpler or more difficult course, or skip the already familiar areas in the current course. Educational literature [9, 13, 15] advocates the use of prerequisite testing in teaching practice.

## 3 STRUCTURE OF A CYBERSECURITY GAME

We use instances of a cybersecurity game following a generic format of a hands-on activity, which is performed in a realistic network environment emulated by the KYPO cyber range [19].

Figure 1 shows the scheme of the game, which is structured into successive levels leading to the final objective, such as data theft. Before the start, each player has access to limited network resources and a brief information about the goal. Every level is finished by finding a correct flag (a short string); this accomplishment is awarded a specified number of points contributing to the player's total score. The game ends upon entering the last flag or when a predefined final check of the system's state succeeds.

The game provides optional scaffolding by offering hints. If the player struggles with a level, these hints can be used in exchange for penalization by negative points. There is evidence that game elements such as points and levels can improve the overall effectiveness of learning [8]. It is also possible to skip the level, display the recommended solution, and quit the game at any time.

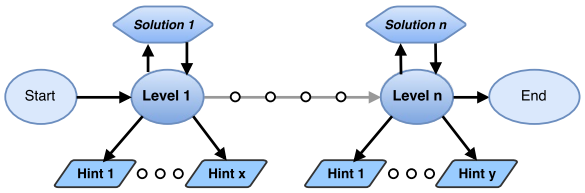


**Figure 1: The general structure of the cybersecurity game used in the experiment**

**Table 1: The design of the experiment with the number of participants and their gender**

| 67 participants | | | | |
|---|---|---|---|---|
| Session 1<br>9 ♂, 1 ♀ | Session 2<br>4 ♂, 3 ♀ | Session 3<br>15 ♂, 3 ♀ | Session 4<br>12 ♂, 0 ♀ | Session 5<br>19 ♂, 1 ♀ |
| Pretest 1<br>5 items | | + 2 items | | Pretest 2<br>4 items |
| Self-assessment<br>3 items | | | | |
| Game 1<br>6 levels | | | | Game 2<br>4 levels |
| Post-game feedback | | | | |

The generic nature of the game format allows us to collect generic game events, regardless of the topic of the particular game and technical infrastructure used. The game events describe the player's interaction with the game interface, namely: starting and ending the game or each level, submitting incorrect flags and their content, using hints, skipping a level, and displaying a solution. Each event contains a timestamp and a unique ID of the player.

## 4 EXPERIMENT DESIGN

Table 1 summarizes the phases and structure of the experiment. At first, each player completed a prerequisite test and a self-assessment questionnaire. The players then proceeded to a game, where their performance was tracked using the generic game events. Finally, the players filled in a post-game feedback questionnaire. The self-assessment and pretest data were used to create a linear regression model of learners' skill, which is expressed by two metrics: the total game score and number of levels finished.

### 4.1 Participants

A total of 67 cybersecurity students and professionals of various levels of expertise, background, and nationality participated, covering a broad spectrum of the games' target audience. The players' only motivation was their interest, as they did not receive any incentives for taking part in the study. The participants were informed about the intended use of the acquired data solely for the purpose of this experiment. The data was anonymized during the processing.

The learners were divided into five game sessions, each lasting two hours. The first session included 10 computer science students

**Table 2: Characteristics of the selected games G1 and G2**

| Level | Score [pts] | | Hints | | Penalty [pts] | |
|---|---|---|---|---|---|---|
| | G1 | G2 | G1 | G2 | G1 | G2 |
| 1 | 8 | 16 | 2 | 2 | −2, −2 | −3, −5 |
| 2 | 12 | 22 | 2 | 2 | −2, −3 | −7, −5 |
| 3 | 23 | 27 | 3 | 4 | −1, −3, −2 | −5, −10, −0, −0 |
| 4 | 20 | 35 | 2 | 4 | −2, −3 | −5, −10, −5, −5 |
| 5 | 22 | — | 2 | — | −3, −4 | — |
| 6 | 15 | — | 2 | — | −5, −2 | — |
| Total | 100 | 100 | 13 | 12 | −34 | −60 |

from St. Pölten University of Applied Sciences, Austria. The second session consisted of 7 employees of the Computer Security Incident Response Team of Pavol Jozef Šafárik University in Košice, Slovakia. The third session included 18 computer science students from Masaryk University, Brno, Czech Republic. The fourth session included 12 finalists of a Czech high school cybersecurity competition. The fifth session included 20 attendees of the AIMS 2017 conference on network security and management, held in Zurich, Switzerland.

## 4.2 Selected cybersecurity games

We have two games G1 and G2 in this experiment to test various game tasks and prerequisite tests. Table 2 details the maximum score, the number of hints, and scoring penalties for taking the hints in each level of both games.

The topic of the G1 game played in the first four game sessions is information theft from a database server of a fictitious bank. Each player initially controls a single Linux host in an unknown network. The player must gradually gain and maintain access to other hosts that are a part of the bank's network infrastructure, and, finally, steal confidential information. This mission is split into six levels, in which the players exercise penetration testing skills.

To provide comparative data, 20 participants of the fifth game session played another game, G2, with the topic of gaining access to a remote server and destroying stored data. This game is split into four levels with learning objectives similar to the levels of G1.

## 4.3 Skill measurement before the game

We created quizzes testing prerequisites for each of the two selected games using a model for question design by Beatty et al. [2]. Details about the pretest's design can be found in [22]. In the first three game sessions, the players completed a pretest consisting of five questions. In the fourth session, this pretest was enhanced by two extra questions. In the fifth session, a different but similar pretest consisting of four questions was used. The prerequisite quizzes capture a representative sample of key knowledge and skills exercised in the games. Below is an example of a question for the first level of both games (an asterisk marks the correct answer):

```
What is the effect of the command ping 10.0.0.3?
a)* Tests the reachability of a host with an IP address 10.0.0.3.
b)  Scans open ports of the server with an IP address 10.0.0.3.
c)  Error, the syntax of the command is incorrect.
d)  Measures the number of network hops to a host with an IP address
    10.0.0.3.
```

To evaluate the test, we used a simple *dichotomous* scoring method awarding one point for a fully correct answer, and zero points for a partially or entirely incorrect answer per question. Moreover, after responding to each question, a learner rated the level of certainty in the answer on a five-step scale developed by Hassmén and Hunt [6, 7]. This scoring method, referred to as *confidence assessment*, yielded another test score. In both cases, a sum of the respective scores was considered as an estimate of each learner's total readiness.

Apart from the prerequisite test, each player completed a 3-item self-assessment questionnaire before starting the game. Since there is no standardized methodology for designing the questions, we created them with respect to the content of the particular games. The survey asked the players to self-evaluate their expertise with using three tools needed in the games: for port scanning (Nmap), vulnerability exploiting (Metasploit), and password attacks (John the Ripper). For each tool, the player selected one of four levels of competence on the following ordered scale: zero experience, beginner (basic knowledge), intermediate (some practical experience), and expert (professional working experience). To aggregate the learners' input, we used the median to express the central tendency of each player's self-assessment. Since the self-assessment data are ordinal, we avoided using the arithmetic mean.

## 4.4 Post-game feedback

After finishing the game, the participants completed a post-game feedback questionnaire. The goal of the survey was to have each player subjectively assess the game's difficulty on a scale from 1 (trivial) to 5 (impossible), and reflect if any learning occurred. This reflection helps to determine if game balance was achieved, and if the player perceived the game as educational. Unfortunately, due to a technical error, we collected the results from only 46 participants.

# 5 RESULTS

Table 3 reports the examined variables and descriptive statistics of the collected data for G1, which are further detailed in Figure 2. The boxplots show distributions of the game score ($T$) grouped by the dichotomous quiz score ($P_d$) and the input from the self-assessment ($S$). The boxplots showing the distributions of finished levels ($L$) are almost identical considering patterns of the medians, and thus were omitted. Also, statistically significant ($p \leq 0.02$) Pearson and Spearman correlations (ranging from 0.36 to 0.60) were reported between the score and pretest and between levels completed and pretest (both scoring methods in both cases). Finally, there was strong evidence that all performance predictors and skill descriptors negatively correlate with how difficult the game is perceived ($p \leq 0.02$, coefficients ranging from $-0.37$ to $-0.52$).

Next, we modeled the learners' skill (dependent variables $T$ and $L$) using independent variables $P_d$, $P_c$, and $S$. Three different sets of regression analyses were performed. First, we used the data of the 47 participants playing G1, ignoring the two extra pretest questions given in session 4. Second, we used only the data of the 12 participants playing G1, using the result of the 7-question pretest. Third, we used the data of the 20 participants playing G2.

Table 4 reports the regression models of the first set of analyses. Statistically significant fits were computed for the score or level

**Table 3: Examined variables and descriptive statistics of the participant data for the four G1 sessions and 5-item pretest**

| Variable | | Possible range | Min | Max | Avg | Med |
|---|---|---|---|---|---|---|
| Self-assessment | $S$ | 0 to 3 | 0 | 2 | 0.9 | 1 |
| Pretest (dich.) | $P_d$ | 0 to 5 | 0 | 5 | 3.9 | 4 |
| Pretest (conf.) | $P_c$ | −300 to 250 | −91 | 240 | 150.6 | 155 |
| Game score | $T$ | 0 to 100 | 0 | 100 | 49.7 | 55 |
| Levels finished | $L$ | 0 to 6 | 0 | 6 | 3.5 | 4 |
| Difficulty | $D$ | 1 to 5 | 2 | 5 | 3.5 | 3 |

**Table 4: The overall best linear regression models (see Table 3 for the description of the variables).**

| Game | Model | $R^2$ | F-statistic | p-value |
|---|---|---|---|---|
| G1 | $T = 8.98 + 10.34 \cdot P_d$ | 0.17 | 9.36 | 0.004 |
| G1 | $T = 10.87 + 0.23 \cdot P_c$ | 0.31 | 15.66 | $< 0.001$ |
| G1 | $L = 1.62 + 0.48 \cdot P_d$ | 0.14 | 7.22 | 0.010 |
| G1 | $L = 1.73 + 0.01 \cdot P_c$ | 0.24 | 10.99 | 0.002 |

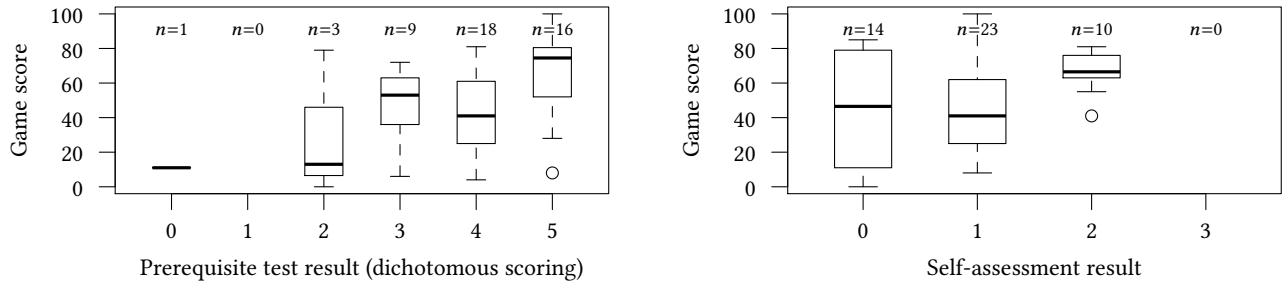


**Figure 2: Boxplots depicting relationships between test score, self-assessment, and game score in G1**

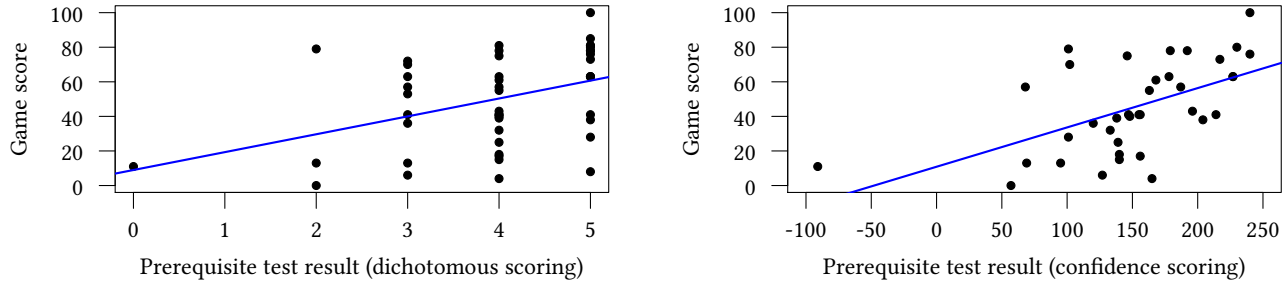


**Figure 3: Linear regression models describing game score by pretest for G1**

prediction based on the dichotomously scored pretest ($R^2 = 0.17$ or $R^2 = 0.14, p \leq 0.01$). An even more promising relationship emerged when incorporating confidence testing ($R^2 = 0.31$ or $R^2 = 0.24$, $p \leq 0.004$). The second set of analyses on the subset of players and two extra questions yielded almost identical results to those on the full sample of players from G1. The coefficients in the remaining models did not show statistical significance, and neither did they in other models for G1 nor in any model for G2.

The two best fits for score prediction are graphed in Figure 3. These plots show a certain degree of linearity. The other two plots for predicting completed levels were largely similar and were omitted to conserve space. The regression diagnostic plots (residual plots and Q-Q plots) confirm that the assumptions of homoscedasticity and multivariate normality were met. Based on the leverage plot, the player scoring 0 points in the pretest was identified as an outlier (Cook's distance > 0.5). However, the removal of the data point did not significantly influence the models (the $R^2$ changed by ±0.01, and the p-value remained up to 0.01). Thus, we decided to keep the data point in the sample.

# 6 DISCUSSION

## 6.1 Quantitative view

Returning to the research questions posed in Section 1, the models accurately predicted the learners' total game score and completed levels when employing prerequisite testing, regardless of the scoring method used, for the game G1. It is interesting that confidence testing revealed guesswork: 8 players randomly guessed at least one correct answer. It also showed misconceptions, since 7 players were quite or absolutely sure about at least one incorrect answer. Most importantly, the confidence testing improved the models, as the value of $R^2$ almost doubled compared to the dichotomous pretest.

Considering G1, the players from the fourth session exhibited similar results as the whole when using the two extra pretest questions. However, the success was not reproduced in G2. We attribute this to the fact that two of the questions were easy for the players to answer. Moreover, due to unclear formulation of the third question we decided to accept even the answers we originally deemed incorrect. As a result, 16 out of 20 players had at least three questions correct, preventing the models from differentiating between them.

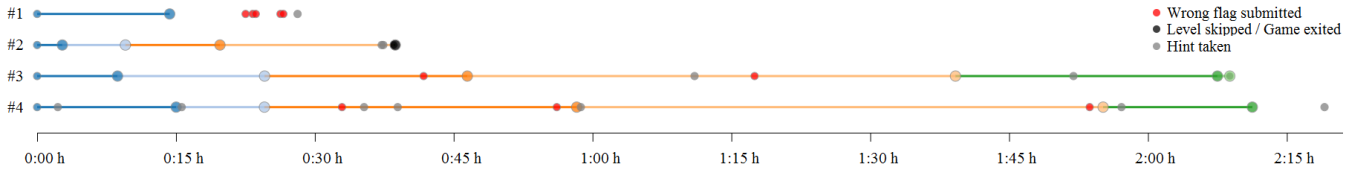


**Figure 4: Game events of selected individual players in G1, distributed over time. Each line represents the actions of one player. Finished levels are marked as colored line segments: dark blue displays level 1, light blue level 2, and so on.**

To complement the discussion of the regression models, we examined medians in boxplots depicting prerequisite test results (Figure 2). On the one hand, the medians do not show a linear trend, as they are not in ascending order. On the other hand, the players who achieved the highest quiz scores also had the highest medians of the game score and finished levels. Overall, the results confirmed the intuition that players with a high score from the pretest would often perform better in the game compared to the others. The reverse was also true.

However, the regression models employing self-assessment were not statistically significant, and their $R^2$ values were below 0.07. The models seem unlikely to reach statistical significance even if the sample size increased. The conclusion that the self-assessment is an unreliable skill predictor is consistent with the previous findings by Mirkovic et. al [11], who used a similar self-assessment scale. One possible explanation is that experience with using in-game tools is not a key factor in deciding learners' readiness. Most of the players had worked with a Linux Terminal before the game, thus were able to discover and understand the application of other command-line tools for themselves. Another plausible justification for the poor self-assessment results is that the scale has only four values and three items, which yields data that is too coarse.

## 6.2 Qualitative view

It can be argued that applying statistical tests and using regression models on a relatively small sample might bias the quantitative results. Therefore, the actions of individual players were further explored from a time perspective (see Figure 4). Several notable anomalies were identified in G1 sessions and are addressed below[1].

Player #1, who we nicknamed "the dropout", reached the full score (5 points) in the dichotomous pretest but got frustrated as early as in level 2. Over the course of less than 6 minutes, he attempted 5 wrong flags, took a hint, and stopped playing. As a result, his game score was only 8 points.

Player #2 ("the achiever") followed the same pattern as "the dropout"; at the beginning, he seemed like a competent learner but got frustrated with the game. "The achiever" scored 4 points in the dichotomous pretest and solved the first three levels quickly. Shortly before the indicative time limit for the fourth level ran out, he took both available hints and then prematurely exited the game. As a result, he scored a, below-average, 43 game points. However, he later reported not knowing that the time limit was only informative and had no impact on the game. Instead, the player thought that if the time runs out, he cannot play anymore, which annoyed him and caused him to quit the game.

[1]All the players are referred to as males, even if their gender is unknown.

We hypothesize that "the achiever" and "the dropout" had possessed the necessary prerequisites for finishing the game. However, they were thwarted by ambiguous game mechanics or design, by insufficient attention paid to the rules, or by some other reason. Due to these unanticipated situations, the dataset includes players scoring well in the pretest but poorly in terms of game score or levels completed. This might have introduced noise in the regression models, as similar misunderstandings could have influenced other players' results.

Another interesting case is player #3 ("the determined one"), who scored 0 in the self-assessment and 2 points in the pretest. Still, he completed 5 levels and scored 79 points in the game: one of the best results in the sample. The player used only 2 hints and attempted only 2 incorrect flags in total, all in the later phases of the game. The time spent in the levels was rather long. This is reflected in the post-game feedback, where he rated the game as hard (4). Although the player did not possess theoretical knowledge from the pretest, his determination allowed him to perform very well.

Finally, the player #4 ("the practitioner") scored 0 in the self-assessment and 3 points in the pretest. However, by taking some hints, using trial and error, and given enough time he was able to complete 5 levels and score 72 game points, which is a good result. This player, like "the determined one", might not have had the theoretical background, but was still able to solve the practical tasks.

These case anomalies show that some unanticipated aspects influence players' performance. An arising challenge is recognizing and deeply understanding all factors that contribute to a successful game. We believe that solving this challenge is essential for designing a useful diagnostic assessment and the whole game.

## 6.3 Addressing the limitations

Despite using a well-established framework for question-writing [2] and following best practices of assessment design, we were confronted by three main challenges of prerequisite testing. The first is calibrating the test to predict the possession of skills most relevant to the game. While the players often performed well in the quiz, no one finished the last level of G1. It seems that theoretical knowledge might not be enough for succeeding in practical tasks.

The second challenge was the limited time frame for assessing a participant. It is impractical and discouraging to perform a lengthy examination when the learners are eager to play the game. Both the test and self-assessment combined were designed to take 8 minutes at most, yet were perceived by some of the players as an inconvenience. The third challenge is embedding the pretest in the game. For further experiments, we propose designing and implementing pretests, and, by extension, cybersecurity games

differently. Inspired by the results of Lee at al. [10], who report positive effects of assessments in educational games, we propose two main improvements. One is dissolving the assessments into the story of the game. Compared to using questionnaires, which distract the players and shift them into a "testing mode", in-game tests are more engaging [10]. They also allow the use of more assessment questions, which, in turn, brings more validity and reliability to the results. This approach necessitates another improvement in the design of the game itself. Individual levels can be created such that each has only one particular learning outcome. Appropriate prerequisites can be tested before or during that level.

While the experiment proved a link between prerequisite testing and players' performance in the game G1, the generalizability of the results might be questionable, as the model for G2 was not statistically significant. One possible explanation is the dependence on the particular game and its scoring method. Ultimately, successful diagnostic assessment largely depends on the quality of game design. This work attempted to prove the validity of the proposed prerequisite test based, to some extent, on its relationship to the game. However, if the game scoring mechanism or individual levels are poorly designed, this can invalidate the pretest. Therefore, we underline the need for careful consideration of educational game design. Another explanation is that the amount and content of questions in pretest 2 were not sufficient to differentiate the players' skill. Nevertheless, adding more questions in pretest 1 did not seem to bring more validity to the results, as they reflected the results of the analysis performed with the subset of questions.

Finally, the self-assessment had only three items with four distinct values, which was too coarse an input for the linear regression. Introducing questions asking for the frequency of use or applicability to a particular task could increase the reliability of results.

## 7 CONCLUSIONS

We performed an experimental study investigating the predictive value of prerequisite tests and self-assessment for identifying learners' readiness before playing a cybersecurity educational game. The analysis of game events and information provided by players showed that only a knowledge quiz, and not self-assessment, can model game score and completed levels. The models based on the pretests significantly improved if the quiz contained confidence assessment. However, educators have to pay special attention to the selection and formulation of quiz questions since they fundamentally affect the accuracy of the model.

The major contribution of this pioneering attempt is the new insights provided into hands-on cybersecurity education, which has not been widely researched. This work also motivated the development of an open-source tool for visualizing generic game events over time [18] (see Figure 4). This tool allowed discovering important patterns that would otherwise stay hidden.

In our future work, we will focus on investigating the means of diagnostic assessment into the game story and structure. Since the players rated the games as educational, practical, and interesting, we believe that active learning in cybersecurity is worthy of both security practitioners' and educators' attention. We also encourage fellow researchers to experiment with diagnostic assessment in other domains than cybersecurity.

## ACKNOWLEDGMENTS

This research was supported by the Security Research Programme of the Czech Republic 2015-2020 (BV III/1–VS) granted by the Ministry of the Interior of the Czech Republic under No. VI20162019014 – Simulation, detection, and mitigation of cyber threats endangering critical infrastructure. The analysis of the experiment results would not have been possible without a visualization tool developed by Juraj Uhlár. Finally, we thank our colleagues and anonymous reviewers for their helpful comments and suggestions.

## REFERENCES

[1] Jim Allen and Rolf Van Der Velden. 2005. The role of self-assessment in measuring skills. In *Transition in Youth Workshop, Valencia, Spain.*
[2] Ian D. Beatty, William J. Gerace, William J. Leonard, and Robert J. Dufresne. 2006. Designing effective questions for classroom response system teaching. *American Journal of Physics* 74, 1 (2006), 31–39.
[3] Alicia Bolívar-Cruz, Domingo Verano-Tacoronte, and Sara M. González-Betancor. 2015. Is University Students' Self-Assessment Accurate? In *Sustainable Learning in Higher Education.* Springer, 21–35.
[4] Mark Gondree, Zachary Peterson, and Portia Pusey. 2016. Talking about talking about cybersecurity games. *USENIX ;login:* 41, 1 (2016), 36–39.
[5] Thavamalar Govindasamy. 2001. Successful implementation of e-learning: Pedagogical considerations. *The internet and higher education* 4, 3 (2001), 287–299.
[6] Peter Hassmén and Darwin P. Hunt. 1994. Human Self-Assessment in Multiple-Choice Testing. *Journal of Educational Measurement* 31, 2 (1994), 149–160.
[7] Darwin P. Hunt. 2003. The concept of knowledge and how to measure it. *Journal of intellectual capital* 4, 1 (2003), 100–113.
[8] Jincheul Jang, Jason J. Y. Park, and Mun Y. Yi. 2015. *Gamification of Online Learning.* Springer International Publishing, 646–649.
[9] Piet Kommers. 2004. *Cognitive support for learning: imagining the unknown.* IOS Press.
[10] Michael J. Lee, Andrew J. Ko, and Irwin Kwan. 2013. In-game Assessments Increase Novice Programmers' Engagement and Level Completion Speed. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research (ICER '13).* ACM, New York, NY, USA, 153–160.
[11] Jelena Mirkovic and Peter A. H. Peterson. 2014. Class Capture-the-Flag Exercises. In *2014 USENIX 3GSE.* San Diego, CA.
[12] Jelena Mirkovic, Aimee Tabor, Simon Woo, and Portia Pusey. 2015. Engaging Novices in Cybersecurity Competitions: A Vision and Lessons Learned at ACM Tapia 2015. In *2015 USENIX 3GSE.* USENIX Association, Washington, D.C.
[13] Gary R. Morrison, Steven M. Ross, Jerrold E. Kemp, and Howard Kalman. 2010. *Designing effective instruction.* John Wiley & Sons.
[14] Ajay Nagarajan, Jan M. Allbeck, Arun Sood, and Terry L. Janssen. 2012. Exploring game design for cybersecurity training. In *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER).* 256–262.
[15] Geoffrey Petty. 2009. *Teaching Today: A Practical Guide.* Nelson Thornes.
[16] Portia Pusey, Sr. David Tobey, and Ralph Soule. 2014. An Argument for Game Balance: Improving Student Engagement by Matching Difficulty Level with Learner Readiness. In *2014 USENIX 3GSE.* USENIX Association, San Diego, CA.
[17] David H. Tobey, Portia Pusey, and Diana L. Burley. 2014. Engaging Learners in Cybersecurity Careers: Lessons from the Launch of the National Cyber League. *ACM Inroads* 5, 1 (2014), 53–56.
[18] Juraj Uhlár. 2017. *Visualization of a run of a security game.* Bachelor's thesis. Masaryk University, Faculty of Informatics, Brno. https://is.muni.cz/th/422160/fi_b/.
[19] Jan Vykopal, Radek Oslejsek, Pavel Celeda, Martin Vizvary, and Daniel Tovarnak. 2017. KYPO Cyber Range: Design and Use Cases. In *Proceedings of the 12th International Conference on Software Technologies - Volume 1: ICSOFT.* INSTICC, SciTePress, 310–321. https://doi.org/10.5220/0006428203100321
[20] Joseph Werther, Michael Zhivich, Tim Leek, and Nickolai Zeldovich. 2011. Experiences in Cyber Security Education: The MIT Lincoln Laboratory Capture-the-flag Exercise. In *Proceedings of the 4th Conference on Cyber Security Experimentation and Test (CSET'11).* USENIX Association, Berkeley, CA, USA, 12–12.
[21] Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D. Van Der Spek. 2013. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology* 105, 2 (2013), 249.
[22] Valdemar Švábenský. 2017. *Prerequisite testing of cybersecurity skills.* Master's thesis. Masaryk University, Faculty of Informatics, Brno. https://is.muni.cz/th/395868/fi_m/.