Can we use the Flipped Classroom Model to teach Black-box Testing to Computer Students?

Leo Natan Paschoal
University of São Paulo
São Carlos - SP, Brazil
paschoalln@usp.br

Brauner R. N. Oliveira
University of São Paulo
São Carlos - SP, Brazil
brauner@usp.br

Elisa Yumi Nakagawa
University of São Paulo
São Carlos - SP, Brazil
elisa@icmc.usp.br

Simone R. S. Souza
University of São Paulo
São Carlos - SP, Brazil
srocio@icmc.usp.br

ABSTRACT

Computer science programs have been delivering newly undergraduate students to the software industry without sufficient knowledge on how to perform software testing activities. It occurs because that activity is usually taught as part of Software Engineering courses, which means that teachers have to put a lot of efforts to teach the main testing techniques and criteria along with many other contents. In this way, the content is mostly addressed in a theoretical way as the teacher has to stick to the course schedule and has no time to address the testing practice. On the other hand, some studies reported the efforts of professors to optimize the time during classes and attract students to the software testing area. These efforts are associated with the employment of activities in which students are actively engaged in the classroom, solving problems, developing real projects, or dealing with real cases. A pedagogical model that seeks to leverage such an approach is the flipped classroom model. However, while some have praised that model, others have criticized it. In the state of the art, there is no consensus about the appropriateness of flipped classroom for specific learning contents. In this paper, we present an experimental study that was conducted to verify the suitability of the flipped classroom model to teach software testing, especially black-box testing. This study comprised an analysis of students' learning gains when subject to the flipped model and the workload introduced when compared to the traditional teaching model.

CCS CONCEPTS

• **Software and its engineering** → *Software verification and validation*; • **Social and professional topics** → Software engineering education;

KEYWORDS

Computer Science Education, Flipped Classroom Model. Software Testing.

SBQS'19, October 28-November 1, 2019, Fortaleza, Brazil

ACM ISBN 978-1-4503-7282-4/19/10…$15.00
https://doi.org/10.1145/3364641.3364659

ACM Reference Format:
Leo Natan Paschoal, Brauner R. N. Oliveira, Elisa Yumi Nakagawa, and Simone R. S. Souza. 2019. Can we use the Flipped Classroom Model to teach Black-box Testing to Computer Students? . In *XVIII Brazilian Symposium on Software Quality (SBQS'19), October 28-November 1, 2019, Fortaleza, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3364641.3364659

1 INTRODUCTION

Software testing refers to activities performed during the software development life cycle with the intent of identifying the presence of faults in the software under development [32]. This is the only activity of the life cycle that is capable of finding faults, offering subsidies to software quality assurance [7]. Faults can produce errors and leave the software system in a state of failure [13]. Defects in software systems bring economic damage to users and the software industry [51], reduce user confidence [38] and can cause irreversible damage to the user's life [42].

Although important, software testing is still a content not normally present in universities' curricula [46]. The software testing activity is usually taught in software engineering courses along with other contents such as requirements engineering, software processes, software reuse, software maintenance, among others [28]. In such courses, teachers usually focus on software projects and software modeling more than on software testing [27]. For these reasons, the majority of students of computing programs graduated without knowing how to test a software system adequately [3, 11]. As a consequence, software companies have to provide additional training to fill the lack of knowledge of students on software testing. If there is no budget to provide enough training, the quality of the software projects under developed may be compromised [9]. Finally, students that had a better training in software testing enable companies to invest their budget in more strategic areas [47].

In order to improve the capability of students to perform testing activities, some institutions are putting some efforts to encourage students to take training in software testing. Some proposals establish free and open courses that are available by means of Massive Open Online Courses (MOOCs) [35]. There are also studies that report on the use of pedagogical models to teach software testing [5, 39]. Preliminary studies describe that it is important and necessary to use pedagogical models that support active learning [40]. These studies advocate the use of project-based learning, collaborative learning and other educational strategies as a way to motivate students to learn software testing in a practical way, beyond the traditional and theoretical form [10].

Despite the existence of initiatives, however, many teachers reveal that they are unable to teach the practice of software testing by due to the time available for classes [34]. Most teachers who work in the software testing area employ traditional teaching methods, and when they teach this subject with another method, they use methods that mirror the traditional model in which most learning activities have to be done by the students at home, without teacher's support [16]. In this sense, activities are still to be done outside the classroom while that the time inside the classroom is mostly spent with lectures in which students are not actively engaged [35].

Seeking to optimize classroom time towards engaging students in active activities (e.g., conducting practical activities, projects, among others), teachers from some areas are making use of a pedagogical model entitled Flipped Classroom (e.g., physics course [2], mathematics course [26], medicine courses [37], and english reading comprehension courses [21]). In this model, passive activities are performed at home by students and active activities are performed in the classroom [23]. At home, students can watch lectures recorded by teachers using a broad range of technologies and formats such as animations, videos, simulations, and slides [6]. Afterward, during the class, the students can have their questions about the material studied at home solved [6]. Then, they are engaged in activities that will exercise the content studied, or sometimes, learn new and more complex concepts that require them to understand the content studied at home [22].

Studies conducted in some areas reported that the Flipped Classroom model manages to motivate students since they become more engaged [25]. There are also areas in which teachers observed that students who study with the flipped classroom model can have better use of the educational content, learning more and consequently surpassing the performance of students who learn from the Traditional model [31]. On the other hand, some studies revealed that the performance of students learning from the flipped model is not superior when compared to students learning from the traditional model [19]. Given such claims, one believes that the flipped classroom model can be employed to improve the learning of certain contents [33]. Consequently, before the teacher simply adopts this pedagogical model, it is necessary to recognize whether it really produces a significant impact on student learning [6].

In a preliminary study on the use of the flipped model to teach software testing, it was observed that students who studied software testing with flipped classroom were motivated to the same extent as the students who studied this content with the traditional model [33].Moreover, Paschoal et al. [33] observed that the performance of students in knowledge tests is different when comparing students that learned with the flipped classroom model to those with the traditional model. Despite the existence of such a study, the authors did not measure how much the students have learned about software testing considering each model. As stated by Kong [24], it is important to evaluate the model efficiency for providing gains to the student's knowledge.

In other domains, other aspects are under investigation. In the work performed by He et al. [20], the authors highlight the need to assess the additional workload generated by the Flipped Classroom model over the students. This additional workload is a consequence from the effective implementation of the model, which requires that the students get prepared before classes. Gannod, Burge and Helmick (2008) [18], for example, present some arguments about their beliefs that students that study with the flipped model need to dedicate more time to get prepared for the classes. In this sense, flipped classroom researchers must measure students' efforts whenever possible. If the additional workload introduced by the model is over too high, it can become less attractive from a pedagogical point of view [20].

Based on what was presented before, we report an experimental study performed to analyze the impact of flipped classroom on the knowledge acquired by the student and whether the realization of practical activities makes the students acquire more knowledge. We consider this study as an extension of the previous research conducted in the study of Paschoal et al. [33]. In this sense, we addressed the equivalence partitioning criterion as teaching subject. It is one of the main criteria for software testing and is classified as a black-box testing [13]. Additionally, it is a very common subject within courses on software testing in Brazil [34].

We have structured our article in the following way. Section 2 presents an overview of the flipped classroom model. The planning and execution phases of our study are detailed in Section 3. In sequence, Section 4 reports the results of our experimental study. During section 5, we describe the threats to validity and how we try to mitigate them. Finally, in Section 6 are summarized the conclusions and future research directions.

2 FLIPPED CLASSROOM MODEL

In the traditional teaching model, most of the time "spent" by the students in the classroom is dedicated to passively watch the teacher exposing the content. As a consequence, little time is left for students to perform practical activities to fix the subject addressed. Incapable of solving exercises left by the teacher in such a small amount of time, students have to finish the activities at home [45]. However, most doubts arise when they are solving the exercises and applying the concepts addressed during the class [6].

Due to the characteristics of the traditional model, some studies and practices were made seeking to optimize the time spent within the classroom and offer better conditions to the students so they can learn course contents with more deepness [4]. This way, the flipped classroom model emerged as an alternative to the traditional one.

The flipped classroom model can be understood in the following way: what is usually made within the classroom is now made at home, and what is traditionally made as homework is now made at the classroom [6]. To achieve this, the students receive the teaching materials about the content that would be addressed in the classroom (e.g., videos, presentations, and podcasts) so they can study and get prepared to solve exercises and participate in discussions proposed by the teacher during the class [8]. This way, the time that was spent before to present content is now destinated to individual or group activities with the support of the teacher [43].

In this model, the professor left his role as the communicator of the knowledge and becomes the learning process mediator. The students, in turn, become responsible for getting prepared before the class and have better conditions to become active during the learning process as they solve exercises and participate in discussions. The overview of the flipped classroom model is shown in Figure 1[1]. The figure highlights the three main moments of study addressed by the model: (i) before the class; (ii) during the class; and (iii) after the class. In each one of these three moments, the student will be exposed to activities (passive or active).

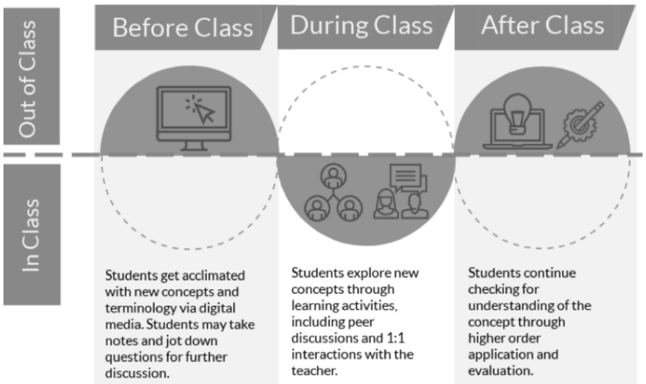

Figure 1: The Flipped Classroom Model

Even though it is a recent teaching model, which have been investigated in some areas, there are studies that present the benefits and drawbacks when employing flipped classroom for undergraduate and graduate courses. Considering the benefits, the studies highlight the following aspects: (i) when the student do the homework in the classroom, the teacher has a better view of the real difficulties of the students [17]; (ii) the students learn on their own pace, using the teaching material recommended by the teacher as many times as they need [41]; (iii) the students begin to reinforce their knowledge as they share it with their pairs during activities in the classroom [50]. About the drawbacks: (i) flipping the classroom requires additional time by the teacher to prepare the teaching material (e.g., produce and/or select video lessons, presentations, etc.) [50]; (ii) the students may not be ready to work with the model and show some resistance to work with it [41]; (iii) some students face difficulties to work with the model as they cannot expose their doubts in the classroom [49]; and (iv) uninterested students may not watch the lectures and end up losing the content presented [50].

Although the recent literature addresses a series of advantages and reinforces the cares that should be taken by the teacher to overcome the disadvantages of the model, it is important to investigate whether the students learn better with the flipped classroom model than with the traditional one [8, 36]. Bergmann e Sams [6] – the creators of the flipped model – state that when teaching chemistry to last year primary education students, the performance of students remains the same, but observed an improvement of the engagement of the students. In this sense, some studies have been investigating the potential of the teaching model within the teaching and learning process of different contents (e.g., for teaching principles of mechanical ventilation to internal medicine residents [44], student performance on psychomotor skill acquisition [14]), especially in the computing area.

In the work of Fassbinder et al. [15], for instance, the authors analyzed the feasibility of the flipped classroom model to teach loop statements when compared to the traditional model. At the end of their study, the authors found no significant difference in the grade means in programming exercises. Similarly, Amresh et al. [1] performed a study to verify the efficiency of the model to teach introductory computer programming. The authors evaluated the performance of the students in exercises throughout the course. The results of the experiment suggest that the flipped classroom model produced better mean scores in the course. The group of students that learned with the traditional model had lower scores in the middle and at the end of the course, when compared to the experimental group.

3 STUDY SETUP

In this section, we present the planning phase of the experiment conducted to assess the flipped classroom model to teach the equivalence partitioning criterion, which is a black-box testing technique. To assess the model, we compared it to the traditional teaching model, paying special attention to instant learning, the effective learning resulted from practical activities and knowledge retention, and the impact of the model in the time spent by the students during the study/instruction. The experimental study was established following the experimental process presented by Wohlin et al. [48].

3.1 Goal definition

The goal of our experimental study is to analyze the flipped classroom model, aiming to verify its impact as a teaching model, with respect to learning, knowledge retention, and study effort, from the researcher's point of view, in the context of software engineering students learning software testing.

3.2 Context selection

This experiment was conducted in the first semester of 2018, with students of the 7th semester of the computer engineering program, offered in partnership between the Institute of Mathematics and Computer Sciences (ICMC-USP) and São Carlos School of Engineering (EESC-USP). The students that participated in this experiment had started to learn the software testing content but were not introduced to testing criteria and techniques. This way, the content addressed in this experiment was the black-box testing technique and the equivalence partitioning criterion.

3.3 Hypotheses formulation

This experiment investigates two specific research questions (Q). For each one of them, we defined research hypothesis:

Q_1**:** Do teaching the black-box testing technique with the flipped classroom model provides more learning than teaching with the traditional model?

- **Null Hypothesis:** There is **no** difference between how much students learn with the flipped classroom and traditional model.

[1]Extracted from <http://bit.ly/2Szvw8Z>.

- **Alternative Hypothesis:** There is a difference between how much students learn with the flipped classroom and traditional model.

Q_2**:** Do the students that learn with the flipped classroom model take more time to study the black-box testing technique than the students learning with the traditional model?

- **Null Hypothesis:** There is **no** difference between the time spent by students learning with the flipped classroom model and students learning with the traditional one.
- **Alternative Hypothesis:** There is a difference between the time spent by students learning with the flipped classroom model and students learning with the traditional one.

3.4 Selection of Subjects

As mentioned before, this experiment was conducted in the context of a software engineering course. That course offering had 41 students enrolled. These students were randomly divided in two groups. Therefore, 21 students were assigned to the control group (learning with the traditional model) and 20 to the experimental group (learning with the flipped model). These students had no previous experience in the area of software testing.

3.5 Variable selection

In experimental studies, the cause and effect relationship is measured using dependent and independent variables. The independent variables are those that affect the dependent variables while the dependent variables are affected by the independent ones. Based on that, we considered the following variables in this experiment:

Independent variables. In this experiment, the following independent variables were considered:

- **Strategy used to teach black-box testing (factor):** It consists of the actions developed that represent and explain the way the content is addressed and is consolidated by means of practices and interactions between the teacher and the students. This experiment had two treatments:
  Treatment A: Teaching is done following the flipped classroom model.
  Treatment B: Teaching is done following the traditional model.

Dependent variables. In this experiment, the following dependent variables were observed:

- **Learning:** This variable refers to the learning obtained by the student when subject to the teaching models. In order to measure learning, we considered the approach employed by Lyra et al. [29], in which learning is measured by calculating deltas of the performance of the student at different stages during the learning process, resulting in three deltas: immediate learning ($\Delta 1$), learning from active activities ($\Delta 2$), and knowledge retained ($\Delta 3$). Each of them is defined as follows:
  Δ1 refers to how much the student learns right after the instruction with the pedagogical model;
  Δ2 refers to how much the student improved their knowledge after performing active activities; and
  Δ3 refers to the knowledge retained by the student at the end of the experiment.
- **Instruction time:** This variables refers to the time in minutes that each student spend to study the content by itself. This variable is observed due to the impact of the flipped model on the students' workload introduced by the before class moment.

3.6 Material

Throughout this study, we developed and utilized different types of instruments. Some materials were used to teach students, while that other were used to collect the data used for analysis. Beyond these resources, the institutional version of the learning virtual environment Moodle was employed, aiming to provide the learning materials to the experimental group.

It was necessary to some educational resources to support the learning process of the students in the "before class moment". In total, we selected and provided two video lessons (one with an overview of black-box testing and another about equivalence partitioning criterion). Both lessons were selected from MOOCs about software testing, summing up 15 minutes and 49 seconds. It is important to highlight that both videos were already employed in previous studies such as the one performed by Paschoal e Souza [35]. For the control group, we employed the presentation used in the experiment performed by Paschoal et al. [33].

To measure learning, we elaborated three tests (pre-test, intermediary test, and post-test). Each one of them is composed of a set of multiple-choice questions with 5 available options (only one is correct). The pre-test is a questionnaire with 10 questions that assess the previous knowledge of the student on the equivalence partitioning criterion. The result of this test is used as the starting value used to measure how much students learned and retained in terms of knowledge throughout the experiment.

The intermediary test is a knowledge questionnaire that has variations of the 10 questions of the pre-test questionnaire. This instrument is applied to students right after introducing them to the testing content using the teaching models. In this sense, the goal is to identify the knowledge obtained after the instruction. Lastly, the post-test questionnaire also contains 10 questions, which are variations of the same questions from the pre-test questionnaire, and is applied to students when they finish the practical activities and solve the doubts that emerge during the activities. The post-test result is used to measure the knowledge retained by the students since the beginning of the interventions. It is important to highlight that the questions used in the three questionnaires were obtained from a certification test from the software testing area and adapted by us.

The time spent by students during the instruction cannot be easily measured. He et al. [20] asked their students (Chemistry and Biology) to track and record the time spent for study (before class moment) and, at the beginning of the class applied a questionnaire to gather that information. Although it is not the most reliable approach to measure the time spent, it can provide valuable information to measure the additional effort resulting from the application of the flipped classroom model. In our experiment, we adopted the same strategy. Therefore, we developed and applied a questionnaire to gather information related to the time spent by each student to get prepared before the class.

Finally, the last instrument developed was the consent form, aiming to make it clear for the students their role in the experiment.

3.7 Experiment Operation – Planning

After finishing the experiment definition and planning, the teacher responsible for teaching the Software Engineering course defined the necessary number of classes for the experiment execution, adjusting the course schedule so the execution phase would not harm the course progress. In this sense, a strategy was established to avoid prejudices for students' learning.

The tactic stipulated to avoid harming the course progress and students' learning was established in a way that all students could participate in all experiment activities. In this sense, the students were notified about the experiment and could choose to have their data analyzed or not. In this way, the experiment was not considered an assessment activity and the students were able to learn the content as if it was a course activity. Students not willing to provide their data for analysis did not fill the questionnaires applied by us.

During the planning of the experiment operation, the institutional Moodle was configured to present the teaching materials available for the study. This configuration was done using the platform–tools for creating groups and limit access to the flipped content just for the experimental group.

The division of the students into two groups required us to use two different classrooms. One of them designated to the experimental group, in which we followed the flipped approach, and the other to the traditional setting.

Lastly, the last part of operation planning consisted of defining how the traditional class should happen. In this sense, the teacher responsible for the course should teach the content using presentations in a traditional manner. The content should be taught in a limited amount of time, after which the students' doubts should be solved while they performed an activity. After the class, the presentation file should be shared with the students of the control group using Moodle.

Students from the experimental group should access Moodle, watch the two lessons and record the time they took to do it and their doubts about the content. In the classroom, the students would have a moment to solve their doubts with the instructor responsible for the experimental classroom and accomplish an activity in pairs with the support of the instructor.

3.8 Experiment Operation – Execution

This experiment was executed over two weeks, comprising 4 classes. The first, the second and the fourth classes occurred on Wednesdays and the third class on a Friday. Due to a holiday between the first and the second class, we had to cancel one class which would occur on a Friday. As a consequence, the second class occurred one week after the first one. The experiment is described in the following, based on the sequence in which it was executed:

Stage 1: At the first meeting, the teacher informed the students that a different teaching model would be used for black-box testing education. The class was divided randomly into two groups. The students were notified that all of them had to accomplish the same activities although their scores would not impact their course grades. They were also notified about the questionnaires that would be applied during the experiment. Finally, the instructions for the activity and the consent form was given to the students.

After filling the consent form, the students were invited to participate in the pre-test. The teacher distributed the pre-test and asked the students to answer the questions individually, without querying any additional knowledge source (e.g., Internet and books). After that, the students from the control group were released from the class. It was necessary because the students from the experimental group had to be notified about the materials available in Moodle and that they had to study the content through the videos, record the time to study and the doubts they had studying.

Stage 2: This stage comprises the activities accomplished by the students from the experimental group (flipped classroom). The students had to watch the video lessons previously selected and available at Moodle (before class moment). The lessons address the equivalency partitioning criterion in a theoretical and practical way, providing examples of how it is applied. At the in-classroom moment, the students of the experimental group had 15 minutes to solve their doubts from the "before class moment". In that occasion, there was a discussion about the interpretation that the students had about the video lessons and a brief review. After solving the doubts, the student had to indicate how much time they spent at the "before class moment".

A knowledge test (intermediary test) was applied. We explained that the goal was to identify how much knowledge the students could obtain about the content addressed with the flipped classroom approach and the resolution of problems and doubts. The test had to be done individually, without additional knowledge resources.

Stage 3: In parallel with the stage 2, the control group was having a traditional class about black-box testing. The class took about 50 minutes, in which 30 minutes were spent to teach the technique and the equivalence partitioning criterion, and to solve students' doubts. The 20 minutes remaining were used to apply the intermediary test to the students.

Stage 4: After the instructions, all the students had the opportunity to accomplish an activity in pairs. However, students could only pair with other students from the same group (control/experimental). During the activity, the students present their doubts that could be solved by the teacher.

Although practices activities within the classroom are less common in the traditional model than in models towards active learning, it was necessary to include the activity during the classroom so we could control most threats to the validity of the performance of the students, such as the help of veterans or share of information between pairs. The goal of this activity was to let students reinforce the knowledge obtained by accomplishing a practical activity.

Stage 5: In the last class, the students from both groups were invited to answer the post-test questionnaire. As in the previous tests, they could not use external sources of information and had to do the test individually. This test was applied to let us measure how much did they learn from both teaching models.

After executing the experiment, the data collected were tabulated and analyzed. The next section presents the analysis performed.

4 DATA ANALYSIS AND RESULTS

At the beginning of the experiment, all course students were invited to participate in the study. We collected data throughout the process with knowledge tests and questionnaire forms. Despite all students enrolled in the course choose to participate in the experiment activities, some opted not to provide their data for analysis, reducing the number of samples. Moreover, some students did not participate in all activities proposed, compromising a few samples.

All tests, forms, and data were reviewed. Incomplete samples were discarded. As a result, we performed an analysis of the data from 27 students (13 from the control group and 14 from the experimental group).

Due to the limited number of complete samples, it was not possible to employ hypothesis tests. In this sense, the data were analyzed with descriptive statistics.

4.1 Learning

The different values associated with the variable learning were collected with knowledge tests and, from that, we calculated the values of Δ1, Δ2 e Δ3. To calculate each Δ, it was necessary to establish three formulas based on the study performed by Lyra et al. [29], which considered learning as an experimental variable during a controlled experiment about infographics.

$$\Delta 1 = Y_{(\mathrm{i})} - X_{(\mathrm{i})}$$

$$\Delta 2 = Z_{(\mathrm{i})} - Y_{(\mathrm{i})}$$

$$\Delta 3 = Z_{(\mathrm{i})} - X_{(\mathrm{i})}$$

in which:
X = Number of correct answers in pre-test
Y = Number of correct answers in intermediary test
Z = Number of correct answers in intermediary post-test
i = {student $_1$, student $_2$, student $_3$, ..., student $_n$}.

After checking the performance of students in the tests, the data was tabulated and the values for Δ1, Δ2 e Δ3 calculated. We performed a descriptive analysis, as shown in Table 1. Regarding the number of correct answers in the pre-test, the students achieved scores superior to 50% of the total number of questions. We believe that it happened due to the characteristics of the pre-test questions. Among the 10 questions, just 3 of them are theoretical. The other 7 questions are dedicated to the application of the criterion, which is considered to be very simple [34] and easily understood by those that are good at set theory.

The analysis of the number of correct answers in the intermediary test reveals that the score mean of students from the experimental group was very near to the score mean of the students from the control group. The median analysis revealed that the median of the number of correct answers from the experimental group was 7 while the median from the control group was 6.

The results regarding the number of correct answers in the post-test show that the experimental group achieved better scores than the control group. The mean of correct answers of the experimental group was 8.07 while that the mean achieved by the control group was 6.54. When it comes to the median scores, the experimental group achieved 9 points while the control group achieved 6 points. The data shows that at least 50% of the students of the experimental group score 3 more points than the control group.

Table 1: Number of correct answers and learning metrics for the teaching models

	Control group			Experimental group		
	Mean	**Median**	**S**	**Mean**	**Median**	**S**
Pre-test	5.15	5.00	0.99	5.86	6.00	1.10
Intermediary test	6.46	6.00	0.78	6.50	7.00	0.94
Post-test	6.54	6.00	2.15	8.07	9.00	2.20
Δ1	1.31	1.00	1.03	0.64	1.00	1.22
Δ2	0.08	0.00	2.29	1.57	2.50	2.06
Δ3	1.38	1.00	2.33	2.21	3.00	2.01

Willing to understand the impact of the teaching model on the student's learning, we analyzed the influence of the model in immediate learning. As a result, we observed that the group of students from the control group better results than that experimental group, considering that Δ1 mean was equal to 1.31 for the control group and 0.64 for the experimental one. Based on that, we understand the traditional model had a better impact on immediate learning.

Concerning Δ2, we observed that the practical activities did not impact positively the learning of the control group, as the mean value was near zero (0.08). On the other hand, these activities promoted the learning of students from the experimental group when considering the mean of 1.57 and a median of 2.5. In this sense, the number of correct answers after the practical activity of each participant increased by 1.57 on average. The median value revealed that more than 50% of the students of that group had an increase of 2.5 points. Therefore, although the practical activities accomplished by the students were the same for both groups, they favored the flipped classroom model.

Finally, Δ3 was calculated an analyzed so we could understand which group achieved the best results in terms of learning. In this analysis, we considered how much the students learned by participating in the experiment. We could observe that the Δ3 mean for the experimental group was superior compared to the control group. It indicates that on average, students of the experimental group could learn more than the control group. The median value of the experimental group is also higher than the control group.

Based on the data analysis performed, we could observe that as the interventions were performed, the learning of students increased significantly. Such a result reinforce that more interventions by the teacher can improve students' learning regardless of the teaching model being used. It should be noted that the students from both groups could achieve the same number of correct answers after the instruction.

The practical activity was essential for the students of the experimental group. It improved the students' learning by a considerable amount so they were able to achieve better scores than the control group.

4.2 Instruction time

The second variable observed throughout this experiment was the time that the students spent to study the content. During the experiment, especially in the second class, the students from the experimental group had to indicate how much time they had spent watching the video lessons to get prepared for the class. To analyze the time spent with instructions, we added 15 minutes (referring to the time spent in class to solve doubts) to the time they had indicated. Table 2 shows an overview of the results obtained.

Table 2: Descriptive statistics of the time spent by the experimental group students

	Mean	Median	S	Max	Min
Experimental group	38 minutes	40 minutes	7.52	50 minutes	22 minutes

To support the analysis, we considered the average time dedicated to giving instructions to the control group. In total, the time dedicated was 30 minutes. This time was enough for the teacher to present the black-box testing technique and get into detail about the equivalence partitioning criterion. The doubts aroused by the students during the lecture were solved on the fly.

The experimental group spent in average 38 minutes, while that the time for the control group was 30 minutes, as explained before. The minimum time spent by the experimental group shows that a student probably did not watch all videos since it took just 22 minutes while that the minimum time expected would be 30 minutes and 49 seconds. On the other hand, the max amount of time spent by the group reveals that a student took 50 minutes to finish the video lessons. The median was also observed for the instruction time variable. The resulting value indicates that more than 50% of the students spent more than 40 minutes to get prepared for the class. Based on that, we believe that the students watched the video lessons (fully or partially) more than one time or paused the video to make annotations.

... 2x video speed exists and is favoured

Finally, seeking to understand whether the students from the experimental group spend more time than the control group, we split the time spent individually by each student into two fragments. The first fragment corresponds to the students that took more than 30 minutes to study the content. The other fragment corresponds to the students that took less than 31 minutes. Figure 2 shows a pie chart with the numbers of each fragment.

We could identify that 79% of the students of the experimental group spent more time on the flipped content than the lecture given to the control group. Only 21% of the students from the experimental group took less time to study the flipped content. Although it can only be considered as a preliminary result, we believe that software testing students under the flipped classroom model spend more time with the content than the student subject to the traditional model.

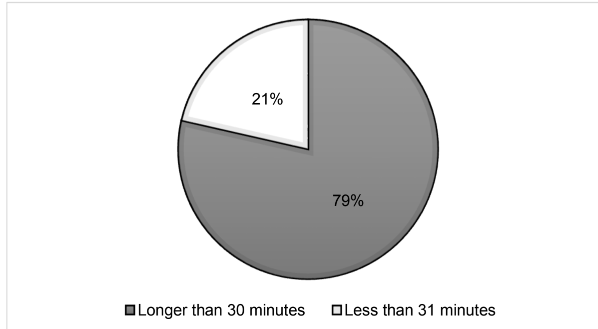

Figure 2: Time spent by the students of the experimental group

5 THREATS TO VALIDITY

Every experimental study is subject to threats that can invalidate the results. Aware of such a concern, we established strategies to mitigate eventual threats that might arise during the experiment. The threats to the validity of this study were classified accordingly to the recommendations of Cook e Campbell [12] (conclusion validity, internal validity, construct validity, and external validity). The threats and the strategies established to mitigate them are presented as follows.

Conclusion Validity: The results obtained in an experimental study may be influenced by factors such as the reliability of the measures employed, statistical tests, operation environment, and random events within the operating environment. In this sense, we took some measures to mitigate such threats. Regarding the reliability of the measures, learning was assessed through knowledge tests that evaluate the memory aspect of the student, according to the definitions of Mayer [30]. The tests have problems that need to be solved by the students. We employed this kind of test since learning occurs when the student is capable of applying the knowledge obtained solving a problem Mayer [30].

As a strategy to identify the workload of the students, we used the same measure employed by He et al. [20]. In this sense, we established that the students of the experimental group had to record the time spent (in minutes) learning with the video lessons. The time spent to solve students' doubts was added to the time provided by the students, resulting in the final value for the measure.

Regarding the statistical tests, initially, we planned to employ hypothesis testing. However, as we could not utilize all samples collected due to incompleteness, we decided to not employ hypothesis testing. Additionally, some students did not provide data from their activities to be part of the experiment analysis. In this sense, we choose only to use descriptive statistics, which gives us just a direction based on central tendency and dispersion measures.

Lastly, to mitigate random events that could occur during the class we selected classrooms in very silent areas of our institute. The students were also notified not to make much noise when entering the classroom if they were late.

Internal Validity: Variations on the variables being observed should be attributed to the treatment and not other factors. In this sense, it is important to identify factors that might interfere with this cause and effect relationship. Factors that could threat internal validity are the quality of the teaching material provided to students, the lecture given by the teacher to the control group, and the student's dropout during the experiment.

The teaching material provided to the experimental group can interfere with the results of the study, mainly affecting the dependent variables. It means that the materials should be carefully defined. For that, the material was selected among a set of other materials owned by the main researchers in this study. Then, the material selected was analyzed by three experts (two of them have more than 10 years of experience teaching and researching software testing).

During material selection, some basic requirements were considered to establish the final material and provide them to the students. We defined that video lessons should be employed to ease students' learning. Also, it should be provided in Portuguese, the native language of the students, and with a duration of fewer than 20 minutes in total, since students tend not to fully watch videos that are too long Bergmann e Sams [6].

The lecture ministered by the teacher to the group control was also a factor to be considered. A professional without experience in the area of software testing could be a negative factor to the learning acquired by the students of the control group. In this sense, the teacher selected had more than 10 years of experience teaching software testing using the traditional model. The class was planned to address the same content studied by the experimental group.

Another problem that can be a threat to the internal validity of this experiment is the dropout of students during the execution phase. To mitigate this threat, the experiment was considered as a usual course activity. This way, we notified the students that a different kind of activity would take place in the next classes to teach software testing and that their final grade would not be affected by the activity.

Construct Validity: The experiment design may offer threats to construct validity. Regarding the instruments used, we defined and utilized only one method to measure each one of the variables (learning and workload). For learning, we employed just knowledge tests. However, there are reports in the literature that show that tests are not the only approach to measure the real knowledge of the students. Other ways such as the resolution of problems, presentations, a list of exercises, and authoring can also be used. Despite that, knowledge tests were used because they are simpler to be applied and we had a limited time to execute the experiment during the course.

The application of knowledge tests can also present threats to construct validity, as the students may be able to memorize the questions and the right answers to perform better in future tests with the same questions. To mitigate this possible threat, we modified the questions, the possible answers or the order of the choices in each variation of the tests employed (pretest, intermediary test, and post-test). We tried to keep the structure and the choices of the questions when developing the variations. This action is similar to the one performed in the experimental study conducted by Paschoal et al. [33], which used tests to measure academic performance.

Regarding the metric used to measure workload, we considered the time spent by students to study the material. Although the time spent to study the material is not enough to measure workload as it does not consider the time spent to accomplish activities, solve lists of exercises, and do homework, we believe that the metric can reflect a period to some extent. Based on He et al.[20], measuring the time spent by the student is not a trivial task and more work should be done to improve the way it is done. In this sense, we observed the time spent by the students on the material as a strategy to measure their workload.

External Validity: Finally, another threat that can be present in experimental studies is the inability to generalize the experiment results to a different context in which the investigation subject and the respective variables are part of. In special, we tried to mitigate this threat selecting the right subjects and the right context. The experiment was executed with Software Engineering students enrolled in an undergraduate Computing program. Accordingly to Paschoal e Souza [34], in general, when there is not a specific course on software testing within an undergraduate program, software testing and the criteria are taught as part of software engineering courses.

6 CONCLUSION AND FUTURE WORK

This paper aimed to present an initial effort to try to understand the impact of the teaching model flipped classroom to teach the black-box testing technique, in particular, the equivalence partitioning criterion. For that, we planned and executed an experimental study to address two variables not yet investigated in this content (learning and workload). Despite these variables not being addressed in studies about software testing, they have been discussed in studies aiming to analyze the suitability of the flipped classroom model in other domains and contents.

From the results of the experiment, we observed that students subject to the flipped classroom model acquired more knowledge than the group of students that learned through the traditional model. Although these students could learn more, they spent more time to do so, as we could observe based on the data collected and analysis performed. However, such a result may not be statistically significant since our sample size is considered to be insufficient. Because of that, we intend to replicate this experiment and encourage other researchers from the software testing education area to do that too.

In order to enable other researchers to replicate this experimental study, we packaged the instruments employed such as the forms, knowledge tests, and activities, and made it available at Zenodo[2], licensed under Creative Commons. The lab package can be downloaded at: <doi.org/10.5281/zenodo.3379068>.

[2]More information at <https://zenodo.org/>.

Beyond encouraging others to replicate this study, we hope that more studies in this research topic appear in the future. There is a strong demand in the community for improvements in teaching practices and many efforts are directed towards in that way. However, there are still many papers in which the authors are adopting teaching models as alternatives to the traditional one without assessing their suitability to teach software testing. The results presented in this paper bring an interesting perspective to optimize the time spent by software testing teachers and, at the same time, bring some clues that the students can learn more than with the traditional model.

Since the flipped classroom model seeks to improve the time of the students in the classroom with the accomplishment of active activities, we believe that this model can be employed to teach other testing criteria such as the ones that require the use of automation tools. This might be an interesting research topic if er consider the reports of teachers about the limited time available to teach good testing practices with testing tools [34].

Another future research perspective on the same topic is to investigate whether students that learned software testing with the flipped classroom model can identify with precision the elements required by software testing criteria. As the elements required for testing criteria is the first step to elaborate testing cases [13], we believe this topic can be further explored. Another aspect is the quality of the testing case sets developed by the students. Based on our results, we believe that the testing case sets produced by students that learned with the flipped classroom approach have more quality when compared to students that learned traditionally. New research is encouraged to investigate our hypothesis.

Finally, an analysis of the use of the flipped classroom model in a course for its full duration is also an item to be investigated. A confront between the flipped and traditional models might offer a rich overview of the benefits and limitations from a pedagogical point of view. In this kind of analysis, it would be interesting to observe dropout rates, the number of students approved or reproved, and the performance of the students throughout the course. We highlight it would be necessary to keep the same teacher for a suitable comparison. Also, the same examples, exercises, and activities should be employed for both groups.

ACKNOWLEDGEMENTS

This study was financed by the University of São Paulo (USP) and the Brazilian funding agencies: the State of São Paulo Research Foundation - FAPESP (under processes no. 2013/07375-0 and 2017/10941-8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and CNPq (under process no. #312922/2018-3).

REFERENCES

[1] A. Amresh, A. R. Carberry, and J. Femiani. 2013. Evaluating the effectiveness of flipped classrooms for teaching CS1. In *IEEE Frontiers in Education Conference*, 733–735.

[2] G. Aşıksoy and F. Özdamlı. 2016. Flipped classroom adapted to the ARCS model of motivation and applied to a physics course. *Eurasia Journal of Mathematics, Science & Technology Education*, 12, 6, 1589–1603.

[3] T. Astigarraga, E. M. Dow, C. Lara, R. Prewitt, and M. R. Ward. 2010. The emerging role of software testing in curricula. In *IEEE Transforming Engineering Education: Creating Interdisciplinary Skills for Complex Global Environments*, 1–26.

[4] J. W. Baker. 2000. The "classroom flip": using web course management tools to become the guide by the side. In *11th International Conference on College Teaching and Learning*, 9–17.

[5] J. R. Barbosa, P. H. Valle, J. Maldonado, M. E. Delamaro, and A. M. R. Vincenzi. 2017. An experimental evaluation of peer testing in the context of the teaching of software testing. In *19th International Symposium on Computers in Education*, 01–06.

[6] J. Bergmann and A. Sams. 2012. *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education, Washington, D.C., EUA.

[7] A. Bertolino and E. Marchetti. 2014. Software testing. In *SWEBOK: Guide to the Software Engineering Body of Knowledge*. P. Bourque and R. E. Fairley, (Eds.) IEEE Computer Society, Washington, EUA, 82–103.

[8] J. Bishop and M. A. Verleger. 2013. The flipped classroom: a survey of the research. In *120th ASEE Annual Conference & Exposition*, 1–18.

[9] F. T. Chan, T. H. Tse, W. H. Tang, and T. Y. Chen. 2005. Software testing education and training in hong kong. In *5th International Conference on Quality Software*, 1–4.

[10] J. F. P. Cheiran, E. M. Rodrigues, E. L. S. Carvalho, and J. P. S. Silva. 2017. Problem-based learning to align theory and practice in software testing teaching. In *31st Brazilian Symposium on Software Engineering*, 328–337.

[11] P. J. Clarke, D. Davis, T. M. King, J. Pava, and E. L. Jones. 2014. Integrating testing into software engineering courses supported by a collaborative learning environment. *ACM Transactions on Computing Education*, 14, 3, 18:1–18:33.

[12] Thomas D. Cook and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, Boston, Massachusetts, United States.

[13] M. E. Delamaro, J. C. Maldonado, and M. Jino. 2016. Conceitos básicos. In *Introdução ao Teste de Software*. M. E. Delamaro, J. C. Maldonado, and M. Jino, (Eds.) Elsevier, Rio de Janeiro, Brasil. Chap. 1, 1–8.

[14] G. A. Dinndorf-Hogenson, C. Hoover, J. L. Berndt, B. Tollefson, J. Peterson, and N. Laudenbach. 2019. Applying the flipped classroom model to psychomotor skill acquisition in nursing. *Nursing education perspectives*, 40, 2, 99–101.

[15] A. G. O. Fassbinder, T. G. Botelho, R. J. Martins, and E. F. Barbosa. 2015. Applying flipped classroom and problem-based learning in a CS1 course. In *45th Annual Frontiers in Education Conference*, 1–7.

[16] C. O. Figueredo, S. C. Santos, P. H. Borba, and G. H. Alexandre. 2011. Using PBL to develop software test engineers. In *14th International Conference on Computers and Advanced Technology in Education*, 1–7.

[17] K. Fulton. 2012. Upside down and inside out: flip your classroom to improve student learning. *Learning & Leading with Technology*, 39, 8, 12–17.

[18] G. Gannod, J. Burge, and M. Helmick. 2008. Using the inverted classroom to teach software engineering. In *30th International Conference on Software Engineering*, 777–786.

[19] E. F. Gehringer and B. W. Peddycord III. 2013. The inverted-lecture model: a case study in computer architecture. In *44th ACM Technical Symposium on Computer Science Education*, 489–494.

[20] W. He, A. Holton, G. Farkas, and M. Warschauer. 2016. The effects of flipped instruction on out-of-class study time, exam performance, and student perceptions. *Learning and Instruction*, 45, 61–71.

[21] Y. Huang and Z. Hong. 2016. The effects of a flipped english classroom intervention on students' information and communication technology and english reading comprehension. *Educational Technology Research and Development*, 64, 2, 175–193.

[22] B. Kerr. 2015. The flipped classroom in engineering education: a survey of the research. In *18th International Conference on Interactive Collaborative Learning*, 815–818.

[23] P. N. Kiat and Y. T. Kwong. 2014. The flipped classroom experience. In *27th Conference on Software Engineering Education and Training*, 39–43.

[24] S. C. Kong. 2014. Developing information literacy and critical thinking skills through domain knowledge learning in digital classrooms: an experience of practicing flipped classroom strategy. *Computers & Education*, 78, 160–173.

[25] C. Kostaris, S. Sergis, D. G. Sampson, M. Í. Giannakos, and L. Pelliccione. 2017. Investigating the potential of the flipped classroom model in k-12 ict teaching and learning: an action research study. *Journal of Educational Technology & Society*, 20, 1, 261–273.

[26] C. Lai and G. Hwang. 2016. A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. *Computers & Education*, 100, 126–140.

[27] O. A. L. Lemos, F. C. Ferrari, F. F. Silveira, and A. Garcia. 2015. Experience report: can software testing education lead to more reliable code? In *26th International Symposium on Software Reliability Engineering*, 359–369.

[28] O. A. L. Lemos, F. F. Silveira, F. C. Ferrari, and A. Garcia. 2018. The impact of software testing education on code reliability: an empirical assessment. *Journal of Systems and Software*, 137, 1, 497–511.

[29] K. T. Lyra, S. Isotani, R. C. D. Reis, L. B. Marques, L. Z. Pedro, P. A. Jaques, and I. I. Bitencourt. 2016. Infographics or graphics+text: which material is best

for robust learning? In *16th International Conference on Advanced Learning Technologies*, 366–370.
[30] R. E. Mayer. 2002. Rote versus meaningful learning. *Theory Into Practice*, 41, 4, 226–232.
[31] K. Missildine, R. Fountain, L. Summers, and K. Gosselin. 2013. Flipping the classroom to improve student performance and satisfaction. *Journal of Nursing Education*, 15, 10, 597–599.
[32] G. J. Myers, C. Sandler, and T. Badgett. 2011. *The Art of Software Testing*. Wiley Publishing, Hoboken, Nova Jersey, EUA.
[33] L. N. Paschoal, L. R. Silva, and S. R. S. Souza. 2017. Abordagem flipped classroom em comparação com o modelo tradicional de ensino: uma investigação empírica no âmbito de teste de software. In *XXVIII Simpósio Brasileiro de Informática na Educação*, 476–485.
[34] L. N. Paschoal and S. R. S. Souza. 2018. A survey on software testing education in brazil. In *17th Brazilian Symposium on Software Quality*, 334–343.
[35] L. N. Paschoal and S. R. S. Souza. 2018. Planejamento e aplicação de flipped classroom para o ensino de teste de software. *RENOTE - Revista Novas Tecnologias na Educação*, 16, 2, 1–10.
[36] D. J. Peterson. 2016. The flipped classroom improves student achievement and course satisfaction in a statistics course: a quasi-experimental study. *Teaching of Psychology*, 43, 1, 10–15.
[37] N. Sharma, C. S. Lau, I. Doherty, and D. Harbutt. 2015. How we flipped the medical classroom. *Medical Teacher*, 37, 4, 327–330.
[38] T. J. Shippey. 2015. *Exploiting Abstract Syntax Trees to Locate Software Defects*. Ph.D. Dissertation. University of Hertfordshire, School of Computer Sciences, United Kingdom.
[39] J. Smith, J. Tessler, E. Kramer, and C. Lin. 2012. Using peer review to teach software testing. In *9th Annual International Conference on International Computing Education Research*, 93–98.
[40] A. Soska, J. Mottok, and C. Wolff. 2016. An experimental card game for software testing: development, design and evaluation of a physical card game to deepen the knowledge of students in academic software testing education. In *7th IEEE Global Engineering Education Conference*, 576–584.
[41] M. I. Syam. 2014. Possibility of applying flipping classroom method in mathematics classes in foundation program at qatar university. In *1st International Conference on Social Sciences and Humanities*, 180–187.
[42] Gang Tan. 2016. A collection of well-known software failures. (2016). http://www.cse.psu.edu/~gxt29/bug/softwarebug.html.
[43] G. Tassey. 2013. CCL GUIDE: LEARNING STORY FLIPPED CLASSROOM: What is the Flipped Classroom model, and how to use it? Tech. rep. Portugal. http://bit.ly/2KXjW5j.
[44] A. Trainor, M. M. Hayes, D. Fobert, A. Cohen, and E. D. Riviello. 2019. Effectiveness of a flipped classroom model for teaching principles of mechanical ventilation to internal medicine residents: an interim analysis. In *C40. CRITICAL CARE: THE ART OF WAR-INNOVATIONS IN EDUCATION*. American Thoracic Society, A4782–A4782.
[45] G. M. Valenzuela, O. Hinterholz Junior, W. A. Silva, M. T. Araujo, D. K. Moraes, A. A. C. Freire, and M. S. Moraes. 2013. Moodle auxiliando progressão parcial em escola pública. In *XVIII Conferência Internacional sobre Informática na Educação*, 531–534.
[46] P. H. D. Valle, E. F. Barbosa, and J. C. Maldonado. 2015. Cs curricula of the most relevant universities in brazil and abroad: perspective of software testing education. In *XVII International Symposium on Computers in Education*, 62–68.
[47] M. Wang, H. Jia, V. Sugumaran, W. Ran, and J. Liao. 2011. A web-based learning system for software test professionals. *IEEE Transactions on Education*, 54, 2, 263–272.
[48] C. Wohlin, P. Runeson, M. Hst, M. C. Ohlsson and B. Regnell, and A. Wessln. 2012. *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated, New York City, USA.
[49] H. Wood. 2016. Experiences and reflections of flipping the classroom. In *2nd Flexible Futures*, 18–26.
[50] A. Zhamanov and Z. Sakhiyeva. 2015. Implementing flipped classroom and gamification teaching methods into computer networks subject, by using cisco networking academy. In *12th International Conference on Electronics Computer and Computation*, 1–4.
[51] M. Zhivich and R. K. Cunningham. 2009. The real cost of software errors. *IEEE Security Privacy*, 7, 2, 87–90.