The use of gamification to support the teaching-learning of software exploratory testing: an experience report based on the application of a framework

Igor Ernesto Ferreira Costa
Graduate Program in Computer Sciense (PPGCC), Institute of Exact and Natural Sciences (ICEN)
Federal University of Pará (UFPA)
Belém – PA, Brazil
igor.ef.costa@gmail.com

Sandro Ronaldo Bezerra Oliveira
Graduate Program in Computer Sciense (PPGCC), Institute of Exact and Natural Sciences (ICEN)
Federal University of Pará (UFPA)
Belém – PA, Brazil
srbo@ufpa.br

Abstract— **This Research to Practice Full Paper presents exploratory testing, an approach that has become quite relevant in the testing software, growing its application in the industrial scenario. The main reason for that is the emerging utilization of agile practices in the software development process to satisfy the needs of the market (Time to Market), which is essential for a company to remain active in the market. However, the systematic mapping study performed found few studies about the application of exploratory testing, a subject little discussed in the academic context. For this reason, this work uses gamification as a systematic strategy in exploratory test teaching and learning in the form of an experiment with two classes. One class has undergraduate students in Computer Science, and the other has students who graduated as Computer Technicians. The aim is to engage students to obtain a better performance, preparing them to use that test approach in the industrial and academic context. As a result, students achieved good overall performance; with reports from students that gamification facilitated and significantly collaborated to achieve a better performance converging with the quantitative data obtained. This can be evidenced mainly by the fact that both runs of the experiment (classes) reached a percentage higher than 70% of achievement, a great overall performance analyzed by the ratio of medals obtained by participation.**

Keywords—gamification, teaching-learning, exploratory testing, software testing

I. Introduction

One of the great challenges faced in the teaching of Software Engineering is to supply the need to use teaching methods that make this process more effective [1]. In these circumstances, several studies have been carried out on software testing, mainly in the applicability of agile methods and systematic approaches to teaching this subject.

Exploratory Testing (ET) is a manual testing approach that emphasizes the tester's responsibility and freedom to explore the system, allowing the tester to gain knowledge of the program while performing the tests, as test cases are not pre-established in a test plan [2] [3] [4] [5].

The ET is flexible and promotes fast feedback. However, the scarcity in the generation of documentation has led to the emergence of test management techniques aimed mainly at structuring this approach [4] [6]. It was observed that the Session-Based Test Management (SBTM) technique is the most widespread among the other techniques, as was evidenced in the performance of a systematic mapping of the literature (SML) on the degree of importance of ET [7]. Thus, this work sought to integrate the SBTM in the application of ET.

In the context of education, the use of gamification allows students to visualize the effects of their actions, their performance in learning and how it happens progressively, becoming a facilitator in the relationship between the parties involved in the practice of teaching, immersed as in a game [8]. Thus, it is defined that gamification is the use of elements of games outside their context, where it is used to mobilize individuals to act, help, solve problems, interact and promote learning [9] [10].

Therefore, this work aims to apply a systematic strategy for teaching ET by using gamification to provide students with better knowledge about the applicability of this approach and enable them to meet market needs. Besides, this present work also seeks to collaborate for future research on this subject. Thus, this study has the following research question: *Does the use of gamification assist in the teaching and learning of the participants in ET?*

In addition to this introductory section, Section II presents the related works, Section III presents the gamified framework, Section IV presents the experiment performed, Section V shows the results obtained, and in Section VI the threats to the validity of this study are discussed. Finally, in Section VII the conclusion of this study is presented.

II. Related Works

Initially, a search was carried out in the specialized literature on works that use gamification to support the teaching of ET. However, no work was found, so the related works cover test teaching in general, showing the importance, relevance and originality of this study.

Herbert [11] presents four software testing teaching standards for professionals who do not develop software. These standards were extracted from the experiences in the test course for undergraduate students at the Federal University of Health Sciences of Porto Alegre in Brazil. The author mentioned above applied risk-based functional testing since the individuals had extensive knowledge of the domain of the system under test. However, the author does not report details of the approach used in teaching, and the standards emphasize more the description of concepts and good testing practices.

Valle, Barbosa and Maldonado [12] conducted an SML to identify the approaches that assist in the teaching of test.

The results indicate more instances of research on the teaching of test with programming and the use of educational games, focusing mainly on the test case design phase. However, few studies were observed, and the majority presented partial results.

Ribeiro and Paiva [13] present an educational game for the learning of software testing. ILearTest is specifically aimed at assisting professionals who aim to obtain ISTQB (International Software Testing Qualifications Board) certification. However, all the content covered is aimed only at the base level (Foundation Level), based on Syllabus.

As can be seen, no work presented addresses the practice of teaching a test using any strategy with elements of gamification. In this context, the present work is different in that it presents a systematic strategy with the use of many playful elements in order to facilitate, improve engagement, minimize the gaps between students and, mainly, boost teaching about ET in the academic context.

III. Research Methodology

The research followed the flow described in Fig. 1. Initially, an SML was conducted to identify the degree of importance of the exploratory test applicability. Then, the game elements (Fig. 2) to be used were defined based on the Octalysis Gamification Framework [14], and then the design of the experiment occurred. In parallel, the content of the training classes was defined based on an adaptation of the content presented in Chapter 4 of SWEBOK Version 3.0 [15] [16] [5], preparation of class presentations, construction of materials to be used in each stage and the elaboration of questions to be discussed at the feedback moment. After that, the data was collected during the experiment's execution, analyzed quantitatively and qualitatively, and described in a final report.

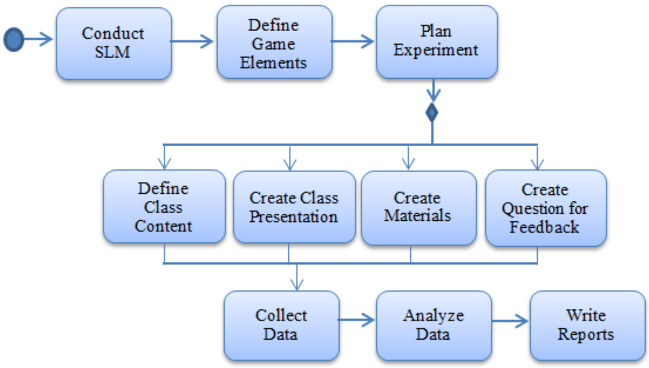

Fig. 1. The flow of research methodology.

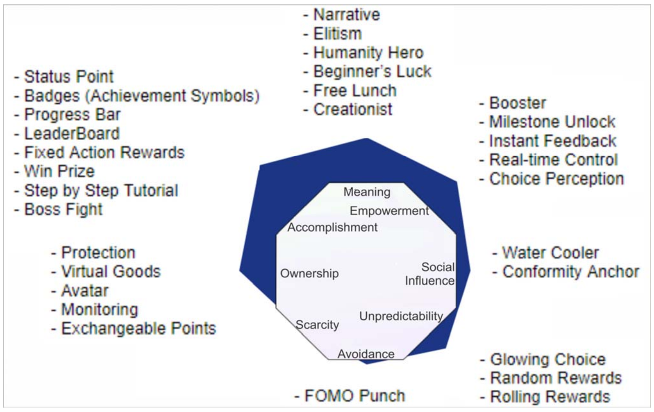

Fig. 2. Mapping of the game elements.

Fig. 2 shows the Core Drivers on the inside of the octagon and the corresponding elements on the outside. In this case, the blue margin is proportional to the number of elements used in each Core Driver.

IV. The Gamified Framework

In the perspective of promoting motivational exploratory test teaching and learning capable of providing students with better knowledge about the applicability of this test approach, as well as enabling them to meet market needs, gamification was adopted to assist in engaging students. Besides, it stands out that the SBTM technique was used to provide a more systematic approach to ET [17].

A. Overview

In gamification, several elements were used alluding to the game of treasure hunt (Fig. 3), as well as the characters were based on the "Pirates of the Caribbean" movie. After Fig. 3, each stage in the experiment flow will be presented. There are three profiles, which are: (a) Specialist, which is the minister and instructor in the experiment, in which has extensive knowledge of the software testing area, (b) Judge, which observes students actions and fills in the Ranking / Gamification worksheet, (c) Testers, which are the students, who made themselves available to participate spontaneously in the experiment [16].

In addition to the profile, each tester receives a fictitious name referring to a pirate who existed in history. In this context, some basic elements involved in the gamified framework are described: (a) Activity Points, rewards granted according to the performance in carrying out the activities, which are specific to each stage, (b) Medals, rewards awarded according to performance in the classroom, (c) Gifts, which are rewards for the personal benefit of the tester, who does not grant advantages in the dynamics, (d) Bitskull coins, virtual coins used to purchase cards, (e) Letters, playful resources used to personalize the tester's avatar, (f) Defect catalog, spreadsheet containing details of defects previously entered on purpose, (g) Bonus points, rewards awarded according to the behavior in the classroom [16].

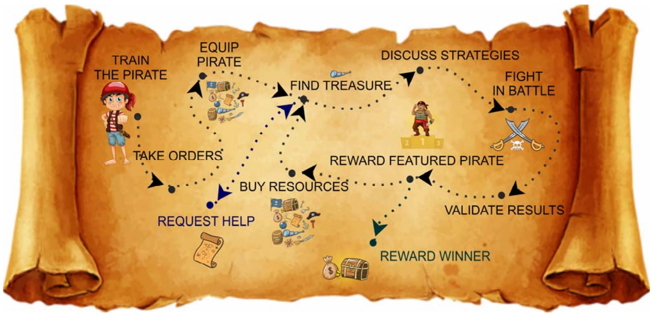

Fig. 3. The Experiment flow.

"Train the Pirate" stage, which serves as training for testers (introductory classes), aims at applying theoretical knowledge in later practical activities. Before starting this step, students filled out a form containing questions about concepts of the test approach in order to collect data to identify the degree of initial knowledge on the subject.

"Take Orders", in this stage the specialist explained the purposes and rules of each stage and the corresponding activities and delivered the materials to be used, thus elucidating any doubt of use.

"Equip Pirate", in this stage each tester selects accessories to personalize their avatar. Resources are obtained by choosing three cards at random. These cards provide resources for attack, defense and clothing, which are related to the world of pirates.

"Find Treasure", in this stage the test session takes place, where testers remain focused on detecting defects, prioritizing them as they register in the test session report.

"Request Help", if testers found it difficult to detect defects ("Find Treasure" stage) there is the possibility to request help, granting the option to consult a system requirement, the description of a feature in the user manual, description of an exploration strategy or a hint of a defect that was purposely inserted by the specialist. It is an optional step, and is inherent to the "Find Treasure" stage.

"Discuss Strategies", in this stage the testers performed the orthographic review of the defects register and self-analysis of the strategies used according to the PROOF technique (Past, Results, Outlook, Obstacles, Feelings). It is a step inherent to the SBTM that has been adapted similar to Debriefings.

"Fight in Battle", the testers exchanged the test session report to carry out a critical analysis of the defects recorded by the opponent. In this analysis, the following aspects are verified: (a) prioritization, identify and justify in the analysis report the reason for disagreeing with the priority level of each registered defect, (b) clarity, analyze whether the script for each registered defect is well inscribed, that is, there are no ambiguous words, wrong words and incomplete sentences, (c) reproducible, check if it is possible to reproduce the defect only with the registered script of each defect, otherwise, the participant highlights the inconsistencies found in the described script. Also, the tester proposes a general score between 0 to 10 on the analyzed test session report, as well as describing a justification for the specialist to subsequently assess the analysis in question.

"Validate Results", it is a stage practically performed only by the specialist, as he/she carries out his/her analysis of each session report, observing the three aspects mentioned above, comparing with the reports in the analysis report, as well as evaluating whether the note and justification were consistent. The specialist also checks whether the defects found are in the defect catalog or not, if present, the participant receives a certain score, otherwise, he / she receives a higher score, precisely because he / she is a native defect of the program and thus collaborates with the quality of the system. Therefore, testers get involved when the results are presented in the classroom, with a discussion around the positive and negative points, in addition to highlighting the points for improvement aiming at improving the exploration and data recording strategies in the aforementioned reports. In addition, there is a ludic activity in which students "fight" each other using the resources of attack and defense obtained in order to contain the opponent's progress or prevent attacks that cause the loss of points.

"Reward Featured Pirate", in this stage the specialist awarded the testers who achieved great performance in the activities related to the "Find Treasure" and "Fight in Battle" stages, and the gifts were only awarded if the puzzles were solved correctly. Such enigmas are theoretical questions about the content taught in the introdutory classes, aiming to fix the concepts.

"Buy Resources", a resource acquisition stage occurs again, but now testers also have the possibility of "unknown" type letters, which grant Bitskull coins, gifts or avatar customization accessories. This stage aims to prepare again for a new detection of defects and later "fight".

"Reward Winner", at the end, the winners are awarded who achieved great performance reaching level 4 in the general experiment. Since inside the classroom, there is also a hidden treasure (chocolate bonbon) that these testers needed to find following the resolution of three puzzles that lead to the location of that treasure. This activity occurs in an attempt to allow the closure of the dynamics of the experiment in an even more playful way.

Therefore, in [18] it is possible to observe how each game element was defined, materials used in each stage, scoring rules, avatar generation rules and more details about the gamified framework at work.

B. Rules and Awards

The rules were defined around the performance of specific activities at each stage of the experiment flow and the students' behavior. Regarding the first factor, students could receive from 0 to 10 points per activity. In the "Find Treasure" stage, the points awarded are conditioned to the number of detected defects, which can be negative, if the registered defect is a false positive, that is, -1000 points are penalized if the student registers a normal flow of the system as if it were an abnormal flow (bug). In relation to the second evaluation factor, the questions asked, presence in the classroom, suggestions, and participation were observed as aspects that granted bonus points, in contrast to absence from classes, hinder the performance of activities and if not perform the activities in question result in the loss of bonus points. In addition to this, there are also rules related to control, such as fulfilling the time allotted for a task and solving puzzles that are prerequisites for receiving some gifts.

For each stage, the student can reach level 1 to level 4 both for the performance factor in carrying out the activities and their performance concerning the assessment items on behavior (participatory actions) in the classroom. The four possible levels (avatar) are: Level 1 corresponds to the Marty avatar, Level 2 corresponds to the Will Turner avatar, Level 3 corresponds to the Joshamee Gibbs avatar and Level 4 refers to the Jack Sparrow avatar.

From the performance in the two evaluative factors, a final avatar for each stage is generated, then the general avatar is created, which is correlated to the student's general performance in the experiment, by calculating the arithmetic mean considering all the stages performed. In this case, the weight score of the avatars proportional to the levels is considered, being weight 1 for level 1, weight 2 for level 2, and so on. Thus, the general avatar based on the result of the arithmetic mean (AM) calculation can be: (a) Level 1, if AM $\leqslant 1$, (b) Level 2, if $1 > \text{AM} \leqslant 2$, (c) Level 3, if $2 > \text{AM} \leqslant 3$, and (d) Level 4, if $3 > \text{AM} \leqslant 4$.

In general, prizes are awarded in medals and coins, which depend on the performance of participatory actions. There is also the concession of gifts, which are prizes that do not grant any advantages in the dynamics, it is just an extra benefit in the game, for example, a book, candy and chocolate bar. Therefore, it is possible to consult more details in [18].

V. THE EXPERIMENT

This section presents details about the experimental study carried out to evaluate the gamified framework. The experiment was performed twice aiming to evaluate whether the gamified framework could assist in teaching and learning about ET, observing its feasibility at the technical and undergraduate level. In addition, based on feedback from participants, the framework has been refined (minimizing gaps in the first execution) to improve its applicability.

A. *Application Environment*

I. First Execution: a computer lab was used, where the students involved in the experiment were three people who were part of a test team, who used this exploratory test approach in a system development project for the Institute of Exact and Natural Sciences, from a Federal University in Brazil.

II. *Second Execution:* it also took place in a computer lab, with a group of six students who had recently graduated in Computer Technician from the Federal Institute in Brazil, in this case, they were students with technical graduation.

B. *Execution of the Experiment*

Tables I and II show the stages that took place on each day of the experiment. Beforehand, it is highlighted that the feedback from students in the first run reported the need for some adjustments, causing the insertion of a day in the second execution.

These two groups were selected to verify whether the referred framework could also be viable in technical schools in Brazil (Federal Institute of Education, Science and Technology - IFET). In this context, as there is a training stage aiming to balance the participants' knowledge about the referred approach, there is the prospect of obtaining results that encourage its viability in both teaching scenarios.

TABLE I. STAGE BY DAY IN THE FIRST EXECUTION

Day	Stage
1st 2nd 3rd 4th	Train the Pirate.
5th	Take Orders; Equip Pirate; Find Treasure; Request Help; Discuss Strategies; Fight in Battle.
6th	Validate Results; Reward Featured Pirate; Buy Resources; Find Treasure; Request Help; Discuss Strategies; Fight in Battle.
7th	Validate Results; Reward Featured Pirate; Reward Winner; Feedback.

TABLE II. STAGE BY DAY IN THE SECOND EXECUTION

Day	Stage
1st 2nd 3rd 4th	Train the Pirate
5th	Take Orders; Equip Pirate; Find Treasure; Request Help; Discuss Strategies; Fight in Battle.
6th	Validate Results; Reward Featured Pirate; Buy Resources; Find Treasure; Request Help; Discuss Strategies; Fight in Battle.
7th	Validate Results; Reward Featured Pirate; Buy Resources; Find Treasure; Request Help; Discuss Strategies; Fight in Battle.
8th	Validate Results; Reward Featured Pirate; Reward Winner; Feedback.

In the two executions on the 4th day, only a traditional test was applied containing objective and subjective questions to assess learning in training. The experiment starts in the “Train the Pirate” stage and ends in the “Reward Winner” stage (Fig. 3), however there were some iterations where it starts in the “Find Treasure” stage until “Reward Featured Pirate” stage.

C. *Software Tools*

Basically, the Google tools were used, being Google Forms to collect data on students' prior knowledge, Google Sheets to prepare the Ranking / Gamification spreadsheet, to extract and store quantitative data and other necessary materials. To create the traditional test and fixation exercises, Microsoft Office Word was used. In relation to qualitative data, a program installed on a mobile device that was recorded in the form of audio was used.

It is noteworthy that the system under test (SUT) was related to the context of generating tests or simulating online with objective and / or subjective questions, whose name is SAW. The system provides two user profiles, one being a “teacher” type, who had access to all features, and another “student” type, which only had access to some features. The SAW allows the “teacher” profile to also generate reports on the answers to the referred tests or simulations [19].

VI. RESULTS AND DISCUSSIONS

This section presents and discusses the quantitative and qualitative results obtained from the experiment. It is emphasized that the experiment was applied twice to better observe the efficiency of the framework from the comparative analysis of the data collected in the two executions involving students at a different level of training.

Regarding the characteristics of the participants, the average age was 21 years ago. The gender (Male - M or Female - F) and the age of the respective students are: Cofresí (F, 23), Henry (M, 22), Anne (M, 21), Kidd (F, 19), Bart (M, 20), Edward (M, 19), Mohamed (F, 27), Felix (M, 19) and Kotaro (F, 23). Those involved participated freely and voluntarily, without offering any benefit to participate. The only requirement was that both groups should be less in the last semester of their course, in order to ensure that they obtained the minimum technical knowledge about computers.

A. *Quantitative Results*

A relationship between grade and concept was elaborated to understand the performance of students in training activities (“Train the Pirate” stage). In this case, a score less than 5 is an Insufficient concept, from 5 to 6.9 is a Regular concept, from 7 to 8.9 is a Good concept and from 9 to 10 is an Excellent concept. There was also the application of the initial form where it was identified that no participant in the two executions of the experiment had previous knowledge about software testing, much less about the exploratory test approach [20].

In the first execution, the three students achieved a good performance in the exercises, reaching the maximum grade. In the traditional test, Anne, Cofresí and Henry obtained grades 8, 6 and 5, respectively, however obtaining the Regular grade by Cofresí and Henry, can be justified because they are absent in some classes. The three students mentioned above did not get a better concept, as they confused what would be the stages of the testing process with the stages of the V&V (Verification and Validation) process [20].

In the second performance, all six students also performed well in the exercises, considering that they scored between 8 and 10 in these tasks. In the traditional test, Kidd and Eduard were more interactive in the classroom, mainly with questions. With this, they obtained a Good concept. Felix and Mohamed obtained a Regular performance and Bart and Kortaro obtained a lower performance than expected, a Bad concept. Those students who performed below were less participatory. In this case, Kotaro was the student who least interacted in the classroom with questions to clarify his doubts due to his shyness. Mohamed was not present on the last day of the introductory class, and Bart was not present in the first two days of the introductory classes.

In both executions, it was asked in the feedback about the reason for such low performance, so the students reported that they did not study for the test, that is, they all answered only with classroom learning. It is also noteworthy that after carrying out the theoretical exercise and the traditional test, there was a moment to discuss the answers to clarify doubts and favor that students were able to apply such concepts in practical activities.

It can be seen in Table III that Anne and Cofresí achieved a good performance for having achieved mostly Gibbs avatar. The fact that Henry is missing two days from the experiment justifies the great part of obtaining the Marty avatar, having only performed well when he was present and participating in the dynamics. Although Henry performed a good analysis of the session report, his poor performance is justified by the fact that he found few defects and did not clearly record some of those defects. In this context, the arithmetic mean of Cofresí, Henry and Anne was 3.07, 1.71 and 3.14, respectively. As a result, only Cofresí and Anne were the ones who were able to look for the treasure in the classroom for having generally reached level 4.

TABLE III. AVATAR BY STAGE IN THE FIRST EXECUTION

	Tester		
Stage	Cofresí	Henry	Anne
Train the Pirate	🟧	🟧	🟨
Take Orders	🟨	🟥	🟨
Equip Pirate	🟨	🟥	🟨
Find Treasure I	🟩	🟥	🟩
Discuss Strategies I	🟨	🟥	🟨
Fight in Battle I	🟨	🟥	🟨
Validate Results I	🟩	🟥	🟩
Reward Featured Pirate I	🟩	🟧	🟩
Buy Resources I	🟨	🟨	🟨
Find Treasure II	🟩	🟨	🟩
Discuss Strategies II	🟨	🟨	🟨
Fight in Battle II	🟨	🟨	🟨
Validate Results II	🟩	🟥	🟩
Reward Featured Pirate II	🟩	🟩	🟩

🟥 1	**Marty**
🟧 2	**Tuner**
🟨 3	**Gibbs**
🟩 4	**Jack**

Among the 31 defects, 4 were considered duplicates, due to two students registering them, in addition to having 2 false positive defects. It summarizes that 27 different defects were detected, 5 of which were in the catalog, thus having 22 defects pertinent to the SUT since its development, that is, around 81.5% of the defects were native to the SUT. In this case, 5 test sessions were carried out with 30 minutes each, with 27 defects detected in 150 minutes. So the average time for each detection and recording of a defect was 5.5 minutes, evidently considering three testers and defects not duplicated.

In the second execution, most students managed to perform well by observing the amount of Jack-level avatar (Table IV). It is noted that these students had an excellent performance in the crucial stages of the dynamics, which had as activity the detection of defects and analysis of the opponent's session report. In this context, Kidd, Bart and Eduard reached the Jack avatar for having the respective arithmetic averages 3.05, 3.15, 3.05, and consequently were able to find the treasure in the classroom. While Mohamed, Felix and Kotaro obtained Gibbs avatar, but very close to reaching level 4 with the respective arithmetic mean 2.95, 3.0 and 2.65.

TABLE IV. AVATAR BY STAGE IN THE SECOND EXECUTION

	Tester					
Stage	Kidd	Bart	Eduard	Mohamed	Felix	kotaro
Train the Pirate	🟩	🟥	🟩	🟨	🟩	🟧
Take Orders	🟨	🟨	🟨	🟨	🟨	🟨
Equip Pirate	🟨	🟨	🟨	🟨	🟨	🟨
Find Treasure I	🟨	🟩	🟩	🟨	🟨	🟨
Discuss Strategies I	🟨	🟨	🟨	🟨	🟨	🟨
Fight in Battle I	🟨	🟨	🟨	🟨	🟨	🟨
Validate Results I	🟥	🟩	🟧	🟥	🟥	🟩
Reward Featured Pirate I	🟩	🟩	🟩	🟩	🟧	🟧
Buy Resources I	🟨	🟨	🟨	🟨	🟨	🟨
Find Treasure II	🟨	🟨	🟨	🟨	🟨	🟨
Discuss Strategies II	🟨	🟨	🟨	🟨	🟨	🟨
Fight in Battle II	🟨	🟨	🟨	🟨	🟨	🟨
Validate Results II	🟨	🟥	🟥	🟥	🟥	🟥
Reward Featured Pirate II	🟩	🟩	🟩	🟩	🟩	🟧
Buy Resources II	🟨	🟨	🟨	🟨	🟨	🟨
Find Treasure III	🟨	🟩	🟨	🟨	🟩	🟨
Discuss Strategies III	🟨	🟨	🟨	🟨	🟨	🟨
Fight in Battle III	🟨	🟨	🟨	🟨	🟨	🟨
Validate Results III	🟧	🟩	🟧	🟨	🟩	🟥
Reward Featured Pirate III	🟩	🟩	🟩	🟩	🟩	🟧

🟥 1	**Marty**
🟧 2	**Tuner**
🟨 3	**Gibbs**
🟩 4	**Jack**

Among the 66 defect records, 6 were false positives, and 14 were duplicates (two or more students reported the same defect). It can be summarized that 46 different defects were detected, and 6 were contained in the catalog, thus having 40 defects pertinent to the program since its development, that is, around 80.5% of the defects were native to the system. In this case, 18 test sessions were performed with 30 minutes each, with 46 defects were detected in 540 minutes. So the average time for each detection and recording of a defect was 11.7 minutes, evidently considering six testers and non-duplicated defects.

Given the data, it is clear that in both executions there was a very similar effectiveness in detecting native defects despite the difference in educational level. In addition, it is noteworthy that all students in both executions had a very similar overall performance, except Henry for not being present in three days of the experiment.

B. Qualitative Results

The data collected at the feedback moment were classified based on the SWOT (Strengths, Weaknesses, Opportunities, Threats) Matrix [21] [22]. In the feedback, the participants expressed themselves by conducting a critical analysis informing the positives, negatives, opportunities and threats, including suggestions for improvements, which were recorded by audio equipment. In general, these data were analyzed on materials, didactics, profiles, dynamics (gamification), rules and each current stage.

In both executions of the experiment, the students considered the “Find Treasure” stage as fundamental, especially when there was a discussion of the results of the exercises and the traditional test. During this discussion, they were able to perceive correctness of their responses, fix the subjects improving learning on how to create missions and how to use exploration strategies, consequently, improving the applicability of ET [20].

According to the students' report, the feedback moment was important to convey their opinions, suggestions and critical view of the experiment. Thus, the factors listed below are in accordance with the SWOT Matrix. The students reported as strong factors:

a) *Demonstrated examples, as well as how to present them*, contributed to improve the learning of the test approach, and that the examples of real cases were of great value to understand about such an approach,

b) *Topics and contents of the introductory classes*, contributed to expand knowledge about testing,

c) *The scenario of treasure hunting, using some characters from the Pirates of the Caribbean movie*, this relationship was quite adequate, as students felt immersed in the pirate world having to explore the search for treasures (defects). In addition, they really liked the characters related to the level they reached, and also the play strategy in general,

d) *The dynamics*, this contributes to a better understanding of the applicability of the ET, being the moment that the students liked most, emphasizing that they wanted to have more practice because it was very fun and exciting as they found more defects and interacted with each others in the “Fight in Battle” stage,

e) *Letters and Avatar*, they liked the cards a lot and even suggested using a real avatar doll to put their accessories on as they got it.

Regarding the strengths, the first four items were mentioned in both executions of the experiment, and item “e” was mentioned only in the second execution.

The students mentioned as weak factors:

a) *Didactics*, in the first execution the specialist caught on several examples on the same subject, so this was reported as a negative factor in some cases just when there was a quick understanding in the first example. In the second execution the specialist explained at times quickly, so this was also seen as a negative factor for understanding some examples,

b) *Practice*, in the first execution it was emphasized that they expected more practice, as they would feel safer and more efficient in the exploration,

c) *Personalization of the avatar*, it was perceived that more time was spent than expected in the “Equip Pirate” and “Buy Resources” stages, due to the students need to personalize the avatar, that is, in the first execution of the experiment one student at a time went to the judges to select the cards because there is a minimum amount of each type of card, thus causing the loss of time in personalizing the avatar.

Regarding the weaknesses, the first two items were mentioned in both executions of the experiment, and item “c” was mentioned only in the first execution.

The students mentioned as opportunities:

a) *Attack and Defense Cards*, diversification of resources destined for attack and defense,

b) *Tool to manage resources*, it would be important to obtain a tool that would allow the letters to be generated at random, mainly to manage all the resources obtained and the possible ones,

c) *Test application*, obtaining more test sessions to improve defect detection,

d) *Avatar*, possession of a doll to personalize it and also be visible to all resources obtained,

e) *Requesting letters*, selecting more cards, as the “fight” moment is quite fun,

f) *Examples*, exemplification making analogies with something from everyday life to facilitate the setting of examples.

Regarding opportunities, the first three items were mentioned in both Executions of the experiment, and the last three items were mentioned only in the second execution.

The students mentioned as threats:

a) *Letters*, having many letters may compromise the monitoring and progress of the “Equip Pirate” and “Buy Resources” stages, due to the fact that it is necessary to reorganize them every time a person removes the letters,

b) *Participants*, having many people (tester profile) can compromise the progress of the “Fight in Battle” and “Validate Results” stages. Because there is a lot of interaction between the participants and a relatively large number of session and analysis reports to validate, respectively,

c) *Paired Exploration*, the students reported that if the “Find Treasure” stage were carried out in pairs it would not be so advantageous, as they would most likely not be able to keep the focus on exploration,

d) *Specialist*, the instructor must in fact be an expert in the test area, otherwise, it may hinder the progress and teaching-learning process of the experiment.

Regarding threats, the first three items were mentioned in both executions of the experiment, item "c" was only mentioned in the first one while item "d" only in the second one.

At the end of the first execution of the experiment there was the addition of the 8th day dedicated to practical activity, and also the inclusion of more cards, use of the card album instead of the resource spreadsheet to minimize the problems with excessive time spent, adjustments in the theoretical content to grant more time for the exercises.

C. Discussion of Results

The analysis of all the results was carried out observing two factors: (a) medals obtained in each execution of the experiment, comparing in percentage terms the quantity of medals expected in relation to the quantity obtained, and (b) activity avatar, participatory action and final avatar, also comparing the amount of avatar possible to obtain in relation to the quantity obtained in each level. The number of medals received depended entirely on the performance of the participatory actions and also if the participants were able to achieve maximum marks in the specific activities at each stage, for this reason the analysis was carried out in the context of obtaining the medals.

In Fig. 4 it is evident that in both executions of the experiment there was an excellent use, since it reached more than 70% of the possible medals, thus being considered satisfactory. In detail, in the first execution of the experiment it was possible to obtain a total of 294 medals (100%), but 210 medals (71.4%) were obtained, with 34 (approximately 16.2%) of these medals being obtained by reaching the maximum score in some activities. In the second execution of the experiment it was possible to obtain 840 medals (100%), however 644 medals (76.6%) were obtained, 110 (approximately 17%) of which were for reaching the maximum score in some activities. All of this shows that the students were very participative and interacted trying to reach the maximum of the scores.

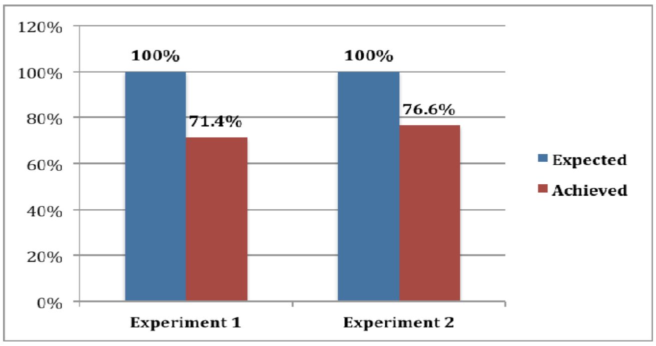

Fig. 4. General performance related to medals obtained.

To generate the charts expressed in Fig. 5, 6 and 7, the following calculations were performed: Experiment 1, three participants, 14 stages being performed due to the two iterations, so there are 48 avatar to be considered as total (100%); Experiment 2, there were six participants, 20 stages were performed due to the three iterations, so there are 120 avatar to be considered as total (100%). In both executions, the "Request Help" and "Reward Winner" stages are disregarded.

Fig. 5 shows a significant amount of level 1 avatar caused by the absence of certain students on some days. On the other hand, the significant amount obtained from level 4 avatar was not so influenced by the fact that the level was automatically granted at times when there was no way to analyze the activities, as this occurred in just two stages, in this case, "Take Orders" and "Buy Resources". Therefore, it is possible to affirm that the general performance in carrying out the activities was quite satisfactory, considering that most of the avatars obtained were level 4.

In Fig. 6 the percentage of level 4 shows that in the second execution the students interacted a little more, something that was congruent to what was observed in the classes. This can be justified by the fact that first-time students consult the website extensively for the purpose of clarifying doubts or remembering rules. In the second execution, the students practically did not consult the website, due to the Internet connection being unstable and thus impelling them to ask more questions.

It is observed (Fig. 7) that the overall performance in the stages was better for the students in the first execution of the experiment because they obtained 31% level 4 avatar. In this context, the fact that Cofresí and Anne managed to find more defects among all students may have corroborated. And when analyzing the performance obtained by adding the percentages of levels 3 and 4, the students of the second execution were better, however, a performance very similar and considered satisfactory to the expected. Obviously, this overall performance could be even better, if by chance, everyone was always present and participatory.

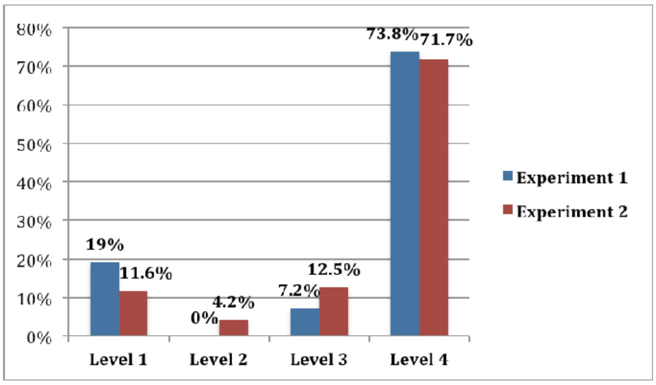

Fig. 5. General performance in the activities.

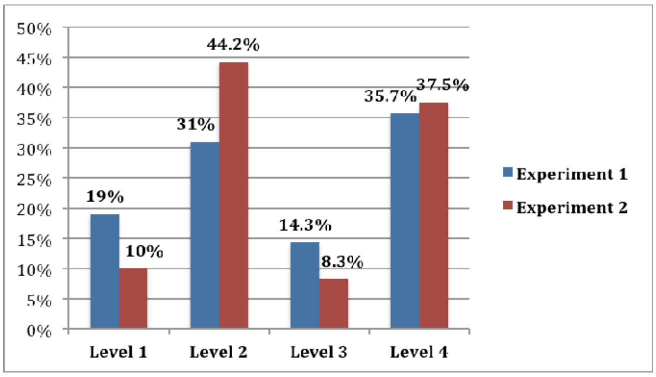

Fig. 6. General performance in the participatory actions.

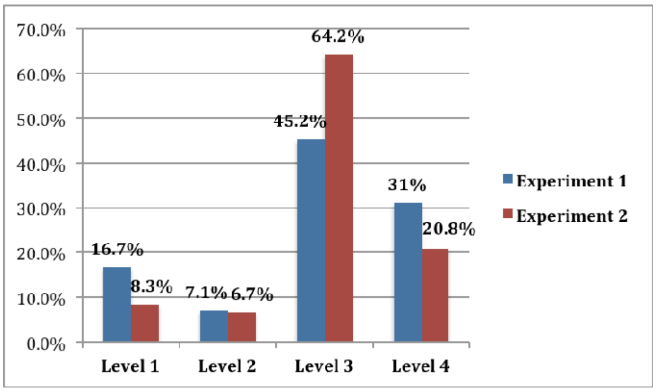

Fig. 7. General performance in the stages.

In view of the data presented, it is clear that gamification was something that provided better teaching and learning, in view of the achievement of an excellent overall performance in the two executions of the experiment. In this context, some reports stand out that corroborate with all these data presented.

The students stated that all the materials were simple and very understandable. The students also emphasized that the presentation and discussion of the results of the analysis of the reports were essential to acquire a better critical view, as well as providing a more creative thought in the exploration of the SUT.

The students reported that the applicability of the ET during the dynamics provided the observation of several flows and details of the Graphical User Interface (GUI) and usability of the SUT. They also stated that with the freedom obtained to explore, they felt more comfortable in order to detect defects of the most diverse levels of criticality. In addition, the use of SBTM actually provided a better structuring and organization of the applicability of the ET, being evidenced mainly in the analysis of the session report, as it would be quite complicated to analyze and reproduce the defects detected by the opponent without the aid of a document with the registration of such defects.

Especially in the second execution of the experiment, the students reported that they felt very motivated and intuitively explored their mobile devices and computer programs when they were using them. This fact caused the detection of two defects in the Microsoft Excel program. This was due to the free and spontaneous action of these students who reported being excited to test any software they used.

Still based on the reports of these students, they were impressed how the ET approach really allows several people to detect totally different defects in the same functionality. In addition to all this, it is also noteworthy that students answered a form to identify the satisfaction degree. All students considered the experiment Excellent, while seven students considered Excellent and two reported that the ET approach, the gamified approach and the didactic approach were good.

In this same form on satisfaction, it was identified that the greatest difficulty for students in the first execution of the experiment was understanding about some concepts, but this difficulty was met with practice. In the second execution the students felt afraid of making mistakes in registering defects in causing the penalty when it was considered a false positive defect, as well as there was insecurity in finding defects, as they were unaware of the system to be tested, even though they were aware of the option of help.

VII. Threats to Validity

This section discusses some threats to validity of the experiment carried out.

A. *Internal Validity*

It is argued that the Gamified Framework was applied properly, as the students obtained a performance proportional to their participation in the experiment. In addition, it was evident that gamification provided students with immersion in the game world and engagement to obtain the best possible performance, all of which could be confirmed by the reports at the feedback moment.

B. *External Validity*

Despite being nine students (testers), it is considered that it is possible to be reproduced in other environments when analyzing some factors listed: (a) the testers did not obtain previous knowledge about software testing, (b) the system used had already been subjected to some initial tests, however, it was still possible to discover a large number of defects native to the program, and (c) the rules and materials are available in other initial studies [18] and can be consulted freely.

C. *Construct Validity*

Initially, it was adopted as an initial strategy to conduct an SML [7] to identify several positive and negative factors of the applicability of ET, from this there was the identification of the games elements, description of rules, the purposes and the necessary materials. This whole process always took place with the monitoring and validation of a specialist in the Software Engineering area with numerous published papers on gamification in education. Thus, the experiment was built and executed, obtaining results with adherence to what was aimed at investigating.

D. *Conclusion Validity*

It is noticeable that the qualitative results converged with the quantitative results obtained, and that this allows us to affirm that the adherence of the results represents a very relevant factor for the applicability of the Gamified Framework on ET teaching-learning. In this context, the conclusive validity is based on the results obtained, showing that gamification really helped in the students' teaching-learning process.

VIII. Conclusion

It is noticed that the students were satisfied with the experiment, however they would like more test sessions, given that it was a more engaging practice with gamification, causing a decrease in any interaction resistance. In view of this, it can be seen that the playful issue caused greater interest in students, when it was evidenced that they felt stimulated to obtain a good performance.

The gamified framework has been planned and designed to train people to autonomously apply the ET approach in a structured and systematic way, both in academia and in industry. In this case, it was important to involve training activities ("Training the Pirate" stage) precisely in the perspective of obtaining similar results as it has occurred.

Regarding the analysis of skills by type of error, no data was collected to specifically measure this, because it will be a study in the future, due to the exploration strategies used to influence the types and severities of defects. In the present study, 7 exploitation strategies [23] were exemplified in order to help in the detection of defects because he / she did not have any knowledge about the SUT, and according to reports of the participants these strategies facilitated the detection of defects, mainly, usability, functionalities and GUI. In addition, we intend to apply the Gamified Framework in an industrial context, involving undergraduate participants for an analysis focused on observing the effectiveness and efficiency of that Framework.

ACKNOWLEDGEMENTS

The authors would like to thank CAPES (Coordination and Improvement of Higher Level or Education Personnel) for granting an institutional Master's and Doctoral scholarship to the author of this work, the students participating (UFPA and IFET) in the experiment and the collaborators who participated in the experiment in the Judge profile.

REFERENCES

[1] R. Santos, P. Santos, C. Werner, and G. Travassos, "Using Experimentation to Support Research in Software Engineering Education in Brazil". In: I Fórum de Educação em Engenharia de Software. 2008. DOI: 10.13140/2.1.2946.4002.

[2] C. Kaner, "A Tutorial in Exploratory Testing". QUEST, 2008.

[3] D. Pfahl, H. Yin, M. Mantyla, and J. Munch, "How is Exploratory Testing Used?: A state of the Practice Survey". EPEM'14, September 18-19, Torino, Italy. 2014. Copyright 2014 ACM. ISBN: 978-1-4503-2774-9/14/09.

[4] J. Bach, "Exploratory Testing". In: The Testing Software engineer, 2nd ed., E. van Veenendaal (Ed.) Den Bosch: UTN Publishers, pp. 253-265. 2004.

[5] IEEE, "SWEBOK V3.0: Guide to the Software Engineering Body of Knowledge". Computer Society. 2014.

[6] J. Itkonen and M. Mantyla, "Are test cases needed? Replicated comparison between exploratory and test-case-based software testing". Empirical Software Engineering, pp. 1-40. 2013.

[7] I. Costa and S. Oliveira, "An Evidence-Based Study on the Efficiency and Efficacy of Exploratory Testing". 16th International Conference on Information Systems & Techology Management – CONTECSI. 2019.

[8] M. Fardo, "Gamification Applied in Learning Environments". Renote - Novas Tecnologias na Educação. 11, 2013. ISSN 1679-1916.

[9] K. Kapp, "The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education". North Carolina: Pfeiffer, 2012, 366 p.

[10] K., Werbach and F. Hunter, "For The Win: How Game Thinking Can Revolutionize Your Business". Filadélfia, Pensilvânia:Wharton Digital Press. 2012.

[11] J. Herbert, "Patterns to Teach Software Testing to Non-developers". SugarLoafPLoP 16 Proceedings of the 11th Latin-American Conference on Pattern Languages of Programming. 2016. HILLSIDE 978-1-941652-05-3.

[12] P. Valle, E. Barbosa, and J. Maldonado, "A Systematic Mapping on Teaching Software Testing". Anais do XXVI Simpósio Brasileiro de Informática na Educação CBIE-LACLO 2015. DOI: 10.5753/cbie.sbie.2015.7171.

[13] T. Ribeiro and A. Paiva, "iLearnTest - Educational Game for Learning Software Testing". Atas da 10ª Conferencia Ibérica de Sistemas y Tecnologías de la Información (CISTI'2015). Universidade do Porto. 2015.

[14] Y. Chou, "Actionable Gamification - Beyond Points, Badges, and Leaderboards". Octalysis Media. 2016.

[15] F. Benitti, "Evaluating Learning Objects for Teaching Software Testing". Nuevas Ideas en Informática Educativa TISE. 2015. ISBN: 978-956-19-0929-8.

[16] I. Costa and S. Oliveira, "A Systematic Strategy to Teaching of Exploratory Testing Using Gamification". 14th International Conference on Evaluation of Novel Approachs to Software Engineering. ENASE. 2019.

[17] J. Bach, "Session-Based Test Management", STQE, vol. 2, no. 6, 2000.

[18] I. Costa and S. Oliveira, "A Gamified and Systematic Approach to Teaching and Learning of Exploratory Testing". 16th International Conference on Information Systems & Techology Management – CONTECSI. 2019.

[19] A. Alcântara, S. Oliveira, R. Junior, W. Cardoso, and L. Rodrigues, "SAW: A Simulation and Assessment Generation System to Aid Teaching and Learning". XXIII conferência Internacional sobre Informatica na Educação (TISE). 2018. ISBN:978-956-19-1111-6.

[20] I. Costa, S. Oliveira, L. Cardoso, A. Ramos, and R. Sousa, "A Gamification for Teaching and Learning Exploratory Software Testing: Application in an Experimental Study". XVIII Simpósio Brasileiro de Jogos e Entretenimento Digital. Rio de Janeiro – RJ. SBC – Proceedings of SBGames 2019 – ISSN: 2179-2259. Education Track – Short Papers. 2019.

[21] E. Takao, N. Copppini, and A. Toregeani, "Application of the SWOT Technique in Enabling a Support System for Executives with Free Software in a Plastic Packaging Industry". Caderno de Administração. v. 14, n.1, p. 9-17. 2006

[22] M. Santos, J. Grechi, and P. Bermejo, "Assessing the Impact of SCRUM on Software Development Using SWOT Analysis". In: XXX ENEGEP. 2010. São Paulo.

[23] M. Micallef, C. Porter and A. Borg, "Do Exploratory Testers Need Formal Training? An Investigation Using HCI Techniques". IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops. 2016. DOI 10.1109/ICSTW.2016.31.