CTF-based Self-study Exam in a STEM Course

Pedro J. Roig[1], Salvador Alcaraz[1], Katja Gilly[1], Cristina Bernad[1], Carlos Juiz[2]
[1]*Miguel Hernández University, Av. Universidad s/n (Ed. Alcudia), Elche*
[2]*University of the Balearic Islands, Ctra. Valldemossa km 7.5, Palma de Mallorca*
Spain

Abstract

Self-study exams are a way for students to assess their knowledge and skills before facing real exams. There are different approaches when it comes to propose a self-study exam, even though the use of a Capture The Flag (CTF) layout provides a way to interact with the questions proposed, as opposed to traditional exams, where no interaction is available. In this paper, a CTF-based self-study exam has been designed for a course on introduction to computer science at college level with the aim of getting students more prepared in order to face the final exam. This CTF-based exam was evaluated according to a weighted scoring system, where the overall score obtained was modified according to the number of hints unveiled along the questions and the time elapsed during the whole exam. The outcome achieved displayed a rise of about 10% when it comes to the results of the final exam related to the previous course held in the previous year where no self-study exam was proposed. On the other hand, an inferential study was performed on the outcome obtained, whose results showed that the increase was not statistically significant. Moreover, the actual sample size was insufficient in order to attain the effect size achieved. Hence, further research must be done with a greater amount of students.

1. Introduction

Capture The Flag (CTF) competitions were initially created to improve the technical skills within the cybersecurity community, as CTF challenges allow to test the knowledge and skills of the contestants related to cybersecurity without having to get into unauthorized networks or systems [1]. This way, participants could face real life issues regarding computing and network security in a controlled environment [2]. Some of the most common topics in this field are cryptography, steganography, web exploitation, or reverse engineering [3].

There are different formats of CTF challenges, such as attack-defense, king of the hill, hack quest, or jeopardy, where the latter consists of a bunch of teams competing to answer questions or solve challenges [4]. A typical format of jeopardy CTF consists of a mission, composed of a series of tasks to be carried out, where some hints may or may not be available for each task [5]. Competitors get a certain amount of points for solving each task, which is inversely proportional to the time spent in order to solve it, such that the lesser the time spent, the more points are obtained [6]. Hence, in this type of format, the CTF winner is usually the first individual competitor or team to complete the whole mission, as it is usually the one getting more points. However, in case there are hints available, each time a participant takes a hint in order to get a clue about how to solve a task, then a certain penalization is applied, thus meaning that a certain amount of points are subtracted from its score, which is directly proportional to the number of hints unveiled [7]. Therefore, the CTF winner might not be the contestant spending less time clearing the mission, as the CTF runner-up might end up with more points if there is not much time difference and the number of hints consumed is lower [8].

In fact, it can be stated that the score of each participant depends on two variables, such as the time elapsed and the number of hints consumed, although the number of points tied to each variable depends on the CTF designer [9]. It is to be noted that CTF-style activities have been adopted in recent years from the cybersecurity field to other education fields because of the versatility of its characteristics and its gamified environment [10]. In fact, CTF-style activities can provide an alternative way to evaluate students, as they can be used as a summative assessment, in a similar way as traditional written exams [11].

In this paper, a CTF-like activity has been designed in order to provide a self-study exam in a computer course at a college level, whose aim is to get learners trained for a traditional written exam to be carried out at a later stage. The main research goals in this paper are to first present a novel layout for self-study exams whose structure is inspired by the CTF competitions, as well as comparing the academic performance in a course dedicated to the introduction to computer science where an instance of this type of self-study exam was introduced with a similar course without it.

The remaining of this paper is organized as follows: section 2 displays the most common types of CTF, section 3 presents the method, section 4 displays

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

the results, section 5 presents the discussion, and finally, section 6 draws the final conclusions.

2. Types of cybersecurity CTF

Focusing on cybersecurity CTF competitions, it is to be noted that each type of CTF has its own characteristics, which allows to focus on specific skills and knowledge. In this sense, attack-defense CTF are high-intensity competitions where teams play the part of attackers and defenders at the same time, as if they were located in a real battleground. This way, a given team must be responsible of managing and protecting its own computing assets, although at the same time that team must also launch attacks against other teams' resources in order to capture their flags, which are often strings in alphanumeric format. In other words, each team is expected to challenge the rest of the teams by trying to exploit their services, and at the same time, they are supposed to secure their own assets.

Sticking to the cybersecurity domain, the attack part is carried out by trying to exploit vulnerabilities on the computers, networks, and services hosted by other teams in order to steal the flags stored inside. However, the defense part is undertaken by trying to patch vulnerabilities on the computers, networks, and services hosted by each particular team in order to avoid the rest of the teams stealing its own flags.

Another common CTF format is king of the hill, where the goal is to maintain control of computing assets for as long as possible. Hence, the goal in this type of competition is to defend against penetration testing, and in order to do so, participants need to consider some of the common trade-offs taken by real-world pen testers, such as which services are run or restricted, or the specific machine within the network employed to gain access to other resources.

A different CTF format is hack quest, which is usually referred to online platforms providing hacking-themed challenges based on the cybersecurity domain. In other words, hack quests are considered hacking missions where contestants need to fulfill a certain goal, which usually involves the capture of different flags. In order to meet the requirements proposed, participants must find vulnerabilities in any of the stages of the hacking process, such as reconnaissance, scanning, gaining access, maintaining access, or clearing track.

Additionally, another CTF format is jeopardy, where individual participants or teams have to solve a series of challenges classified in different domains, such as web security or forensics. In this context, each challenge solved accounts for a given point value, such that participants get points when they solve a challenge, where the winner is the participant getting the greater amount of points. This format provides flexibility, as challenges can be tackled in any order, which makes it an interesting learning environment for different skills. Additionally, this format can be exported towards other domains, such as any educational field. Furthermore, there might be some hints available, whose unveiling provokes the subtraction of certain amount of points out of the original points awarded to solve a given challenge.

The main challenges in any kind of CTF format go about testing the participants' skills related to coding, cybersecurity, and vulnerability research. This way, coding difficulties revolve around debugging source code, identifying logical errors, or trying to solve programming issues. In order to deal with this, the mastery of different programming language is required, as well as the ability to analyze source code under pressure.

Similarly, cybersecurity difficulties revolve about real-word security measures in the form of firewalls, encryption, or access controls to be bypassed. The typical challenges to be dealt with contain vulnerability detection, detection system evasion, and the ability to exploit flaws without triggering alarms. On the other hand, a typical scenario usually starts with a reconnaissance stage, where participants research the target in order to uncover details about the defense structure and gather intelligence so as to build up a potential attack plan.

Among the multiple benefits of taking part in CTF are skills development, versatility, accessibility, and teamwork. The first one is related to improve practical skills in a controlled environment, which tries to mirror real-life environments. The second one is related to the possibility of focusing on specific categories or a set of categories in each CTF. The third one is related to the use of this type of format for any sort of user, no matter if a given participant is a beginner or an expert in the field. And the fourth one is related to the possibility of taking the CTF as a team competition, which allows us to foster teamwork and collaboration as well.

Furthermore, challenges may be structured in layers, which force participants to circumvent different types of security controls. Those layers are preventive, detective, responsive, and recovery. An instance of the first one is to block access to a certain resource, whereas an instance of the second one is to monitor for intrusions, whilst an instance of the third one is to activate countermeasures, while an instance of the fourth one is to restore or mitigate loss.

Some examples of game challenges are the process of debugging faulty code in order to uncover secret access credentials, or the process of identifying vulnerabilities to bypass restrictions and also to unlock higher levels, or the process of outwitting automated security agents which adapt to the actions of the participants, or the process of using reconnaissance and social engineering to gain insider information.

Regarding the most effective strategies to face CTF competitions, those are careful observation and

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

logical problem-solving, along with preparation and adaptability. This way, attention should be paid to details, such that clues and information need to be examined closely in order to spot any hidden message or unusual pattern. Moreover, logical thinking and problem solving are key players, thus each challenge should be broken down into manageable parts, whereas multiple angles may need to be applied if the initial approach taken leads to a deadlock or to growing difficulties.

Besides, mental preparation and technical flexibility are to be used in order to raise success rates, whereas community knowledge should be leveraged through the read of guides, walkthroughs, or user tips so as to take advantage of other's experience, thus bypassing common mistakes. Also, collaboration and strategy sharing can be established through different forums in specific online platforms for knowledge sharing. Furthermore, adapting to the type of challenge found is crucial as it may help find the proper solution shortly. For instance, regarding coding problems, the fact of reviewing sample code of error messages on a systematic basis gets participants more used to those type of situations. On the other hand, regarding social engineering, the consideration of alternative views and paying special attention to context may make a difference, as well as carefully examining the instructions in order to catch possible narrative hints.

3. Method

We were running a college course about the introduction to computer science during the present academic year 2024-2025. The evaluation scheme included the delivery of a number of practical cases, where coding designs had to be developed, along with a final exam to be completed through a Learning Management System (LMS). Such an exam consisted on a series of multi-choice questions, where the correct answer had to be selected among a group of possible statements, along with short-answer questions, where the correct answer must be written. In the last academic year, students were unsure about the type of questions they would face during the final exam. Therefore, a CTF-like self-study exam was designed in order to help them get ready for it.

The structure of this self-study exam was composed of 10 questions, randomly drawn out of a pool of a thousand questions in order to minimize the chances for any pair of students to get the same questions. Each question was presented on a separate screen, such that an answer was required before passing to the next question. Regardless of the type of question, namely multi-choice or short-answer, three hints were available for each question in case an answer was not known by a student. However, those hints were initially hidden, although a student could unveil them in a sequential manner so as to get extra information about the question proposed. This way, the first clue was more generic, whilst the second one was slightly more specific, whereas the third one was more focused.

The points assigned for each question was 1 point for a correct-answer question without using any hint, or otherwise, 0 points for an incorrect-answer question regardless of the number of hints being used. However, each time a clue was unveiled, then 0.25 points were taken out of a possible correct-answer question. This way, if one hint was unveiled, then a correct-answer question was worth 0.75, whilst if two hints were unveiled, then a correct-answer question was worth 0.50, whereas if three hints were unveiled, then a correct-answer question was worth 0.25.

Regarding the evaluation system used, the assessment of multi-choice questions was straightforward, as if the amount of possible right answers was higher than one, then such an amount was quoted within each question. Hence, for multi-choice single answer (MCSA) question types, the way to get a correct-answer question was just to select the right answer. On the other hand, for multi-choice multiple answers (MCMA), the way to get a correct-answer question was to select all the right answers, or otherwise, it was considered as an incorrect-answer question.

With respect to short-answer questions, the assessment was done through cosine similarity. This way, the string with the right answer established by the teacher and the string with the answer given by a student were confronted for text matching by means of cosine similarity. The outcome ranged from 0 to 1, where the former implied no matching at all between both strings and 1 implied full matching between both strings. The threshold established was 0.70, which is quite commonly used in this kind of measure of similarity for any given pair of non-zero vectors. The reason for the popularity of this threshold is because the cosine of 45º is the square root of 2 divided by 2, which occurs to be approximately 0.707, and 45º is located just in the middle of the range of angles between 0º and 90º. It is to be noted that the cosine of 0º is 1, while the cosine of 90º is 0, which are basically the upper and lower bounds of the range defined in this case. Hence, if a cosine value is 0.7, then it can be said that it corresponds to an angle of nearly 45º. Therefore, for a short-answer question to be considered as correct-answer question, then the cosine similarity between the answer given by the student and the right answer established by the teacher must be at least 0.7. Obviously, this process is internally dealt with by the server where the application of self-study exams is hosted, which is also the case for the multi-choice questions. It is to be noted that this type of format applied to these self-study exams is typically used in CTF competitions. However, in that kind of event, the points obtained by the winner is commonly related to both the amount of time spent

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

completing a mission and the number of clues being unveiled during the path to do so. Regarding the time, the shorter, the better, thus the number of points assigned is higher as the shorter time is elapsed to get a mission done. With respect to the clues used, the lesser clues are used, the lesser points are subtracted from the missions accomplished, thus the number of points assigned is higher.

Focusing on the self-study exams proposed, in order to take into consideration both variables, namely the number of points obtained, and the time spent, a weighted score was put in place. The system of weights used was a variation of the Pareto principle, also known as 80/20 rule, such that 80% was assigned to the points achieved through the questions and the other 20% was achieved through the time elapsed in the self-study exam. On the one hand, the percentage assigned to the points achieved relates to the number of points scored during all the questions are accounted. On the other hand, the percentage assigned to the time units, or units, spent relates to the amount of time elapsed from the beginning of an exam until the exam completion.

It is to be remarked that the time units are assigned such that spending at most 10 minutes in the exam accounts for full marks, whilst 1 unit is subtracted per each extra minute, or any portion of a minute, spent in the exam after overcoming the threshold of 10 minutes. However, the outcome of this subtraction has a lower bound of zero, thus it will never go under zero in any case. The motivation for this grading system of units for the time elapsed is because it is commonly considered that 1 minute per question is a sufficient time in a test exam, hence as there are 10 questions, then 10 minutes should be a fair threshold.

For instance, if a self-study exam contains 10 questions and the score achieved is 7.5 points, then those points represent 80% of the final score, which accounts for $7.5 \times 80/100 = 6$ points in the weighted score. Furthermore, if the time elapsed in this exam is 8 minutes, then the score associated to the time spent is 10 units, which is the top mark as it is lower or equal than 10 minutes. Those 10 units represent 20% of the final score, which account for $10 \times 20/100 = 2$ points in the weighted score. Hence, the overall weighted score obtained is $6 + 2 = 8$ points, which accounts for the final score achieved.

Alternatively, another instance would be a self-study exam with also 10 questions, where the score attained is 5 points, which account for $5 \times 80/100 = 4$ points in the weighted score. Besides, if the time elapsed is 14 minutes and 30 seconds, which gets ceiled up to the nearest integer, namely 15 minutes, hence it exceeds to threshold of 10 minutes to get the full exam done. In this case, a penalization of 1 unit per minute is applied after the 10th minute, such that the penalization is subtracted from 10 units. This way, the units attained due to the elapsed time yield $10 – (15 – 10) = 10 – 5 = 5$ units due to the elapsed time. These 5 units represent 20% of the final score, which account for $5 \times 20/100 = 1$ point in the weighted score. Therefore, the overall weighted score achieved is $4 + 1 = 5$ points, which accounts for the final score obtained.

4. Results

The course devoted to the introduction to computer science was run in the present academic year, namely 2024-2025, and it was attended by 25 learners, where a self-study exam was implemented as a CTF-style activity. On the other hand, in the previous academic year, namely 2023-2024, the course was also run and it was also attended by 25 learners, although there was not any self-study exam.

The next tables show the relevant statistical measurements obtained out of the results related to the final exam in both academic years. Hence, the scores related to the final exam of the current academic year got influenced by the extra training taken by the students for preparing and sitting the CTF-style self-study exam.

Furthermore, as the college is based in Spain, the Spanish grading system has been used in order to evaluate students. This scoring system ranges from 0 all the way to 10, while a score of 5 represents the passing score [12].

Table 1 displays a summary of the most relevant measurements regarding descriptive statistics. The first one is the sample size, which accounts for the number of learners in each of the courses delivered. The second one is the mean, which stands for the average score obtained by all students in each course, thus being a centralization measurement. The third one is the standard deviation, which stands for the dispersion of the scores corresponding to each course with respect to its average, thus being a dispersion measurement [13].

Table 1. Relevant measurements regarding descriptive statistics

	N	Mean	Standard deviation
Past academic year	25	6.12	1.82
Current academic year	25	6.75	1.75

Table 2 exhibits the steps to calculate the two-sided t-test at the 95% confidence interval in order to find out if the differences in the distribution of scores regarding each academic year are statistically

Table 2. T-test applied to both distributions of scores

t-student test value	1.25
degrees of freedom (df)	25+25-2 = 48
p-value	0.22
s-value	0.05
Interpretation of the p-value	Not significant

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

significant or otherwise [14]. The t-test is also known as significance test and is a relevant part of inference statistics.

Table 3 shows the effect size achieved with the values of the means and standard deviations related to the distribution of scores corresponding to each academic year [15]. The target is to measure the strength of the statistical claim due to the difference between groups in inferential statistics.

Table 3. Effect size achieved with both distributions of scores

Difference of means	0.63
Pooled standard deviation	1.79
Cohen's d (effect size)	0.35
Interpretation of Cohen's d	Small to medium

Table 4 exposes the sample size needed to get the effect size achieved with the distribution of scores corresponding to each academic year [16]. The target is to minimize possible errors in inferential statistics.

Table 4. Sample size needed to get the effect size achieved with both distributions of scores

Cohen's d	0.35
Alpha value	0.05
Power	0.80
Sample size needed	127
Sample size achieved	25
Interpretation of sample size	Insufficient

Table 5 compares the value of the sample size needed obtained through different methods. In this sense, the values chosen for alpha and power have been the most commonly used in the literature, such that former is 0.05 and latter is 0.80, respectively [17]. Furthermore, in order to make fairer comparisons among those methods, the effect size is going to be considered with up to four decimal digits, thus being 0.3529.

Table 5. Sample size needed obtained through different methods

Mathematical expression	127
Online sample size calculator	127
Table with effect size -vs- sample size	129
Lehr's rule of thumb	128

5. Discussion

With respect to Table 1, it shows that the mean obtained in the current academic year, namely 6.75, is higher than the mean attained in the last academic year, namely 6.12, thus accounting for a rise slightly higher than 10%. Hence, the academic performance related to the overall students has improved in the current academic year. On the other hand, the number of attendants in both courses was the same, whereas the difference in the standard deviations was around a 4% lower in the current academic year. Therefore, the rise in the mean lead to think that the establishment of the CTF-style self-study exam may well be the responsible for this enhancement, as students got better prepared for the final exam.

With regards to Table 2, it exhibits a value of 1.25 for the t-student value of 1.25 and a value of 48 for the degree of freedom, which leads to a p-value of 0.22. However, this value is higher than s-value chosen, which is 0.05, hence it can be concluded that there is not a statistically significant difference between the distribution of marks attained in both academic years. Therefore, it cannot be concluded that the improvement in the scores was due to the effect of the self-study exams, as that enhancement could have been due to chance.

Regarding Table 3, the difference of means obtained is 0.63, whereas the pooled standard deviation is 1.79, which actually was the average spread of scores around their group mean. The division of both values leads to calculating Cohen's d, which accounts for the standardized effect size. The value obtained was 0.35, thus it was right in the middle of the interval with a lower bound of 0.20, which accounts for small effect, and an upper bound of 0.50, which stands for medium effect. It is to be noted that a value of 0.50 represents an effect visible to the naked eye, considering a careful observer. Moreover, it is to be highlighted that the lower the effect size, the greater the sample size is needed.

With respect to Table 4, the Cohen's d is 0.35, as stated in the previous table, whereas the alpha value taken is 0.05 and the value for statistical power is 0.80. With those values, the sample size needed is 127, which is higher than the sample size achieved, namely the number of attendants in each course, which are 25 students. Hence, this sample size is insufficient for the effect size obtained, namely 0.35, which means that a greater number of students should be considered in order to state that the effect size achieved was due to the implementation of the self-study exam.

Regarding Table 5, the four methods exposed present very similar results, although the mathematical expression is meant to be most precise one. In the first place, the outcome was obtained through a specific mathematical expression written in a spreadsheet [18], whose result was the one shown in the previous table. In the second place, the outcome was attained through an online sample size calculator [19], which happens to also carry out effect size calculations.

In the third place, the outcome was achieved through a table showing the relationship between effect size, namely the independent variable x, and sample size, namely the dependent variable y [20]. In order to obtain a more precise value for y, a trend line has been found out with the data pairs involved, namely x and y, leading to the following equation:

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

$y=16.755x^{-1.963}$, which allowed calculating the outcome shown in the table. And in the fourth place, the outcome was obtained through Lehr's rule of thumb [21], which is a shortcut offering a quick and close approximation to the sample size.

In summary, the implementation of the CTF-style self-study exam in order for students to better prepare for their final exam improved their scores by around 10%. However, this improvement is not statistically significant, which means that it cannot be concluded that such an improvement was due to the implementation of the self-study exam, as it could have been due to chance. Likewise, the standardized effect size achieved of 0.35 requires a sample size of 127, whereas the actual sample size was 25. Hence, a larger number of students should be included in further research in order to conclude that the implementation of the self-study exam led to the effect size achieved.

Looking at future courses, and focusing on the scoring model used, namely the weighed score, the system of weights put in place has been considered as a reasonable balance between test results and time spent. However, further research might lead to better adjust the 80/20 rule established between the outcome of the exam and the time spent to get it completed in order to give more weight to either the points achieved in the exam, or otherwise, the units associated with the time elapsed.

On the other hand, it is to be noted that a weighted score could be considered as a normalized score, as two variables, namely the test outcome and the time spent, are used to calculate the final score. However, other options for scoring are available. Further research might lead to the convenience of changing the scoring method.

One option is a direct score, which just takes into account the points achieved in the exam, whereas time is not considered. However, if the time limit is reached, then the exams not yet submitted are automatically submitted at that moment.

Another option is learning curve score, which assigns a determined time to each question according to its difficulty, as opposed to assigning a global time for the overall exam. In this case, if the deadline for a given question is reached, then this question is considered as unanswered, and the following question is presented to the student.

A further option is applying the Rasch scoring, which depends on both the student's knowledge and the item difficulty, where the probability of getting a correct answer is modeled as a logistic function [22]. This method is considered as more objective, although the key point is the difficulty in establishing the statistical model to estimate the values for the ability of a person and the difficulty of an item.

Furthermore, the importance of a self-study exam in future courses could be improved by assigning some weight in the course final score. For instance, a certain percentage of the course overall evaluation score may be assigned to the self-study exam in order for learners to take it more seriously. This percentage could be something like 5% or 10%.

Besides, another improvement could be to penalize the wrong answers, which is not done in the model proposed herein. A typical penalization in multi-choice questions is to subtract the inverse of the predecessor of the number of choices available per each incorrect-answer question, so this could be applied in future editions in order to restrain random answers. For instance, in a multi-choice test made of questions with 4 possible answers, this penalization system subtracts 1/3 points per each wrong-answer question out of the overall correct-answer questions.

Additionally, another enhancement could be the chance for students to take a self-study exam more than once in order for them to get acquainted with the layout at their first attempt, such that they may improve their performance in successive attempts. Nonetheless, in order to avoid taking the exam too many times, a maximum number of attempts may need to be established, such as just 3.

Moreover, in those circumstances, the fairest selection of the final score assigned to the self-study exam may be the score at the last attempt, even though another option would be to stick to the highest score achieved in any attempt.

All those possible changes listed can be grouped into four different categories, where the option proposed can be swapped with the actual option implemented in the CTF-based self-exam proposed.

In this sense, the scoring system can be considered as one of those categories, where the direct score, the learning curve score, and the Rasch score would be included, along with the weighted scoring system based on the 80/20 rule. This way, the scoring system for the next course could be any of those, or otherwise, it could be the weighted scoring system with different weights. Hence, there are plenty of options to take in future editions, even though the one implemented in this course gets a good balance between the clarity of the scoring, the fairness of the scoring system, and the difficulty to get it implemented.

On the other hand, the percentage of the CTF-based self-study exam in the final score of the course can be seen as another category, as the importance assigned by the students to this activity will rise as it increases the percentage in the final score of the course assigned to it. Additionally, it is to be reminded that the CTF-based self-study exam proposed is taken at home, thus the extra pressure of taking a final exam at school is not present. Furthermore, another possible improvement could be to set the final exam as a CTF-based exam. This would actually be the long-term goal of this proposal, even though some more tests need to be undertaken in order to get it done.

Another category could be the penalization system applied to wrong answers in the exam. A

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

common penalization scheme has been described above as an improvement, which it is directly related to the number of possible options in a multichoice question, thus making it a fairer system. However, other penalization schemes could be worked out for multichoice questions, such as a more restrictive one where a wrong answer is compensated by a right answer, or a looser one where a sort of a waiting period could be deployed, such that a certain number of wrong answers are allowed without getting penalized. Likewise, similar penalization systems could be implemented in short-answer questions in order to restrict the random answers if participants do not know or are not sure of the correct answer to certain questions.

The last category could be the scoring attempt, as it would influence the number of attempts taken by a given student in a given exam. In this edition, there was no discussion about it as there was just one attempt to take the activity. However, more than one attempt can be granted, which may involve different strategies in order to calculate the score obtained. In those cases, a common scoring strategy is to select the highest score among all attempts taken, which is the best situation for students, as they do not have to worry about bad performance in future attempts. Another common strategy is to take the last attempt, which might seem a bit fairer, as if a student takes the chance to get a higher mark, then this student must be conscious that underperformance has negative consequences.

Another strategy is to find out the average score of all attempts taken by a given student. This way, students would get motivated to make more attempts if the first attempts got low performance, thus forcing learners to study harder to get higher marks. A less common strategy is to stick to the intermediate score obtained in different attempts, such that the highest and lowest attempts are discarded, which might seem a fair trade-off. Another strategy is to get the average of the two highest scores, such that the best attempts achieved by each student are considered when it comes to establishing the score. Any of those scoring strategies can be taken into consideration for further editions of this course.

6. Conclusions

In this paper, a CTF-style self-study exam has been proposed in order for students to get better prepared to the final exam of a college course devoted to the introduction to computer science. The self-study exam was composed of 10 questions, whose answers could be either multiple-choice or short answer. Each of those questions was worth 1 point, hence the highest score obtained was 10 points. Furthermore, the format of the exam followed that of a jeopardy CTF, where questions were presented to the participants and some hints were offered per each question with different clues in order to help then find the correct answer.

A weighted score system was put in place, where a variation of the Pareto principle was used. In fact, the score achieved with the questions was worth 80%, whereas the time spent was worth the other 20%. Regarding the former, the score achieved for each question depended on the number of hints unveiled before answering, such as a correct response without unveiling any hint accounted for 100% of the value of that question, whilst a correct response where one hint was unveiled accounted for 75%, a correct response where two hints were unveiled stood for 50%, a correct response where three hints were unveiled was worth 25%, and obviously, a wrong response was worth nothing.

Regarding the latter, the deadline time was considered as 10 minutes, as it accounts for 1 minute per each of the 10 questions proposed. Hence, if a students completed an exam in less than 10 minutes, then 10 time units, namely units, were assigned, whereas 1 unit was taken per each minute after the deadline time, which also applies for any fraction of a minute left. However, a lower bound of zero was set up for the scoring related to time, such that it could not detract any points from the those obtained with the correct answers to the questions.

With respect to the results, the mean score in the current academic year was 6.75, where the self-study exam was set up, while the mean score in the last academic year was 6.12, where no self-study exam was found. Hence, an improvement of about 10% was experienced with the implementation of the self-study exam, thus from the point of view of descriptive statistics, it seems to be an good result.

However, a t-test was performed, leading to a p-value of 0.22, which is higher than the s-value selected, namely 0.05. Hence, this improvement cannot be seen as statistically significant from the point of view of inferential statistics, thus meaning that the enhancement could be due to chance.

Furthermore, the effect size achieved was 0.35, which accounts for a value in the middle of the interval defined by 0.20 as the lower bound, which accounts for small effect, and 0.50 as the upper bound, which stands for medium effect. Additionally, the sample size considered was 25 students, although the sample size needed was 127, considering the effect size, the alpha value of 0.05 and the statistical power of 0.80. Hence, the sample size used was insufficient for the effect size achieved. Therefore, further research must be carried out with a larger amount of students in order to account for the effect size achieved.

Moreover, some improvements could be applied regarding four categories in future editions of the course. Those categories include the scoring system used in the activity, the percentage of the activity within the final scoring of the course, the penalization

 | DOI: 10.20533/ijicr.2042.4655.2025.0163

system applied to wrong answers, and the scoring strategy in order to deal with more than one attempt in the activity.

7. References

[1] Karagiannis, S., and Magkos, E. (2020). Adapting CTF challenges into virtual cybersecurity learning environments. Information and Computer Security, 29(1), 105-132.

[2] Mansurov, A. (2016). A CTF-Based Approach in Information Security Education: An Extracurricular Activity in Teaching Students at Altai State University, Russia. Modern Applied Sciences, 10(11), 0159.

[3] Collins, J., Ford, V. (2023). Teaching by Practice: Shaping Secure Coding Mentalities through Cybersecurity CTFs. Journal of Cybersecurity Education, Research and Practice, 2, 09.

[4] Zafar, A., Yamin, M. M., Katt, B., Torseth, E. (2024). All flags are not created equal: A deep look into CTF Scoring Algorithms. Expert Systems and Applications, 254, 124436.

[5] Ford, V., Siraj, A., Haynes, A., Brown, E. (2017). Capture the Flag Unplugged: An Offline Cyber Competition. In: Proceedings of SIGCSE. March 8-11, Seattle (WA), United States.

[6] Minagar, S., Sakzad, A. (2023). Automatic Problem Generation for CTF-Style Assessments in IT Forensics Courses. In: Proceedings of ITiCSE 2023, July 7-12, 2023, Turku, Finland.

[7] Lagorio, G., Ribaudo, M., Armando, A. (2021). Capture the Flag competitions for Higher Education. In: Proceedings of ITASEC 2021, April 7-9, 2021, Rome, Italy.

[8] Beauchamp, C., Matusovich, H. (2024). A Mixed-Method Study Exploring Student Motivation for Participating in Cybersecurity CTF Competitions. Cybersecurity Pedagogy and Practice Journal, 3(1), 4-26.

[9] Nelson, C., Shoshitaishvili, Y. (2024). PWN The Learning Curve: Education-First CTF Challenges. In: Proceedings of SIGCSE 2024, March 20-23, 2024, Portland (OR), United States.

[10] Cole, S. V. (2022). Impact of Capture The Flag (CTF)-style vs. Traditional Exercises in an Introductory Computer Security Class. In: Proceedings of ITiSCE, July 8-13, Dublin, Ireland.

[11] Meinsma, D., Barjas, S., Gilhespy, M., Björkqvist, M. (2022). Effectiveness of CTF education. Measuring the learning outcomes of Jeopardy CTF's. The Hague University of Applied Sciences, The Netherlands.

[12] Polytechnic University of Valencia. Spanish grading system, https://www.upv.es/entidades/OPII/infoweb/pi/info/1147768normali.html (Access Date: 28 February 2025).

[13] Roig, P. J., Alcaraz, S., Gilly, K., Bernad, C., Juiz, C. (n.d.). Design and Assessment of an Active Learning-Based Seminar. Education Sciences, 14(4), 0371.

[14] Roig, P. J., Alcaraz, S., Gilly, K., Bernad, C., Juiz, C. (n.d.). A Generic Multilevel Structure for Educational Escape Rooms. Education Sciences, 14(5), 0488.

[15] Schäfer, T., Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. Frontiers in Psychology, 10, 0813.

[16] Psyteachr Github. (n.d.). Power and Effect Sizes. https://psyteachr.github.io/quant-fun-v2/power-and-effect-sizes.html (Access Date: February 28, 2025).

[17] Praharaj, S. K., Ameen, S. (2024). Sample size estimation in research: necessity or compromise? Kerala Journal of Psychiatry, 37(1), 66-71.

[18] Sommet, N., Weissman, D. L., Cheutin, N., Elliot, A. J. (2023). How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. Advances in Methods and Practices in Psychological Science, 6(3).

[19] Highyieldmed. (n.d.). Sample size calculation. https://www.highyieldmed.org/sample-size-calculation-2/ (Access Date: 28 February 2025).

[20] Jackauty. (n.d.). Power calculation Cohen's d. https://jackauty.com/power-calculator-cohens-d/ (Access Date: 28 February 2025).

[21] McConnell, M. M., Monteiro, S., Bryson, G. L. (2019). Sample size calculations for educational interventions: principles and methods. Canadian Journal of Anesthesia, 66, 864-873.

[22] Verdú-Soriano, J., González-de la Torre, H. (2024). Rasch analysis implementation in nursing research: A methodological approach. Enfermería Clínica, 34, 493-506.

 | DOI: 10.20533/ijicr.2042.4655.2025.0163