Journal of Software Engineering Research and Development, 2025, 13:2, doi: 10.5753/jserd.2025.4103

ADoTe: Approach to teaching and learning functional testing technique criteria supported by Testing Dojo

Vladimir Belinski [**Universidade Tecnológica Federal do Paraná** | ***vladimir_belinski@hotmail.com***]
Adolfo G. Serra Seca Neto [**Universidade Tecnológica Federal do Paraná** | ***adolfo@utfpr.edu.br***]
Maria Claudia Emer [**Universidade Tecnológica Federal do Paraná** | ***mcemer@utfpr.edu.br***]

Abstract

Context: Students of computer science and related courses (including those who have already graduated) often have knowledge gaps on software testing and low levels of motivation and interest in learning this subject. Functional testing technique criteria, for example, are usually required and used in industry, but their teaching still needs support. Faced with the problems caused by such gaps and a historical shortage of qualified testing professionals, improvements in testing education are demanded by the industry for the academy. Dojos may contribute in this context since positive results are usually reported regarding their impact on the motivation and perception of learning of students. However, their use is still little explored in the testing area. **Objective:** In this study we aimed to define and evaluate an approach supported by testing dojo to teaching and learning functional testing technique criteria in higher education. **Methods:** We defined the approach (named ADoTe) iteratively, supported by literature data and by quali-quantitative analysis of the results of two executions of a controlled experiment that evaluated the impact of ADoTe (compared to a traditional teaching approach) on the learning (measured by knowledge tests) and motivation (measured by a questionnaire named Intrinsic Motivation Inventory (IMI)) of students. **Results:** From the analysis of data collected from 44 participants we noticed that, although the averages for learning and general motivation of the groups submitted to ADoTe were greater than those of the groups submitted to a traditional teaching approach, it was not possible to state that, for these variables, there was a statistically significant difference between the approaches evaluated. However, the results of the IMI subscale related to intrinsic motivation were statistically significant, indicating that students feel more interest/enjoyment in learning functional testing technique criteria through ADoTe (the averages of the groups submitted to ADoTe were 15.80% and 19.71% higher than those of the control groups). Additionally, the results of thematic analyses carried out on the answers to a retrospective questionnaire showed that ADoTe was well accepted by students and reinforced the importance of its steps and guidelines. **Conclusion:** ADoTe positively impacts the learning and motivation of students and, compared to traditional teaching, tends to lead to greater levels of interest/enjoyment in learning functional testing technique criteria.

Keywords: *Software Engineering Education, Software Testing, Functional Testing Criteria, Testing Dojo*

1 Introduction

Despite the importance of software testing (Garousi et al., 2020b) and its presence in academic curricula (Elgrably and de Oliveira, 2021; Valle et al., 2015a), students of computer science and related courses (including those who have already graduated) often have knowledge gaps and inadequate habits regarding this subject (Melo et al., 2020; Scatalon et al., 2018; Sherif et al., 2020), as well as low levels of motivation and interest in learning it (Garousi et al., 2020b; Melo et al., 2020).

Functional testing technique criteria, for example, are usually required and used in industry (Kassab et al., 2017; Melo et al., 2022). However, according to Scatalon et al. (2018), there is still a high demand related to the teaching of this topic, since the generation of test cases based on these criteria and on artifacts used by them (such as client requirements) figures among the testing topics with the highest levels of knowledge deficiency[1]. Furthermore, considering the mistakes commonly made by students when learning software testing, Aniche et al. (2019) indicate that students typically have difficulty in creating tests using functional testing technique criteria, which reinforces the need to support the teaching of this topic more effectively.

Faced with the problems caused by such gaps and a historical shortage of qualified testing professionals (Melo et al., 2020), improvements in testing education are demanded by the industry for the academy (Garousi et al., 2020a).

In this scenario, studies have been conducted aiming to support software testing education through alternative teaching and learning approaches (e.g., gamification, problem-based learning, and flipped classroom) (Melo et al., 2020; Valle et al., 2015b). Despite the existence of these efforts, further investigation on the topic is still necessary, with the experimentation of different approaches (Valle et al., 2015b) and new pedagogical models (Elgrably and de Oliveira, 2021) being essential.

Active, collaborative, and cooperative learning strategies can help in this context since positive results are reported regarding their impact on the learning and motivation of students (Gehringer, 2007; McConnell, 2005). Due to their practical, collaborative (Sato et al., 2008), and cooperative (Rooksby et al., 2014) nature, dojos are adherent to such strategies.

In computing, a dojo is a collaborative meeting or activity

[1] The term "knowledge deficiency" indicates that the tocpic is more requested/used in industry than taught in academy, which exposes an imbalance between these environments and the need for improvements regarding the teaching of the topic (Scatalon et al., 2018).

where participants actively engage with challenges related to a specific topic, seeking to learn from each other and improve their skills (Santos et al., 2012; Sato et al., 2008). A type of dojo where the challenges are focused on software testing is named testing dojo (Gaertner, 2010), which is a possible way to learn and exercise testing topics in a practical and collaborative way (Gregory and Crispin, 2014).

Although few studies evaluate the use of dojos for teaching and learning content related to software testing in higher education, we evidenced through a Systematic Literature Mapping (SLM) that the reported results are usually positive for aspects such as motivation and perception of learning of participants. At the same time, due to methodological or statistical issues, such results cannot be generalized. Hence, there is a need for additional scientific evidence to better understand the effects and applicability of dojos in the context of testing education.

Considering that the use of dojos for teaching testing may contribute to solving the aforementioned problems, in this work we aimed to define and evaluate an approach supported by testing dojo to teaching and learning functional testing technique criteria in higher education. The choice of functional testing technique criteria as the focal topic of the approach is justified by the following reasons: (a) as already detailed at the beginning of this section, this content is usually required and used in industry (Kassab et al., 2017; Melo et al., 2022) and figures among the testing topics that most need support regarding teaching (Scatalon et al., 2018; Aniche et al., 2019); (b) since this is a fundamental content that is included in several academic curricula (Elgrably and de Oliveira, 2021; Melo et al., 2020), considering it as a focal topic may make the approach defined in this work beneficial to a greater number of students and teachers (which might not occur with other testing topics); and (c) limitations regarding the ability to validate the approach for a larger number of testing subjects in the context of this work, since it was carried out within the scope of a master's research with limited resources and time.

The approach, named ADoTe, was defined iteratively, supported by literature data and by quali-quantitative analysis of the results of two executions of a controlled experiment that evaluated the impact of ADoTe (compared to a traditional teaching approach) on the learning (measured by knowledge tests) and motivation (measured by an Intrinsic Motivation Inventory questionnaire) of students.

Therefore, this paper contributes to the improvement of software testing education by (i) defining an approach that is supported by testing dojos for the teaching and learning of functional testing technique criteria and (ii) presenting evidence regarding the applicability of the defined approach (and consequently evidence about the use of testing dojos) in the teaching and learning of functional testing technique criteria in higher education, as well as its impact on the learning and motivation of students.

The remainder of this paper is structured as follows. Section 2 presents a SLM on the use of dojos and Section 3 reviews related work. Section 4 introduces background knowledge. Section 5 describes the research methods. Section 6 defines ADoTe. Section 7 presents the study results, which are discussed in Section 8. Section 9 presents the limitations and threats to the validity of the study. Section 10 concludes this paper and identifies topics for future work.

2 A SLM on the use of dojos

To identify related work and support the definition of the approach, we conducted a SLM — following the guidelines by Petersen et al. (2015), Kitchenham and Charters (2007), and Felizardo et al. (2017) — that aimed to describe what has been investigated and reported on the use of dojos for teaching, learning, or knowledge sharing about computing-related content in higher education or professional contexts.

Subsections 2.1, 2.2, and 2.3 present the SLM protocol, SLM execution, and summary of key findings, respectively.

2.1 SLM protocol

Seeking to meet the goal presented at the beginning of Section 2, some research questions were defined: (1) How dojos have been used/organized in the context investigated? and (2) What insights/results about dojos have been reported?

Considering the SLM goal and its research questions a three-stage search strategy was used to identify primary studies:

1. Stage 1 — Primary searches (to identify keywords): we initially searched Google Scholar for "dojo AND (teaching OR learning) AND software". Based on the reading of the titles of the studies retrieved by this search, an initial set of studies was established and their titles, abstracts, and author keywords were analyzed to identify the terms commonly used in studies of interest to the SLM. Since such terms have been defined, pilot searches and iterative refinements of search strings were performed in research sources (those detailed in the next item). The resulting keywords from this process were: dojo, teaching, learning, training, and education. Furthermore, Dojo Toolkit[2] was defined as a term to be excluded from searches, since it was not related to the object of study of the SLM.
2. Stage 2 — Automatic searches: the research sources used in this stage were IEEE, ACM, Scopus, Compendex, Science Direct, and Web of Science, i.e., a combination of popular databases and search engines in the field of Software Engineering, as recommended by Dybå et al. (2007) and Felizardo et al. (2017). In the automatic searches, we looked for studies that presented the term dojo in any field, as long as the title, abstract, or keywords also included terms related to teaching and learning. More details on the execution of this stage will be presented in Subsection 2.2.
3. Stage 3 — Supplementary searches: iterations of backward and forward snowballing (Wohlin, 2014) were performed until no more relevant studies were identified. Thus, the references of the studies included in the previous stage, as well as their citations (returned in Google Scholar), were evaluated to identify possible primary

[2] Dojo Toolkit (https://dojotoolkit.org/) is a JavaScript toolkit that provides resources related to the development of web applications.

studies of interest. Details of the execution of this stage will be presented in Subsection 2.2.

Regarding the search strings used, their basic syntax was "dojo AND (teaching OR learning OR training OR education) NOT "dojo toolkit"". Searches were performed: (a) in the title, abstract, and author keywords fields for the terms teaching, learning, training, and education; and (b) in all fields for the terms dojo and dojo toolkit. Furthermore, when necessary (according to the behavior of each research source), adaptations were made to the strings, such as: (1) inclusion of the term "dojos", when the source did not automatically consider the plural of "dojo" (case of Scopus and Web of Science); (2) broadening the search for terms related to teaching and learning, when the reasearch source allowed the use of the * operator for the necessary number of terms; and (3) filtering by knowledge area and type of study, when there was no loss of possible studies of interest to the SLM. The search strings used in each research source are presented in Appendix A.1.

Aiming to select relevant studies to the SLM, the following inclusion (I) and exclusion (E) criteria were used: (I1) the study proposes or analyzes the use of dojos for teaching, learning, or knowledge sharing about computing-related content in higher education or professional contexts; (I2) the study is as a foundation work about dojos (e.g., publication usually referenced as the first on the subject or commonly used by other publications for foundation/definitions); (E1) the study is a book or is not a primary study (e.g., it is an editorial, summary of proceedings of a scientific event, tutorial, or secondary study); (E2) the study does not mention or is not related to dojos in the sense (technique, practice, approach, or meeting/event for teaching, learning, or knowledge sharing about computing-related content) and contexts investigated in the SLM (higher education or professional contexts); (E3) the study discusses dojos superficially or does not add new information to the state of the art (e.g., the term appears only in the references or in an enumeration; its use is not sufficiently detailed; it just presents results from another study also retrieved in the searches conducted for the SLM); (E4) it was not possible to obtain access to the full text of the study; (E5) the study is not written in English; and (E6) the study is a summary or an older version of another publication already considered in the SLM.

The study selection and data extraction processes were documented in an electronic spreadsheet and reviewed to ensure consistency of decisions.

2.2 SLM execution

As mentioned in Subsection 2.1, the search strategy used in the SLM was composed of three stages. This subsection details the automatic and supplementary search stages and the selection of studies.

Figure 1 presents a graphical representation of the study search and selection process. At the top of the figure are exposed the filtering steps applied during selection:

- Step 0: removal of duplicate studies.
- Step 1: initial selection of studies by reading their titles and searching for the term "dojo" in the full text. When the term "dojo" existed, the excerpts where it was used were analyzed to understand its meaning/context in the study and the granularity of the information presented.
- Step 2: selection of studies by reading their full texts.
- Application of inclusion and exclusion criteria (described in Section 2.1): occurred concurrently with steps 1 and 2.

Through **Figure 1** it can be evidenced that automatic searches were initially performed (on September 22, 2022) in the research sources indicated in Section 2.1. The retrieved studies were then analyzed through a process that involved the execution of the filtering steps described above. After filtering, the resulting studies were used as seed for supplementary searches, which were also performed in 2022 by executing iterations of backward and forward snowballing until no more relevant studies were identified (which occurred with the execution of two iterations of each type). Twenty-eight studies of interest were selected through this process.

From **Figure 1** it is also possible to notice that, in order to update the SLM, new iterations of forward snowballing were performed (on September 7, 2024) until no new relevant studies were identified (which occurred with the execution of two iterations). In this process, we analyzed the citations returned by Google Scholar for the twenty-eight studies previously selected (we considered only the citations referring to publications that had not yet been returned in the 2022 searches). Through this process, seven new studies were identified, totaling thirty-five studies of interest to the SLM. The references of these studies are exposed in Appendix A.2.

2.3 Summary of key findings

By synthesizing data from the studies selected through the process described in Subsection 2.2, we were able to understand how dojos have been used/organized (e.g., dojo structure, formats, and roles), thus answering the first research question of the SLM. This information will be detailed in Subsection 4.2 (for a better organization).

Furthermore, we could also identify positive and negative/challenging aspects reported on the use of dojos, answering the second research question of the SLM. This subsection summarizes these findings.

Regarding the positive aspects/results reported, the most frequent ones indicate that dojos:

- Favor the learning and acquisition of skills (mainly practical) (Sato et al., 2008; Bravo and Goldman, 2010; Santos et al., 2012, 2015; Aniche and Silveira, 2011; Luz et al., 2013; Heinonen et al., 2013; Estácio et al., 2015b,a, 2016b; Oliveira et al., 2016; Soomlek, 2015; Lee et al., 2017; Rodrigues et al., 2017; Elgrably and Oliveira, 2018; Santos and Oliveira, 2019; Elgrably and Oliveira, 2022b; Garcia et al., 2022; Meireles et al., 2022b,a; Garcia et al., 2023; Costa et al., 2023).
- Enhance relationships, cooperation, collaboration, teamwork, interaction, and communication among participants (Santos et al., 2015; Luz et al., 2013; Estácio et al., 2015b, 2016b; Soomlek, 2015; Lee et al., 2017; de Oliveira et al., 2018; Fonseca et al., 2019; Meireles et al., 2022b,a; Garcia et al., 2023).

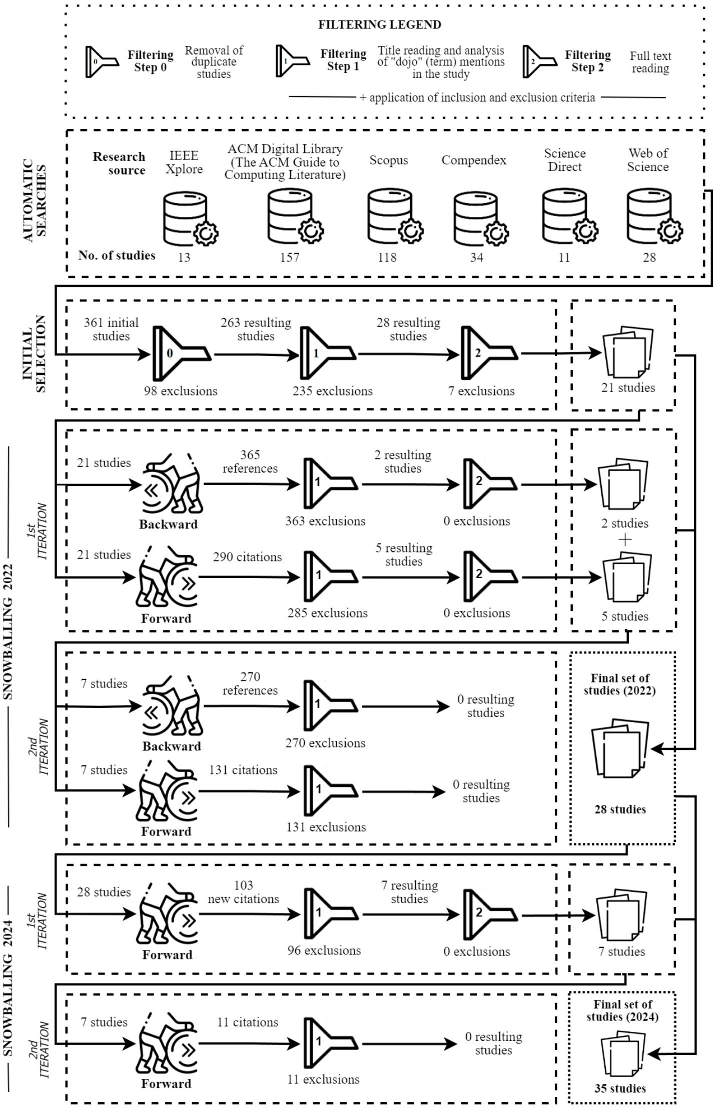


Figure 1. Search and selection of studies

- Increase the motivation, interest, participation, engagement, or commitment of participants (Luz et al., 2013; Estácio et al., 2015a; Soomlek, 2015; Lee et al., 2017; Rodrigues et al., 2017; Garcia et al., 2022; Meireles et al., 2022b,a; Garcia et al., 2023; Costa et al., 2023).
- Favor knowledge sharing (Sato et al., 2008; Santos et al., 2012; Luz et al., 2013; de Oliveira et al., 2018; Santos and Oliveira, 2019; Meireles et al., 2022b,a).
- It is considered as a fun/enjoyable activity by participants (Santos et al., 2015; Aniche and Silveira, 2011; Estácio et al., 2015b, 2016b).

Furthermore, studies also mention that dojos are a good option to be used in academic contexts (Sato et al., 2008; Luz et al., 2013; Soomlek, 2015; Rodrigues et al., 2017; Elgrably and Oliveira, 2020, 2022b), would be an interesting way to complement traditional methods, and could be used in teaching and training programs (Sato et al., 2008; Luz et al., 2013).

In turn, the main negative/challenging aspects were:

- Complexity of katas (also known as exercises, challenges, or missions): Santos et al. (2012), Estácio et al. (2016b), and Elgrably and Oliveira (2022b) mentioned that some participants considered the katas complex. Thus, it is important to elaborate katas with an adequate level of complexity.
- Size of baby steps: Aniche and Silveira (2011) reported that participants did not feel they were having progress in the activity, as their steps were very small. This indicates that attention should be paid to the size of the steps used to solve the katas.
- Cycle/iteration time: in some studies, the time to work on a kata (usually five to seven minutes) was considered insufficient by participants (Santos et al., 2012; Heinonen et al., 2013; Estácio et al., 2015b; Lee et al., 2017; Rodrigues et al., 2017; Meireles et al., 2022b), while in Filho and de Toledo (2015) it is reported that ten minutes is a long time for the same person at the keyboard. In Estácio et al. (2016b) the use of fixed time cycles was perceived as negative, while in Sato et al. (2008) it was understood as necessary. Furthermore, in Furtado and Oliveira (2020) it was reported a lack of balance between cycle time and the difficulty of katas. Thus, cycle time and its balancing with the size/complexity of katas is a challenging aspect.

- Number of participants: in the Randori format (see Subsection 4.2.4), sessions with large groups can be problematic, since a huge number of participants slows down the dynamics and decision-making (Sato et al., 2008; Estácio et al., 2016b). Moreover, in these scenarios participants spend a lot of time as observers, which can bore them (Sato et al., 2008) and generate situations where dojo sessions are finished without everyone having assumed the role of pilot at least once (Rodrigues et al., 2017). Thus, it is necessary to adapt the dynamics for larger groups (Santos et al., 2012) or use the Kake format (Rodrigues et al., 2017). At the same time, according to Estácio et al. (2016b), conducting dojos with very small groups (e.g., with only three participants) is also not interesting.
- Other challenges and points of attention: according to Rodrigues et al. (2017), the exposure that occurs in dojos (referring to performing in front of other people) can cause discomfort in some participants; Heinonen et al. (2013) and Rodrigues et al. (2017) point to the need to take care so that competition does not occur in dojos, since they must be collaborative; de Oliveira et al. (2018) presents that, for the dojo to be effective, it is necessary to carry out prior planning and follow the established dojo rules; and in Santos et al. (2012) and Estácio et al. (2015b) the participants complained about the lack of visibility of the code, which indicates that attention should be paid to infrastructure issues.

Specific results relating dojos and software testing will be presented in Section 3.

3 Related work

Some studies identified in the SLM described in Section 2 approach or at least mention the possibility of teaching or learning software testing related content in higher education through dojo sessions (Sato et al., 2008; Luz et al., 2013; Heinonen et al., 2013; Estácio et al., 2016a,b; Lee et al., 2017; Rodrigues et al., 2017; de Oliveira et al., 2018; Elgrably and Oliveira, 2018, 2020, 2022a,b; Costa and Oliveira, 2022; Costa et al., 2023) and, therefore, can be considered as related work.

Regarding the testing related content explored or mentioned in these studies as possible to be taught or learned through dojos, Test-Driven Development (TDD) is the most common (Sato et al., 2008; Luz et al., 2013; Heinonen et al., 2013; Estácio et al., 2016a,b; Lee et al., 2017; Rodrigues et al., 2017; de Oliveira et al., 2018; Elgrably and Oliveira, 2018, 2020, 2022a). Other topics also explored or mentioned are: unit testing (Luz et al., 2013; Rodrigues et al., 2017; de Oliveira et al., 2018; Elgrably and Oliveira, 2022b,a), exploratory testing (Elgrably and Oliveira, 2020; Costa and Oliveira, 2022; Costa et al., 2023), test automation (de Oliveira et al., 2018), Behavior-Driven Development (BDD) (Sato et al., 2008), test coverage (Sato et al., 2008), and code coverage (Lee et al., 2017). This indicates that a variety of topics can be approached in dojos, but also demonstrates that some testing topics have not yet been explored.

Considering studies that presented specific results on the approach of software testing related content in dojo sessions conducted in academic contexts: (a) Elgrably and Oliveira (2022a) focused on presenting the perception of professors about the use of active methodologies (including dojo) to teach software testing remotely; (b) Luz et al. (2013), Estácio et al. (2016b), and Lee et al. (2017) had a greater emphasis on teaching and learning of TDD; and (c) Sato et al. (2008), Heinonen et al. (2013), Rodrigues et al. (2017), and de Oliveira et al. (2018) exposed results about testing, but without focusing on the subject.

Still considering the publications of the previous paragraph: (a) in most of them, the evaluation methods were non-experimental. Estácio et al. (2016b) used Grounded Theory for data analysis. In the remaining studies dojo sessions were usually conducted and opinions/perceptions from participants or authors (about the dojo and, sometimes, its impact on learning and motivation) were collected (through questionnaires, interviews, or observation) and analyzed quantitatively and qualitatively; and (b) Lee et al. (2017) was the only one in which controlled experiments were conducted and the results (regarding code coverage and TDD) obtained from dojos were compared to another approach (individual programming). However, Lee et al. (2017) presented results simplistically, and it was unclear whether their statistical significance was assessed.

Regarding learning results, considerations usually positive were presented about the impact of dojos on the learning of unit testing (Rodrigues et al., 2017), test automation (de Oliveira et al., 2018), test coverage (Sato et al., 2008), code coverage (Lee et al., 2017), and TDD (Sato et al., 2008; Luz et al., 2013; Heinonen et al., 2013; Estácio et al., 2016b; Lee et al., 2017; Rodrigues et al., 2017). Positive results were also reported for student motivation, interest, participation, and engagement (Luz et al., 2013; Lee et al., 2017; Rodrigues et al., 2017).

In turn, regarding teaching results (less common in studies), Luz et al. (2013) and Heinonen et al. (2013) mention that the coding dojo activity favors the teaching of TDD. Additionally, considering the use of the dojo formats known as Kake and Randori for teaching testing in remote contexts, Elgrably and Oliveira (2022a) present that most professors did not considered this approach difficult to construct a subject, evaluate student participation, make students interact, and collect their feedback. However, some professors indicated that in remote contexts: dojo interactions may be hampered, that the activity requires a high degree of planning, and that it may be difficult to provide mentoring (in Kake) and organize the practice if there are many students (in Randori).

Our work differs from the aforementioned by presenting an approach to teaching and learning functional testing technique criteria supported by testing dojo. To the best of our knowledge, our study is the first to investigate the teaching and learning of functional testing technique criteria using dojos. Furthermore, this study also differs from most of the previous ones by conducting a controlled experiment (comparing the proposed approach to traditional teaching) and verifying the statistical significance of the results. More differences between our approach and solutions from other studies will be exposed in Section 6, after the presentation of ADoTe.

4 Background

In this section, we provide background information related to software testing concepts, dojo, and testing dojo.

4.1 Software testing concepts

According to Myers et al. (2012), testing is the process of executing a program aiming to find errors. To ensure through testing that there are no defects in a program P, the authors point out that P should be executed with all elements of its input domain $D(P)$. However, due to the cardinality of $D(P)$ (usually very large or infinite), this approach becomes impractical, making it necessary to identify and use for the testing of P only a reduced subset of $D(P)$, that requires the minimum amount of time and effort, but has a high probability of uncovering defects (Myers et al., 2012).

In this context, the definition of test subdomains is essential. As presented by Weyuker and Ostrand (1980), a test/revealing subdomain consists of a subset of $D(P)$ with similar test data, i.e., for which the same behavior of P is expected. The idea is that a test set composed of one element from each subdomain would be remarkably reduced compared to $D(P)$, but in a certain way would represent it entirely (Weyuker and Jeng, 1991).

To identify subdomains, rules known as testing criteria are established. Testing criteria usually define requirements to decide when test data should or should not be in the same subdomain and are grouped into techniques differentiated by the source or type of information used to establish the subdomains (Delamaro et al., 2006). The functional technique and its criteria are usually the most used in the software industry (Kassab et al., 2017; Melo et al., 2022) and also the most included in academic curricula (Elgrably and de Oliveira, 2021; Melo et al., 2020).

Myers et al. (2012) state that in the functional technique P is considered a black box, since test data are established based only on its specifications/requirements, without considering implementation details or the internal structure of the program. Thus, for the authors, this technique focuses on finding circumstances in which P does not behave according to its specifications. Non-exhaustive examples of functional testing technique criteria are equivalence partitioning (Myers et al., 2012), boundary value analysis (Myers et al., 2012), and systematic functional testing (Linkman et al., 2003).

4.2 Dojo and testing dojo

The first publications about dojos in computing define them as periodic meetings (called sessions) where programmers with different knowledge levels come together to solve small exercises (called katas) collaboratively, seeking to learn from each other and improve their skills through deliberate practice (Bossavit and Gaillot, 2005; Sato et al., 2008).

Dojos are fundamentally social and cooperative (Rooksby et al., 2014). The idea behind them is to create a safe and fun environment where people can test new ideas and learn continuously, in a collaborative, inclusive, and non-competitive way (Sato et al., 2008). Additionally, dojos would be aligned with principles such as: (a) failure: it is acceptable to fail when learning something new; (b) redundancy: new understandings can be obtained by approaching the same problem using different strategies; and (c) Baby Steps: each step towards solving the kata must be small enough so that everyone can understand it and replicate it later (Sato et al., 2008).

Although dojo was originally defined as a periodic meeting focused mainly on the practice and knowledge sharing of programming, the dissemination of the proposal of Bossavit and Gaillot inspired the creation of new types of dojos. Among the existing variations, the one known as testing dojo is focused on software testing. According to Gaertner (2010), in this type of dojo participants practice their skills in groups, working on test missions/katas that can vary from testing an application, evaluating tools, learning new testing approaches, etc. Thus, this type of dojo consists of a collaborative way to train new professionals in the area and disseminate knowledge about testing (Gaertner, 2010).

Characteristic dojo/testing dojo elements will be presented in the next sections.

4.2.1 Infrastructure

The infrastructure required for a dojo session depends on the activities that will take place during it (Bache, 2013). Despite this, certain elements are usually mentioned in the literature (Bache, 2013; Gaertner, 2010; Sato et al., 2008): (a) a room large enough to accommodate the participants; (b) at least one computer, where katas will be solved; (c) a projector, to display in the room what is being done on the computer; and (d) materials for taking notes and supporting discussions (e.g., pen, paper, whiteboard, and flip chart).

4.2.2 Roles

In dojo sessions, participants play roles, learning from their direct action, observation, and interactions (Heinonen et al., 2013). The common roles in dojos are outlined below.

Facilitator. The person playing this role (also named moderator, organizer, presenter, Dojo Master, or Sensei) is responsible for organizing and facilitating the dojo, usually acting: (a) before the session, adjusting infrastructure issues; (b) during the session, ensuring that dojo rules are being followed, moderating discussions, presenting examples related to katas, and conducting retrospectives; and (c) after the session, organizing the room (Sato et al., 2008; Gaertner, 2010; Heinonen et al., 2013). Although their main activity is to facilitate the dojo, the participant who acts as a facilitator can also play other roles during the session (Gaertner, 2010). In groups that are more experienced in running dojos, this role may not exist (Bache, 2013).

Pilot/driver/tester. Participant who is responsible (at a given moment) for solving the kata, being the only person who can use the keyboard and mouse (Rodrigues et al., 2017). While working on the kata, the pilot must describe his mental model (e.g., ideas and justifications about what to do or not do) to the other participants so that everyone can understand what is being done and why (Gaertner, 2010).

Copilot/navigator/recorder. Works alongside the pilot, providing support through suggestions, code inspections, and making notes about the activity (e.g., test ideas and defects found) and the steps taken on it, ensuring that these can be reproduced (Gaertner, 2010; Rodrigues et al., 2017).

Audience/public/observers. Participants playing this role must follow the work of the pilot and copilot (the solution they are creating and their interactions) in an engaged manner, being allowed at certain moments to ask questions to the pair, provide them support, feedback, and improvement suggestions (Gaertner, 2010; Heinonen et al., 2013).

4.2.3 Structure/moments

The structure of a dojo session is not consensual in the literature (Rodrigues et al., 2017). However, some moments are typically mentioned and are presented below in the order in which they usually occur.

Dojo introduction. The purpose of this moment is to help participants feel safe to experiment and learn. To this end, the facilitator can introduce them to the concept of dojo, its structure, the rules of the format that will be used in the session (see Section 4.2.4), as well as the expected behaviors and attitudes from the participants. The introduction can take from two to fifteen minutes, depending on the level of prior knowledge of the group about dojos (Bache, 2013).

Initial agreements. It usually includes a two to ten-minute discussion between the facilitator and the other participants about: (a) the date, time, and location of the next meeting (this definition can also be made at the end of the session); (b) a brief discussion about what happened in the last meeting; and (c) the definition of what will be accomplished in the current session (e.g., choice of kata) (Bache, 2013; Rodrigues et al., 2017).

Execution. This moment may include (Sato et al., 2008): (a) a few initial minutes (ten to twenty) to discuss the different approaches that can be used to solve the kata, the choice of which one will be adopted, and the creation of a script to guide participants during the activity; and (b) a longer period (one to two hours) to work on solving the kata following the dynamics of a dojo format.

Retrospective/reflection. At this moment — that usually takes five to fifteen minutes (Bache, 2013) — participants discuss/reflect on what happened in the session and what they learned from it (Sato et al., 2008). To encourage discussion, each participant can be asked to mention something they learned, something that surprised them, and something they still don't understand (Bache, 2013) or to answer questions such as "What went well?", "What could be improved?", "What have we learned?" and "What has hindered learning?" (Sato et al., 2008). From this discussion, the group can make decisions to improve the dojo (Gaertner, 2010).

4.2.4 Formats

Different dynamics (usually called formats) can be used while working on a kata. The most common in the literature are described below.

Kata/Prepared Kata. **Figure 2** illustrates the dynamics of Kata/Prepared Kata, a format where the facilitator prepares the kata solution before the session (Bache, 2013).

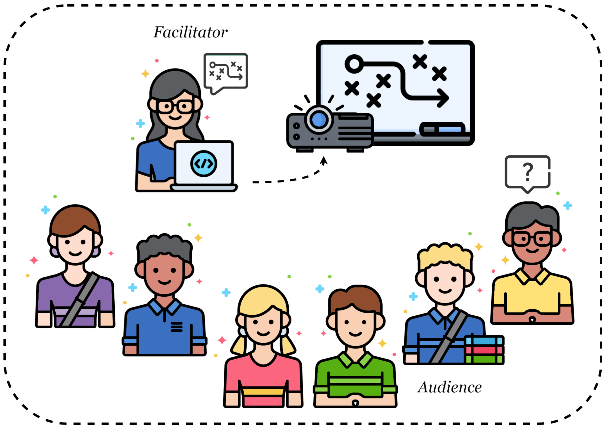


Figure 2. Dynamics of the Kata/Prepared Kata format

During the session, the facilitator shows the kata solution step by step to the audience, reproducing it from scratch and exposing his/her mental model (e.g., the decisions made throughout the process and their motivation), so that everyone is capable of understanding it (note that in **Figure 2** the facilitator is explaining the solution being created, which is also projected in the room). In turn, the audience must ensure that they understand what is being presented and can ask the facilitator for explanations when they have questions (Rodrigues et al., 2017). The objective is that, at the end of the session, everyone will be able to reproduce the steps performed and solve the kata on their own at another time (Bache, 2013; Sato et al., 2008).

Randori/Paired Session. The basic dynamics of a dojo session in this format are illustrated in **Figure 3**.

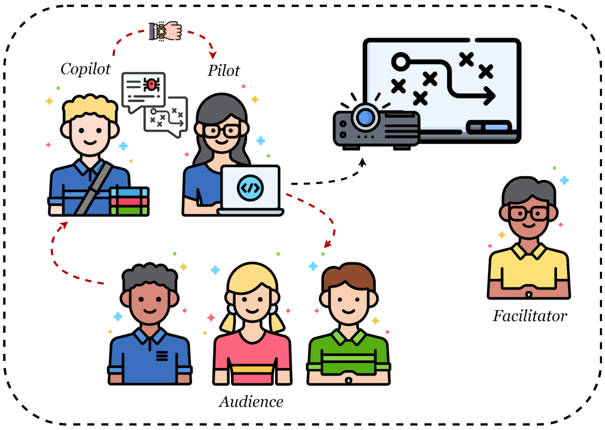


Figure 3. Dynamics of the Randori format

According to Bache (2013) and Rodrigues et al. (2017), in this format a single computer is used, where the kata is solved and whose screen is projected in the room. At the start of the session, a pair (pilot and copilot) is chosen to begin solving the kata and a pair exchange strategy is defined. During execution, the pilot controls the keyboard and mouse and is accompanied by a copilot, who assists him in inspecting what is being created and giving suggestions. Pilot and copilot must interact and explain what they are doing so that everyone can follow and understand their actions. In turn, the audience must pay attention to the actions of the pair and provide support when requested (usually when the pair does not know what to do). The facilitator performs the activities presented in Section 4.2.2 (e.g., ensures that the dojo rules are being followed, moderates discussions, and presents examples/tips when necessary).

To ensure that participants can play different roles in the same session, moments of pair exchange are held. Timebox (represented in **Figure 3** by the clock and red dotted arrows) is a common strategy for exchanging pairs: each pair works on the kata for a short period (five to ten minutes); at the end of this time, the pilot returns to the audience, the copilot starts to act as pilot, and a participant from the audience joins the new pilot and starts to act as copilot (Gaertner, 2010).

Finally, according to Bache (2013), the Randori format is more suitable for groups of four to ten participants since a larger number can make discussions extensive and reduce the time each participant can work as a pilot.

Kake. In this format participants are organized into multiple pairs and work simultaneously on the same kata on different computers. Due to this characteristic, Kake eliminates a problem that can occur in Randori when a session has many participants: the idleness of the audience for long periods (Rodrigues et al., 2017).

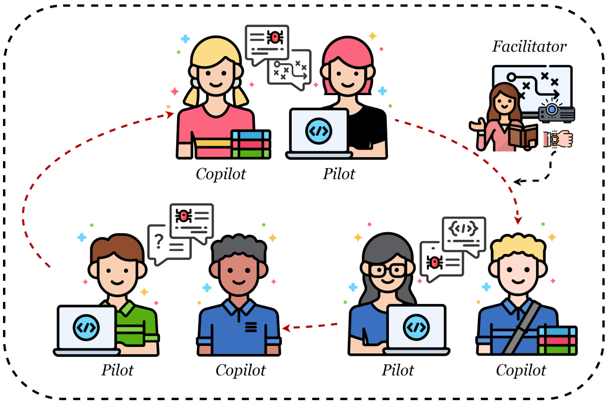


Figure 4. Dynamics of the Kake format

Figure 4 illustrates the dynamics of the Kake format, where it is possible to notice the existence of multiple pairs. The roles of pilot and copilot are maintained (as in Randori), but the cycle of interactions (which occur in the pair during the cycle and lasts a few minutes) is a little different: at the end of each round, participants who played the role of pilot become copilots in another pair; copilots stay at the same workstation, but now working as pilots; participants are encouraged to form new pairs each round. The facilitator, when existing, performs the functions described in Section 4.2.2.

5 Methods

This section describes the research methods, including the study goal, research questions, and details about the definition and evaluation of ADoTe.

5.1 Goal and research questions

As previously stated, this work aims to define and evaluate an approach supported by testing dojo to teaching and learning functional testing technique criteria in higher education. We seek answers to the following research questions (RQs):

RQ_1: How to define an approach[3] supported by testing dojo to teaching and learning functional testing technique criteria in higher education? Through this RQ we seek to determine the elements (e.g., moments, dynamics, and guidelines) the new approach must have and the relationship between them, aiming to maximize the approach results in terms of supporting the teaching and learning of functional testing technique criteria.

RQ_2: Is the learning of students who study functional testing technique criteria through an approach supported by testing dojo greater than the learning of those who study the same content through a traditional teaching[4] approach? This RQ aims to evaluate the impact of the new approach on student learning. The results may indicate the applicability (or not) of the approach and how much it differs from traditional teaching (which is usually used in higher education for teaching this content) regarding this variable.

RQ_3: Is the motivation of students who study functional testing technique criteria through an approach supported by testing dojo greater than the motivation of those who study the same content through a traditional teaching approach? This RQ aims to evaluate the impact of the new approach on student motivation. Since students often report low motivation and interest in learning software testing, better results for the new approach may indicate that it helps to solve this problem.

The approach definition (related to RQ_1) and evaluation (related to RQ_2 and RQ_3) processes are presented in sections 5.2 and 5.3, respectively.

5.2 Approach definition

To answer RQ_1, we first sought to understand what already existed in the literature regarding dojos in computing, to identify positive and negative aspects of the topic, so that these could be considered when defining the approach. To this purpose, we conducted the SLM detailed in Section 2.

The analysis of the studies identified in the SLM made it possible to understand how dojos have been used for teaching and learning computing topics (e.g., dojo structure, formats, and roles; exposed in Subsection 4.2). Furthermore, the synthesis of the reported results allowed the identification of both positive and negative/challenging aspects related to dojos (summarized in Subsection 2.3).

Considering the SLM findings and literature information (especially from foundation works and references identified through the studies from the SLM of Section 2) on software testing and active, collaborative, and cooperative learning, it was established which elements the approach should contain

[3] In this study, the term approach refers to the way or method used to plan, implement, and evaluate a teaching and learning process.

[4] Traditional teaching is typically defined as an approach in which the student passively receives information from an instructor (Prince, 2004). This approach is also known as conventional teaching.

(see Section 6) and then defined ADoTe. Moreover, the analysis of the results of a controlled experiment (see Section 5.3) made it possible to evaluate and improve the approach. The final version of ADoTe is presented in Section 6.

5.3 Approach evaluation

According to Wohlin et al. (2012), in Software Engineering new proposals must be not only suggested but also compared to existing ones. Experimentation makes this possible, as it "provides a systematic, disciplined, quantifiable, and controlled way of evaluating human-based activities" (Wohlin et al., 2012, p. 5). Thus, to answer RQ_2 and RQ_3 two executions of a controlled experiment were conducted.

The experiment was planned by the first author and, before its application, validated and reviewed by the second and third authors, two professors with experience in Software Engineering teaching and research (including software testing and dojos). Two executions were conducted because, aiming to enrich the results of the study, (a) we sought to obtain a larger sample (number of participants), which would not be possible in the context of the research by conducting just one execution and (b) to validate the cyclical structure of ADoTe (detailed in Section 6) we wanted to consider the outputs from the first execution (which, in a certain way, can also be considered a pilot study, since we were unable to conduct a prior execution to those described in this work, especially due to time limitations and definitions of the ethical project) in a subsequent execution.

The decisions related to the experimentation process are presented in sections 5.3.1 to 5.3.3. The experimental package (which contains all the instruments mentioned in the next sections and the detailed description of treatments) is available in an open repository[5].

Regarding ethical aspects, this study was evaluated and approved by a Research Ethics Committee. Approval can be verified on Plataforma Brasil[6] by consulting the Certificate of Presentation for Ethical Assessment (CAAE) number 69036823.2.0000.5547.

5.3.1 Experimental design

This section describes the outcome of the scoping and planning stages.

Goal. The goal of the experiment was: *analyze* ADoTe and a traditional teaching approach, *for the purpose of* evaluation, *with respect to* student learning and motivation, *from the point of view of the* researcher, *in the context of* higher education students (from Computer Science or related courses) who are studying functional testing technique criteria.

Context and subjects. Two executions of the experiment were conducted, both in a real teaching and learning context: in the classroom, with higher education students (legally capable and aged 18 or older) who were enrolled in a subject in which functional testing technique criteria are often taught.

The technique used for subject selection was convenience sampling. Participants were recruited from two Software Engineering classes (2023/1 and 2023/2, i.e., from different semesters), both under the responsibility of the third author of this work and offered in an Information Systems course at a Brazilian federal university. Students who agreed to participate in the study and met the criteria set out in the previous paragraph were admitted.

Hypotheses. For each research question (RQ_2 and RQ_3) were formulated at least: (a) a null hypothesis (H_0), assuming the absence of a significant difference (to a given effect under study) between ADoTe and the traditional teaching approach; and (b) an alternative hypothesis (H_1), assuming the existence of a significant difference between the approaches.

The hypotheses for RQ_2 were:

$$2H_0 : \mu_{\text{Learning}_{\text{ADoTe}}} = \mu_{\text{Learning}_{\text{Traditional}}}$$
$$2H_1 : \mu_{\text{Learning}_{\text{ADoTe}}} \neq \mu_{\text{Learning}_{\text{Traditional}}}$$

To measure learning, two knowledge tests were applied: one before (pre-test) and another after (post-test) participants were submitted to the approaches. The final scores of students on a given test were equal to their number of correct answers. In turn, the learning of students was calculated as the difference between their final scores in the post-test and pre-test. Finally, μ_{Learning} corresponded to the mean of the learning values obtained by the students submitted to a certain approach.

The hypotheses for RQ_3 were:

$$3H_0 : \mu_{\text{Motivation}_{\text{ADoTe}}} = \mu_{\text{Motivation}_{\text{Traditional}}}$$
$$3H_1 : \mu_{\text{Motivation}_{\text{ADoTe}}} \neq \mu_{\text{Motivation}_{\text{Traditional}}}$$

To measure motivation, a questionnaire named Intrinsic Motivation Inventory (IMI) was used. IMI is a multidimensional measuring instrument that evaluates the subjective experience of participants in experiments concerning a given activity (CSDT, 2023; McAuley et al., 1989). IMI was chosen due to (a) the absence of a psychologist to measure the motivation of each participant during the experiment, (b) the reliability of the instrument, since it was already validated in several studies in the field of psychology (CSDT, 2023), (c) the malleability of its items (which allows its adaptation to different types of activities) and simplicity in constructing the questionnaire (CSDT, 2023; McAuley et al., 1989), and (d) its use in other research on software testing education, such as the study by Jesus et al. (2019).

According to CSDT (2023), the IMI is composed of items/statements organized into subscales. The IMI questionnaire used in this work covered five[7] subscales: (1) interest/enjoyment (the self-report measure of intrinsic motivation, an important type of motivation in educational contexts; assesses the interest/enjoyment of participants in doing the activity); (2) perceived competence (positive predictor of intrinsic motivation; evaluates how competent the participants perceived themselves when carrying out the activity); (3) pressure/tension (negative predictor of intrinsic motivation; measures how pressured/tense participants felt while

[5] https://zenodo.org/doi/10.5281/zenodo.10510437

[6] https://plataformabrasil.saude.gov.br

[7] The full version of IMI has seven subscales. In this study, we chose not to include the subscales known as perceived choice (since participation in the experiment was optional and, therefore, all participants chose to do the activities) and relatedness (since the validity of this subscale has yet to be established). This type of adaptation is allowed and valid, as reported in (CSDT, 2023).

performing the activity); (4) effort/importance (assesses the perception of applied effort of participants, how important it was to themselves to perform the activity well); and (5) value/usefulness (evaluates the perception of participants about the value/usefulness of the activity). In this work, we used all the items/sentences from the subscales above, maintaining their original descriptions from the version of interest to this study, the "The post-experimental Intrinsic Motivation Inventory" (CSDT, 2023).

Motivation results were calculated according to CSDT (2023) guidelines, as follows: (a) participants evaluated each questionnaire item using a 7-point Likert scale (values from 1 to 7); (b) the score of an item was the value attributed to it by the participant, except for items with reverse meaning, whose score corresponded to the modulus of the outcome of subtracting the value chosen by the participant from the constant 8 (e.g., the score of a reverse item valued 2 was $|2 - 8| = 6$); (c) the result of a participant for a given IMI subscale corresponded to the mean of the scores obtained for its items; and (d) the motivation of a participant was defined as the sum of the subscales results (inverting pressure/tension, as it is a negative predictor of motivation).

Since the results for each subscale were calculated separately, as described in step (c), it was also possible to test hypotheses for subscales individually, allowing a better understanding of the results. Therefore, specific hypotheses were defined for each of the factors considered in the IMI, as presented in **Table 1**.

Finally, to test $3H_0$ and $3H_1$, $\mu_{\text{Motivation}}$ corresponded to the mean of the motivation values (i.e., $\mu_{\text{INT}}+\mu_{\text{CMP}}-\mu_{\text{PRS}}+\mu_{\text{EFF}}+\mu_{\text{VAL}}$) from students submitted to a certain approach (AdoTe or traditional).

Variables and design type. The variables of an experiment can be: (a) independent: also known as input variables, they must be controllable and have some effect on the dependent variables; or (b) dependent: also known as output variables, they are affected by the independent variables and represent the effect/result of the experiment.

The independent variables of this study were:

- *Teaching and learning approach*: corresponds to what was performed and used (e.g., practices, artifacts, and interactions) to promote the teaching and learning of functional testing technique criteria. This variable is directly related to the experiment design: one factor (the approach used for teaching and learning) with two treatments (listed below and detailed here[5]). Furthermore, a completely randomized design was used, since the subjects were assigned to the same objects (e.g., content and exercises), but randomly to a single treatment:
  - Treatment A (applied to the experimental group): teaching and learning occurred through ADoTe.
  - Treatment B (applied to the control group): teaching and learning occurred through a traditional teaching approach.
- *Academic Performance Coefficient (APC)*[8]: since students with higher APC tend to obtain better learning results, in each execution of the experiment students were randomly assigned — following the principles of randomization, blocking, and balancing (Wohlin et al., 2012) — to two groups (experimental and control) balanced regarding the APC of their members and the number of participants.

The dependent variables were:

- *Learning*: corresponds to how much the participants could learn from the treatment to which they were submitted. As previously stated, this variable was measured through knowledge tests. Since the learning of a student was calculated as the difference between his/her final scores (0 to 10) in the post-test and pre-test, learning values may vary from -10 to 10 (worst and best results).
- *Motivation*: the level of motivation of the participant regarding the activities performed in the experiment. As motivation was measured using the IMI questionnaire, the aspects assessed by the subscales of this instrument can also be considered dependent variables. For each IMI subscale, results may vary from 1 to 7 (the closer to 7 the better; except for PRS, which has the opposite behavior). In turn, since the motivation of a participant was defined as the sum of the subscales results (inverting pressure/tension), motivation may vary from -3 to 27 (worst and best values, respectively).

Instrumentation. The objects, guidelines, and measurement instruments used were:

- *Free and Informed Consent Form (ICF)*: used to describe the research to the recruited subjects and collect the consent of participants about their collaboration.
- *Participant characterization questionnaire*: used to characterize the profiles of the participants based on six closed questions about their age, gender, course, professional experience, and prior level of knowledge about functional testing technique criteria.
- *Retrospective questionnaire*: composed of three open questions, it was used to collect the perceptions of participants about what went well or not in the activity (favoring or hindering their learning), as well as improvement suggestions.
- *Instrument for measuring motivation*: as already mentioned, an IMI questionnaire was used.
- *Instruments for measuring learning*: as previously stated, knowledge tests were applied: a pre-test (to assess the prior knowledge of participants about the content) and a post-test (to assess the knowledge of participants after their submission to an approach). Both tests comprised ten multiple-choice questions and evaluated theoretical and practical knowledge. To reduce biases that could occur due to a greater number of correct answers as a result of the random choice of answers, each question presented an escape alternative (e.g., "I don't know") and participants were instructed to choose it if they did not know the answer.
- *Materials/objects used to teaching and learning*: supported the teaching and learning process and varied according to treatments. They included:

[8] APC is an index used by the university where the experiment took place to measure the overall academic performance of students on the course.

Table 1. Hypotheses associated to the IMI subscales

IMI subscale	Null hypothesis (H_0)	Alternative hypothesis (H_1)
Interest/enjoyment (INT)	$3.1H_0 : \mu_{\text{INT ADoTe}} = \mu_{\text{INT Traditional}}$	$3.1H_1 : \mu_{\text{INT ADoTe}} \neq \mu_{\text{INT Traditional}}$
Perceived competence (CMP)	$3.2H_0 : \mu_{\text{CMP ADoTe}} = \mu_{\text{CMP Traditional}}$	$3.2H_1 : \mu_{\text{CMP ADoTe}} \neq \mu_{\text{CMP Traditional}}$
Pressure/tension (PRS)	$3.3H_0 : \mu_{\text{PRS ADoTe}} = \mu_{\text{PRS Traditional}}$	$3.3H_1 : \mu_{\text{PRS ADoTe}} \neq \mu_{\text{PRS Traditional}}$
Effort/importance (EFF)	$3.4H_0 : \mu_{\text{EFF ADoTe}} = \mu_{\text{EFF Traditional}}$	$3.4H_1 : \mu_{\text{EFF ADoTe}} \neq \mu_{\text{EFF Traditional}}$
Value/usefulness (VAL)	$3.5H_0 : \mu_{\text{VAL ADoTe}} = \mu_{\text{VAL Traditional}}$	$3.5H_1 : \mu_{\text{VAL ADoTe}} \neq \mu_{\text{VAL Traditional}}$

- For both groups: a video and a set of slides about testing concepts and functional testing technique criteria, specifically equivalence partitioning, boundary value analysis, and systematic functional testing. These three criteria were chosen because the first two are the most used in Brazilian companies (Melo et al., 2022), and, together with the third one, they are also the most approached in higher education computing-related courses offered around the world (Melo et al., 2020).
- For the experimental group: an online form with katas related to the three testing criteria mentioned above, a dynamic online presentation (with explanations about testing dojos and resources that allowed participants to share their responses to the katas), and a development environment.
- For the control group: a list of exercises (the same organized as katas for the experimental group) and an answer sheet (containing detailed solutions to the exercises, written by the researchers).

Regarding the selection of the dojo formats and katas used in the experimental group:

- The three dojo formats mentioned in Subsection 4.2.4 were used in the experiment because (a) we considered that the dynamics of Prepared Kata (used to solve the first activity) would be the most appropriate to introduce the participants to the content and the dojo itself, since this format is simpler, closer to the teaching-learning model that students were used to, and expose them less to their peers (which could minimize possible discomforts and help them to build confidence for other dojo formats). However, as using only Prepared Kata could make the activity monotonous (since the role of facilitator was performed by the teacher), to solve the second kata and the following ones the Randori and Kake formats were used, allowing participants to act more actively; and (b) collecting the perceptions of participants about the dynamics of different dojo formats could enrich the results and discussion of the study.
- Aiming to maximize student learning, the katas were created by the authors of this study so that the activities could be good representatives for each testing criteria, could have descriptions related to the same context, and be organized in a progressive order of complexity. Furthermore, the katas covered several stages of a testing process (e.g., test design, automation, and execution, as will be detailed in Section 6) so that students could have a more complete learning experience.

5.3.2 Experiment executions and data collection

Experiment executions took place at university labs, during class periods. Each execution was conducted in two days: the first on June 12th and 19th, 2023, and the second on October 16th and 23rd, 2023. For comparability, the same test was administered to both classes/executions, as detailed below.

Initially, students were invited to participate in the research, whose details were explained to them by the first author through a guided reading of the ICF. The students who agreed to collaborate with the study filled out the ICF and, after the consent collection, execution began.

On the first day, the activities shown in the upper lane of **Figure 5** were carried out. The activities and the time[9] allocated to them in each execution are next described. First, students filled out the participant characterization questionnaire (5 min) and the knowledge pre-test (40 min in the first execution — hereinafter referenced as 1E — and 30 min in the second execution — hereinafter labeled 2E). Then, they were submitted to the first part of the treatments (the same for both), which had a theoretical focus: a video (25 min) about testing concepts and functional testing technique criteria followed by a moment of discussion, where students could question the researchers about the content (20 min in 1E and 30 min in 2E).

On the second day, the activities shown in the lower lane of **Figure 5** were conducted. Each participant was allocated to the control or experimental group (15 min in 1E and 10 min in 2E). After moving to separate laboratories, each group was submitted to the second part of their appropriate treatments, which had a practical focus:

- Experimental group: first, the dojo introduction (7 min) and the initial agreements (3 min) were carried out. Next, four katas were solved during the dojo execution (50 min in 1E and 90 min in 2E): the first using the Prepared Kata format, the second and third using Kake, and the last using Randori. Finally, students filled out the retrospective questionnaire (10 min).
- Control group: each student initially solved a list of exercises (the same organized as katas for the experimental group) individually (40 min in 1E and 60 min in 2E). Then, an answer sheet was made available to the students, who could compare their answers to those on this document and question the researcher who accompanied the group about the exercises (20 min in 1E and 40 min in 2E). Finally, students filled out the retrospective questionnaire (10 min).

[9] Time is expressed in minutes. For some activities, it differs between the first (1E) and second execution (2E) because the 2E times were adjusted based on the perceptions of the researchers and participants regarding 1E.

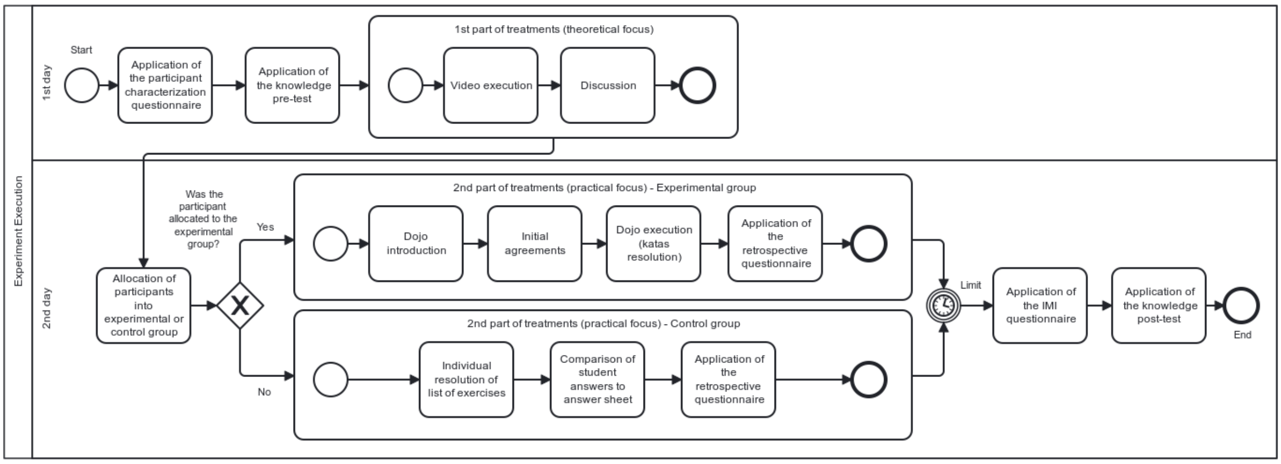

Experiment Execution
1st day
Start
Application of the participant characterization questionnaire
Application of the knowledge pre-test
1st part of treatments (theoretical focus)
Video execution
Discussion
2nd day
Allocation of participants into experimental or control group
Was the participant allocated to the experimental group?
Yes
No
2nd part of treatments (practical focus) - Experimental group
Dojo introduction
Initial agreements
Dojo execution (katas resolution)
Application of the retrospective questionnaire
2nd part of treatments (practical focus) - Control group
Individual resolution of list of exercises
Comparison of student answers to answer sheet
Application of the retrospective questionnaire
Limit
Application of the IMI questionnaire
Application of the knowledge post-test
End

Figure 5. Experiment execution

After that, participants answered the IMI questionnaire (10 min) and the knowledge post-test (40 min in 1E and 30 min in 2E), which was the last activity of the execution process.

5.3.3 Data analysis procedure

In this work, quali-quantitative analyses were conducted.

Quantitative analysis. Carried out on the data collected to answer RQ_2 and RQ_3, i.e., data from the knowledge tests and the IMI questionnaire, respectively. The quantitative analysis included the following steps:

1. Calculation of the learning and motivation results (using the formulas described next to the hypotheses in Section 5.3.1).
2. Data characterization using descriptive statistics: to understand general aspects of the distribution of results, descriptive statistics (e.g., minimum, maximum, median, first and third quartiles, mean, variance, and sample standard deviation values) were calculated and analyzed along with box plots.
3. Data set reduction: outliers were identified, analyzed and, when appropriate, removed from the data set.
4. Hypothesis testing: first, Shapiro-Wilk test[10] (Shapiro and Wilk, 1965) was used to test the data for normality. In the case of a normal/Gaussian distribution, Student's t-test was applied next; otherwise, the Mann-Whitney test was used. After applying the statistical tests, it was possible for each pair of hypotheses: (a) if p-value $\leq \alpha$ (0.05), reject the null hypothesis, concluding the existence of a statistically significant difference between the treatments; or (b) if the null hypothesis could not be rejected (p-value $> \alpha$), conclude that there is no statistically significant difference between the treatments.
5. Finally, when p-value $\leq \alpha$, the effect size (i.e. the practical significance of the result) was calculated. This study used Hedges' g_s (Lakens, 2013), whose values were interpreted according to the ranges by Cohen (1988).

Qualitative analysis. According to Seaman (2008), qualitative methods enable richer and more informative analyses of variables that are difficult to quantify, such as human perceptions. Additionally, qualitative data are useful in supplementing the quantitative data obtained from an experiment (Robson and McCartan, 2016). Therefore, in this work qualitative analysis was used to better understand the perceptions of participants regarding ADoTe. To this purpose, the answers of the experimental group to the retrospective questionnaire were evaluated through a thematic analysis (Robson and McCartan, 2016) composed of the following steps:

1. Data familiarization and validation: the answers were read by the researchers and inconsistent data was removed (e.g., answers resulting from incorrect interpretation of questions).
2. Code generation: relevant excerpts from the answers (i.e., those related to what was being investigated) were associated with codes (identifiers) that represented topics of interest for analysis. This action was carried out iteratively and systematically across the entire data set, with the same codes being assigned to excerpts with equivalent meaning.
3. Identification of themes: the codes were grouped into larger categories called subthemes and themes. Such categories were defined through an iterative process that encompassed the analysis of codes and the search for relationships between them that were relevant to the topics investigated (i.e., positive and negative aspects of ADoTe, as well as improvement suggestions for it).
4. Integration and interpretation: considering the results of the previous steps, data was interpreted by comparing the excerpts with each other, identifying patterns, and verifying their relationship with literature and the elements of ADoTe. This step was also supported by the integration of data into tables and diagrams.

6 ADoTe

The approach was named "Approach to teaching and learning functional testing technique criteria supported by Testing Dojo" (in Portuguese, "Abordagem para o ensino e apren-

[10] BioEstat software (https://www.mamiraua.org.br/downloads/programas; version 5.3) was used for statistical tests. A confidence interval of 95% was adopted (significance level $\alpha \leq 0.05$).

dizagem de critérios de teste da técnica funcional apoiada por Dojo de Teste"), whose abbreviation is ADoTe.

First, it is important to mention that ADoTe is not completely prescriptive. The authors understand that, when organizing a teaching and learning process, teachers must consider at least: (a) their own experience; (b) the learning objectives to be achieved (i.e., what students must know and be able to accomplish at the end of the process); and (c) the teaching and learning context (e.g., the profile and number of students in their classes and the time available to teaching). Thus, instead of exhaustively describing what should be done, it was considered more appropriate to provide through ADoTe a base structure that guides how the teaching and learning process should occur, but that allows teachers to instantiate the approach according to their contexts, aiming to maximize its results.

Figure 6 illustrates the base structure of ADoTe, whose activities are organized into four stages — plan, do, check, and act —, a definition based on the PDCA cycle, an "iterative method for conducting improvement activities" (Carpinetti and Gerolamo, 2016, p. 13). The decision of organizing ADoTe activities in these stages occured due to the following reasons: (1) as identified in the results of the SLM detailed in Section 2, for the dojo to be effective, it is necessary to carry out prior planning (de Oliveira et al., 2018); and (2) the retrospective moment of dojos aims to allow reflections on the activity, but also to enable improvements on it (Gaertner, 2010). However, since it does not prescribe a minimum structure, it may end up not fulfilling the latter function (e.g., if improvement actions are suggested, but not planned and implemented). Furthermore, since the retrospective occurs for just a few minutes at the end of the dojo session, the teacher's assessment of teaching and learning results may be limited if carried out only at that moment. Thus, in addition to the execution of the teaching and learning activities, we also perceived the need for formal moments of planning and control/action for continuous improvement. The PDCA cycle is a method already consolidated in other areas of knowledge and includes stages related to such needs, and, therefore, we concluded that it would be beneficial to organize ADoTe activities in the four stages of a PDCA cycle. Consequently, when carried out cyclically, these stages enable teachers to plan the teaching and learning process, execute it, evaluate it, and, after that, consolidate their knowledge about the process and improve it, if necessary, for new executions.

Studies mention a supposed inefficiency and insufficiency of traditional approaches (usually theoretical-expository) in software testing teaching (Jesus et al., 2019). According to them, when just such approaches are used it is difficult for teachers to motivate the students and transmit in-depth and practical knowledge (Costa and Oliveira, 2019; Gomes and Lelli, 2021). Thus, given the need to support a more active, practical (Elgrably and Oliveira, 2022b; Scatalon et al., 2018), and motivating (Garousi et al., 2020b) teaching, the use of testing dojos was suggested in this work. However, according to Luz et al. (2013), dojos would not replace theoretical classes, but their combined use would be positive. Additionally, for Sato et al. (2008), dojos can complement traditional teaching methods. Considering the above, we realized that ADoTe should cover theoretical and practical aspects.

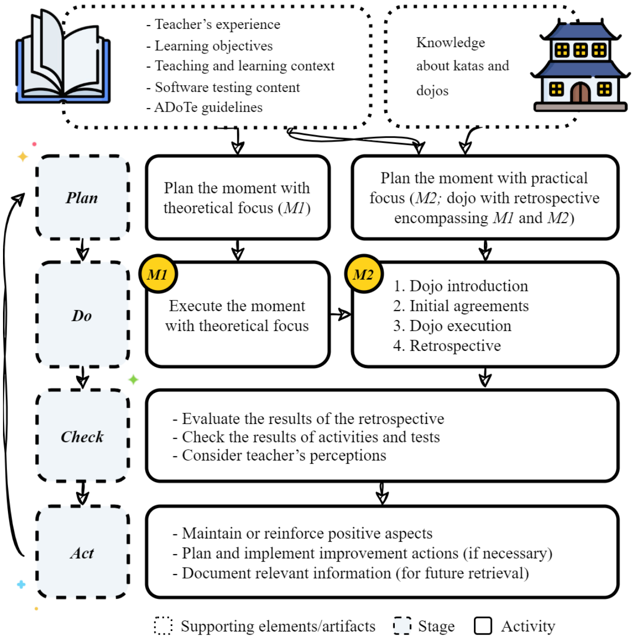

- Teacher's experience
- Learning objectives
- Teaching and learning context
- Software testing content
- ADoTe guidelines
Knowledge about katas and dojos
Plan
Plan the moment with theoretical focus (*M1*)
Plan the moment with practical focus (*M2*; dojo with retrospective encompassing *M1* and *M2*)
Do
M1
Execute the moment with theoretical focus
M2
1. Dojo introduction
2. Initial agreements
3. Dojo execution
4. Retrospective
Check
- Evaluate the results of the retrospective
- Check the results of activities and tests
- Consider teacher's perceptions
Act
- Maintain or reinforce positive aspects
- Plan and implement improvement actions (if necessary)
- Document relevant information (for future retrieval)
Supporting elements/artifacts
Stage
Activity

Figure 6. Base structure of ADoTe

Therefore, its planning and execution stages were organized around two moments (highlighted in **Figure 6**) of teaching and learning:

- $M1$: this moment has a theoretical focus. In $M1$, concepts of software testing and functional testing technique criteria must be presented to the students. The goal is to introduce them to the subject, as well as provide them with the theoretical basis necessary for $M2$. The content can be presented through an expository class supported by resources such as slides or videos. Students should also be allowed to question teachers.
- $M2$: this moment has a practical focus. $M2$ must occur after $M1$ (not necessarily on the same day) through the execution of one or more testing dojo sessions focused on practical exercises about functional testing technique criteria.

In the dojo sessions from $M2$ participants play roles and are involved in dynamics that are adherent to active, collaborative, and cooperative learning strategies (the alignment of ADoTe with these strategies is detailed in **Table 2**), since the katas are solved through the involvement of all participants, in a practical, collaborative, and non-competitive way. Furthermore, the guidelines of ADoTe (presented later in this section) include recommendations for ensuring the practical, collaborative, and cooperative nature of dojos in the approach, reinforcing the importance of these aspects. Consequently, it is expected that ADoTe can lead to positive results in student learning and motivation, such as those reported by Gehringer (2007) and McConnell (2005) for the other strategies aforementioned.

At the top of **Figure 6** supporting elements are presented, which are input artifacts to the stages of ADoTe, must be considered in their activities, and can be related to one or more moments of the approach (the relationships are indicated by arrows in the figure). The basic elements to be considered are: (a) the teacher's experience, the learning objectives, and the teaching and learning context, as mentioned at the beginning of this section; (b) the content to be taught (i.e., software

Table 2. Alignment of ADoTe with active, collaborative, and cooperative learning strategies

Learning Strategy	**Theoretical Principle (Prince, 2004)**	**Implementation in ADoTe**
Active learning	Students actively engaging in classroom (e.g., solving problems, discussing, analyzing, creating...) instead of just listening	- In the moment $M2$ of ADoTe, students solve problems/katas actively. When playing roles such as pilot, copilot, and audience (see Subsection 4.2) they do not just act as listeners, but ask questions, discuss the actions necessary to solve the katas, take notes about the activity, and participate directly in the construction of the solutions (e.g., using the keyboard and mouse). - Guidelines such as $G1$, $G3$, and $G5$ include recommendations that help to ensure active student participation, since they reinforce the structure, roles and behaviors expected for $M2$.
Collaborative learning	Students collaborating/ acting as a group instead of acting alone	- In the moment $M2$ of ADoTe, students solve katas collaboratively, as a group. The attributions of each role and the dynamics of the dojo formats (see Subsection 4.2) promote knowledge sharing and the resolution of katas in a collective manner. - Guidelines such as $G1$ and $G3$ include recommendations that help to ensure collaboration.
Cooperative learning	Students cooperating instead of competing	- In the moment $M2$ of ADoTe, students solve katas cooperatively (not competitively). The attributions of each role and the dynamics of the dojo formats encourage relationships of mutual assistance between the students (e.g., the audience, when necessary, and the copilot help the pilot in the construction of the kata solution). - Guidelines such as $G1$ and $G3$ include recommendations that help to ensure cooperation.

testing, focusing on functional testing technique criteria); (c) a set of guidelines (presented below) on how $M1$ and $M2$ should be planned and executed; and (d) knowledge about katas and dojos (see Subsection 4.2)[11].

The guidelines (G) to be considered in the planning and execution of $M2$ (some also applicable to $M1$) were initially defined based on the analysis of the results of the SLM described in Section 2 and, subsequently, refined and validated based on the results of the experiment conducted to evaluate ADoTe (see Section 7). The guidelines are:

G1: In order not to mischaracterize the dojo (especially its practical, collaborative, and cooperative nature), the teacher should follow the basic dojo definitions and structure (presented in subsection 4.2). Furthermore, it is important to reinforce: (a) the cooperative nature of the activity, if competitiveness is noticed between the participants; and (b) the fact that students are expected and allowed to make mistakes throughout the learning process (to reduce discomforts that can be caused by interactions or exposure in the activity).

G2: The role of facilitator must be played by a teacher or professional with knowledge about functional testing technique criteria. This person must mediate the dojo and provide the necessary explanations, models, and examples so that students can work on the katas.

G3: If the facilitator notices that a participant is not able to act actively in the dojo (e.g., due to excessive collaboration from other students with more knowledge or experience) he/she must intervene to ensure that everyone can share their opinions with the group.

G4: When defining the dojo formats to be used, the amount of facilitation, and the katas to be worked on in a session, the teacher must consider the guidelines of ADoTe and the teaching and learning context (e.g., the time available for teaching and the profile and number of students).

G5: When using a dojo format with multiple roles (e.g., pilot, copilot, and audience) it must be ensured that each participant can play each role at least once.

G6: Attention must be paid to the size of the steps used to solve the katas: they must be small enough for participants to be able to follow what is being done, but not so small as to cause the feeling of lack of progress.

G7: The time allocated to a cycle/round of work on a kata must be proportional to the difficulty and size of the activities to be carried out on it (also keeping in mind the concept of baby steps). As the definition of this time is not trivial, it is suggested that, when conducting multiple dojo sessions, teachers must pay attention to the perceptions of participants about the time of each round, so they can adjust it accordingly from a historical basis.

G8: Katas must be solved in a non-decreasing order of difficulty. Furthermore, if there is time for multiple sessions, it is recommended to solve katas with different and increasing levels of difficulty in each one.

G9: The facilitator should provide feedback to participants regarding their solutions for a kata (after each round, for example), so that they can correct misunderstandings and, when using dojo formats with multiple pairs (as Kake), maintain joint progress in the activity.

G10: An adequate infrastructure must be ensured (to avoid problems related to this aspect, such as difficulties in understanding due to low quality of audio or video). When projecting the computer screen in the room, it must be ensured that participants can view the content without difficulty. Therefore, attention should be paid to the font size of materials. Students should also be instructed to inform the teacher/facilitator if they are unable to see something (in this scenario, appropriate actions must be taken to solve the problem). This guideline is also applicable to $M1$.

Finally, the activities of ADoTe must occur in this order:

- Plan: the planning stage includes the definition of what will be done in $M1$ and $M2$ and how their actions will be conducted. At the end of $M2$, a retrospective must be performed (it can be based on the examples of subsubsection 4.2.3, but must be adapted to evaluate $M1$ in addition to the dojo/$M2$).

[11] If necessary, teachers can complement their knowledge about katas and dojos with supplementary material (e.g., books and articles) or by participating in dojo sessions promoted by other professionals. We also believe the use of ADoTe over time will expand the knowledge of the teacher about these topics through practical experience.

- Do: includes the execution of $M1$ and $M2$ according to the definitions of the previous stage. In $M2$, at least the typical moments of a dojo session must be carried out: introduction, initial agreements, execution, and retrospective.
- Check: relates to the evaluation of the teaching and learning process and its results (to determine whether the learning objectives were achieved and identify aspects to be maintained or improved). Teachers must check the results of the retrospective and instruments that have been used (e.g., activities and tests). They can also consider their perception of what was accomplished and the behavior and performance of students.
- Act: considering the outcomes of the previous stage, activities that support new teaching and learning moments must be executed, improving them when possible. Therefore, it must be: (a) maintained or reinforced positive aspects; (b) planned and implemented improvement actions (if necessary and possible); and (c) documented information that is relevant to future uses of ADoTe.

Considering the above, and complementing the differences already described in Section 3, we can state that ADoTe differs from other dojo approaches used in various contexts since it presents a base structure that has not yet been reported in other studies. Furthermore, by including in its structure the formal use of a PDCA cycle, a set of guidelines, and moments such as $M1$ — defined based on results from the literature and from the experiment conducted in this research, as already described in this section —, ADoTe presents elements that can help avoid or at least reduce the problems/challenges regarding dojos reported in other studies (described in Section 2.3).

A detailed instance of ADoTe can be found in the experimental package[5], where information covering the planning and execution stages is exposed. Following the structure and guidelines of ADoTe, in this instance $M1$ includes the reproduction of a video about testing concepts and functional testing technique criteria followed by a moment of discussion, where students could question the researchers about the content. In turn, $M2$ includes a dojo session with katas that are focused on functional testing technique criteria and cover several stages of a testing process: test design, automation of test cases (using JUnit 5[12]), and test execution (manually and in an automated way, including the comparison of the results obtained with those expected and the recording and analysis of failures). As this instance was used to evaluate ADoTe, in Section 7 are presented examples of the evaluation and action stages of the approach.

7 Results

This section describes the results of the executions of the experiment. Subsection 7.1 presents the characterization of participants. Subsections 7.2, 7.3, and 7.4 expose the learning, motivation, and retrospective results, respectively.

7.1 Characterization of participants

Forty four students participated in the experiment, twenty-two in each execution.

In the first execution nineteen (86.4%) students declared themselves to be male, two (9.1%) to be female, and one (4.5%) preferred not to indicate gender. In the second execution nineteen (86.4%) declared themselves to be male and three (13.6%) to be female.

The ages of the participants from the first execution ranged from 20 to 29 years old, with median and mode equal to 22 years old. In the second, ages ranged from 20 to 32 years old, the set of ages was bimodal (modes equal to 21 and 24), and the median was 23 years old.

Regarding academic background, in the first execution twenty participants (90.9%) were from an Information Systems (IS) course and two (9.1%) from Computer Engineering (CE). In the second, seventeen (77.3%) were from an IS course and five (22.7%) from CE. As shown in **Figure 7**, in both executions most students were in the final periods of their courses, since IS was an eight-period course and CE a ten-period course.

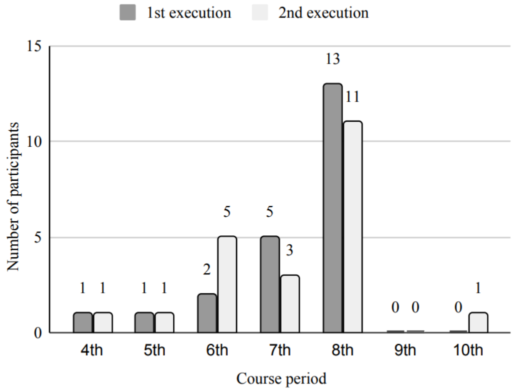


Figure 7. Course periods of participants

The professional experience of participants with software testing (results in **Figure 8**) ranged, in the first execution, from none (fourteen participants; 63.6%) to more than six years (eight participants already had some experience; 36.4%). In the second execution the professional experience was equal to none (sixteen participants; 72.7%), up to two years (five participants; 22.7%), or ranging from two to four years (one participant; 4.6%).

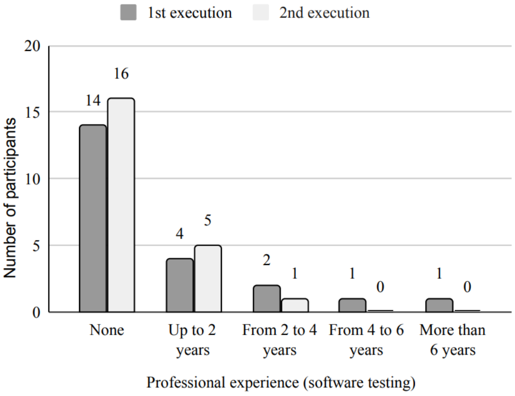


Figure 8. Professional experience of participants (software testing)

[12] https://junit.org/junit5/

Finally, regarding the level of prior knowledge of participants about the functional testing technique criteria known as equivalence partitioning, boundary value analysis, and systematic functional testing (results in **Figure 9**): (1) in the first execution, the majority of respondents (sixteen; 72.7%) had only theoretical knowledge on the topic, three (13.6%) had already applied at least one of these criteria in the industry, two (9.1%) reported not knowing the topic and one (4.5%) had already applied (in practice) one or more of the criteria at university, but never professionally; and (2) in the second execution, the majority of respondents (fifteen; 68.2%) reported not knowing the topic, six (27.3%) had only theoretical knowledge, and one (4.5%) had already applied at least one of these criteria in the industry.

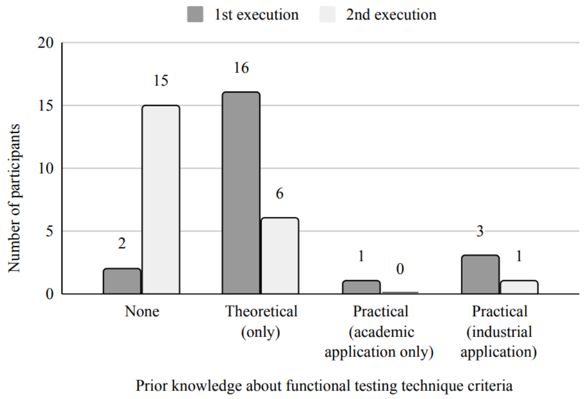


Figure 9. Prior knowledge of participants about functional testing technique criteria

The demographic results exposed above are discussed in Subsection 8.1.

7.2 Learning results

The learning results evidenced for the participants of the experimental (G_E) and control (G_C) groups (both of size $n = 11$ in each execution) were: (1) in the first execution, G_E = {0, 0, 1, 1, 1, 1, 2, 3, 3, 4, 6}, G_C = {-1, 0, 0, 0, 1, 1, 2, 2, 2, 5, 5}; and (2) in the second execution, G_E = {2, 2, 3, 3, 4, 4, 5, 6, 7, 7, 8}, G_C = {-1, 0, 3, 3, 3, 4, 4, 6, 6, 6, 7}. The distributions of these results are exposed in **Figures 10** and **11** (data from the first and second executions, respectively).

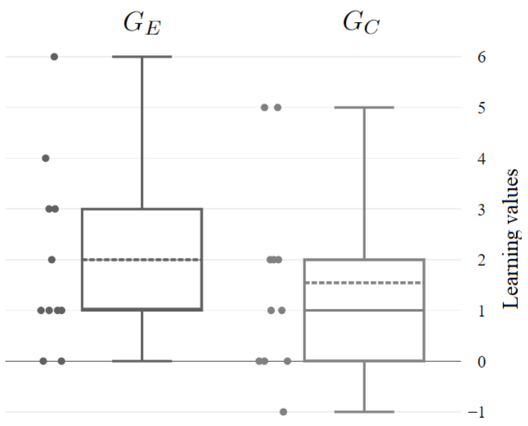


Figure 10. Distributions of learning results — 1st execution

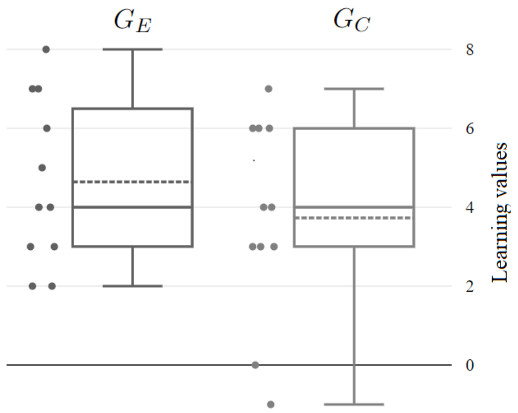


Figure 11. Distributions of learning results — 2nd execution

Analyzing the box plots of **Figure 10** it is possible to notice that, in the first execution, in both groups (each one of size $n = 11$) the majority of the results were positive. Results equal to zero were also verified in G_E and G_C, while negative ones occurred just in G_C. The minimum learning values were -1 for G_C and 0 for G_E, which indicates that the worst learning result was from a participant of G_C. The maximum value was higher in the experimental group (6, compared to 5 in G_C), which indicates that the best learning result was from a participant of G_E. The quartiles of G_E were equal to or higher than those of G_C and the average learning of G_E (2.0000) was higher than that of the control group (1.5455). Furthermore, the result set of G_E was the most uniform, as it presented smaller values of variance (s^2) and sample standard deviation (s) (G_E: $s^2 = 3.4000$ and $s = 1.8439$; G_C: $s^2 = 3.8727$ and $s = 1.9679$).

In turn, through the box plots of **Figure 11** it is possible to notice that in the second execution the results of G_E were all positive, while in G_C there were positive, null, and negative results. The minimum learning values were -1 for G_C and 2 for G_E, which indicates that the worst learning result was from a participant of G_C. The maximum value was higher in the experimental group (8, compared to 7 in G_C), which indicates that the best learning result was from a participant of G_E. The quartiles of G_E were equal to or higher than those of G_C and the average learning of G_E (4.6364) was higher than that of the control group (3.7273). Furthermore, the result set of G_E was the most uniform, as it presented smaller values of variance (s^2) and sample standard deviation (s) (G_E: $s^2 = 4.4545$ and $s = 2.1106$; G_C: $s^2 = 6.4182$ and $s = 2.5334$). These results indicate that in both executions participants submitted to ADoTe obtained, on average, greater learning than those submitted to the traditional approach.

To test the hypotheses related to learning ($2H_0$ and $2H_1$), data of each execution was first tested for normality with the Shapiro-Wilk test. The results were (1) in the first execution, $W = 0.8877$ and p-value $= 0.1733$ (for G_E) and $W = 0.8763$ and p-value $= 0.1016$ (for G_C); (2) in the second execution, $W = 0.9254$ and p-value $= 0.4085$ (for G_E) and $W = 0.9109$ and p-value $= 0.3179$ (for G_C). Since the results for p-value were greater than α (0.05), it was verified a normal/Gaussian distribution of data in both groups of each execution. Then, the Student's t-test for two independent samples was applied (results in **Table 3**).

Table 3. Student's t-test results (learning)

Statistics	1st execution	2nd execution
Homoscedasticity	Yes	Yes
Aggregate variance (s_a^2)	3.6364	5.4364
t	0.5590	0.9144
Degrees of freedom (ν)	20	20
p-value (two-sided)	0.5823	0.3714
Power (0.05)	0.1373	0.2323
Confidence Interval (95%)	-1.2416 to 2.1507	-1.1648 to 2.9830

Considering data from **Table 3**, since the results for p-value (two-sided) were greater than α (0.05), in both executions it was not possible to reject the null hypothesis $(2H_0)$.

Thus, in the experiment it was not evidenced a statistically significant difference between the treatments for the learning variable.

7.3 Motivation results

The motivation results evidenced for the participants of the experimental (G_E) and control (G_C) groups in each execution are presented in **Tables 4** and **5**. The descriptive statistics and the distributions of these results are respectively exposed in **Table 6** and **Figure 12** (data from the first execution of the experiment) and in **Table 7** and **Figure 13** (data from the second execution of the experiment). In addition, to allow the reader a more detailed understanding of the results, the means of the answers obtained for each IMI item are presented in Appendix B.

Regarding the values obtained for the IMI subscales, it is possible to notice from **Table 6** and **Figure 12** that in the first execution of the experiment:

- For INT, CMP, and VAL the means of G_E were higher than those of G_C. Furthermore, for these subscales the minimum, maximum, Q1, Q2, and Q3 values of G_E were mostly higher than those of G_C, being equal in some cases. In turn, for PRS the mean of G_E was lower than that of G_C, the minimum and Q1 values of the first group were equal to those of the second, and the maximum, Q2, and Q3 values were lower. These statistics indicate that, on average, the results for INT, CMP, VAL, and PRS were better in the group submitted to ADoTe.
- For EFF, the mean of G_C was higher than that of G_E. The minimum, Q1, Q2, and Q3 values of G_C were also the highest, with the maximum value being the same for both groups. Such statistics indicate that, on average, EFF results were better in the group submitted to the traditional teaching approach.
- For INT and PRS, the sets of results of G_E were the most uniform, as they presented smaller values of s^2 and s. For the other subscales, the opposite occurred.
- Differences between group means ranged from 1.30% (VAL) to 15.80% (INT).
- Outliers were verified only for PRS (one in G_E and two in G_C). However, as some participants may have felt very tense/pressured (differing from the results shown for the others), it was decided not to consider such data as false, keeping them in the analyses.

In turn, considering the results from **Table 7** and **Figure 13**, in the second execution it was verified for the IMI subscales that:

- For INT, CMP, and VAL the means, minimum, maximum, Q1, Q2, and Q3 values of G_E were mostly higher than those of G_C (the exceptions were the maximum value of VAL and Q1 of CMP, which were equal to and lower than those of G_C). These statistics indicate that, on average, the results for INT, CMP, and VAL were better in the group submitted to ADoTe.
- For EFF, the mean, minimum, and maximum values of G_C were higher than those of G_E, while Q1, Q2, and Q3 values were equal to or lower than those of the experimental group. In turn, for PRS the mean of G_C was lower than that of G_E, with the minimum value of the first group being the same as that of the second, and the maximum, Q1, Q2, and Q3 values of G_C being lower. Such statistics indicate that, on average, the results for EFF and PRS were better in the group submitted to the traditional teaching approach.
- Except for VAL, the sets of results of G_C were the most uniform, as they presented smaller values of s^2 and s.
- Differences between group means ranged from 1.82% (ESF) to 19.71% (INT).
- One result (for PRS, in G_C) was considered an outlier. However, as the participant may have felt very tense/pressured, it was decided not to consider such data as false, keeping it in the analyses.

Regarding general motivation, it was verified from **Tables 6**-**7** and **Figures 12**-**13** that:

- The most uniform result sets were G_E in the first and G_C in the second execution, as they presented smaller values of s^2 and s.
- Except for one subject from G_C, in the first execution motivation levels[13] were positive, ranging from 63.33% (since the values were higher than 16) to 98.57% (the maximum value was equal to 26.57). In the second execution motivation levels ranged (a) in G_E, from 37.30% (minimum value $= 8.19$) to 95.33% (maximum $= 25.60$), and (b) in G_C, from 37.90% (minimum $= 8.37$) to 82.67% (maximum $= 21.80$).
- In the first execution the lowest motivation value occurred in G_C and the highest in G_E. Q1, Q2, Q3, and mean values were higher in G_E (the difference between the group means was equal to 1.4964, i.e., 4.99%). In the second execution the lowest and highest motivation values occurred in G_E. Q1 and Q2 were higher in G_C, while the highest Q3 and mean values were verified in G_E (the difference between the group means was equal to 0.8582, i.e., 2.86%). These results indicate that, on average, in both executions participants submitted to ADoTe felt more motivated.

[13] The possible range of motivation values (-3 to 27) was converted into a percentage scale to calculate these results.

Table 4. Motivation results - 1st execution

Variable	**Experimental group (G_E)**	**Control group (G_C)**
Motivat.	18.10, 19.96, 25.21, 16.93, 23.33, 17.86, 20.93, 18.04, 22.24, 26.57, 20.35	16.54, 22.20, 16.96, 21.41, 25.26, 22.47, 19.61, 12.31, 18.57, 17.66, 20.07
INT	6.43, 5.43, 6.71, 4.86, 6.86, 5.00, 5.00, 5.71, 6.14, 6.57, 5.71	4.71, 4.57, 4.86, 4.57, 6.57, 5.57, 5.57, 3.14, 5.57, 3.43, 5.43
CMP	4.67, 5.50, 6.50, 5.50, 4.67, 5.00, 6.50, 4.67, 5.67, 7.00, 4.50	4.17, 6.17, 4.67, 5.33, 5.83, 5.67, 4.67, 3.83, 4.83, 5.17, 4.67
PRS	4.80, 2.40, 1.00, 1.00, 1.40, 2.20, 1.00, 2.60, 1.60, 1.00, 1.20	6.00, 2.40, 2.60, 1.00, 1.00, 2.00, 2.20, 5.40, 1.60, 1.00, 1.00
EFF	4.80, 6.00, 6.00, 2.00, 6.20, 4.20, 4.00, 4.40, 5.60, 7.00, 5.20	6.80, 7.00, 4.60, 5.80, 7.00, 6.80, 5.00, 4.60, 4.20, 4.20, 5.40
VAL	7.00, 5.43, 7.00, 5.57, 7.00, 5.86, 6.43, 5.86, 6.43, 7.00, 6.14	6.86, 6.86, 5.43, 6.71, 6.86, 6.43, 6.57, 6.14, 5.57, 5.86, 5.57

Table 5. Motivation results - 2nd execution

Variable	**Experimental group (G_E)**	**Control group (G_C)**
Motivat.	12.00, 25.60, 21.27, 16.61, 13.18, 11.61, 10.74, 23.57, 22.57, 23.89, 8.19	21.80, 8.37, 13.31, 10.06, 18.31, 18.31, 18.87, 21.49, 12.56, 21.45, 15.26
INT	4.43, 7.00, 5.29, 4.71, 5.86, 4.14, 5.14, 6.57, 6.14, 6.86, 2.57	5.43, 3.14, 4.00, 2.43, 5.71, 4.71, 3.57, 4.57, 3.00, 5.71, 3.43
CMP	3.00, 6.00, 5.67, 5.33, 3.83, 2.50, 3.00, 6.00, 4.83, 6.17, 3.33	5.17, 3.00, 3.17, 4.00, 2.00, 5.17, 4.50, 5.83, 4.33, 5.17, 3.83
PRS	5.00, 1.00, 2.80, 2.40, 5.00, 4.80, 5.80, 2.40, 1.00, 1.00, 4.00	2.40, 2.80, 5.20, 2.60, 1.00, 1.40, 1.00, 1.00, 1.00, 2.20, 1.60
EFF	5.00, 6.60, 6.40, 3.40, 3.20, 3.20, 3.40, 6.40, 5.60, 5.00, 1.00	6.60, 1.60, 5.20, 2.80, 4.60, 3.40, 6.80, 5.80, 2.80, 6.20, 4.60
VAL	4.57, 7.00, 6.71, 5.57, 5.29, 6.57, 5.00, 7.00, 7.00, 6.86, 5.29	7.00, 3.43, 6.14, 3.43, 7.00, 6.43, 5.00, 6.29, 3.43, 6.57, 5.00

Table 6. Motivation statistics — 1st execution

Statistics	**INT**		**CMP**		**PRS**		**EFF**		**VAL**		**Motivation**	
	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C
Minimum value	4.86	3.14	4.50	3.83	1.00	1.00	2.00	4.20	5.43	5.43	16.93	12.31
Maximum value	6.86	6.57	7.00	6.17	4.80	6.00	7.00	7.00	7.00	6.86	26.57	25.26
1st quartile (Q1)	5.2150	4.5700	4.6700	4.6700	1.0000	1.0000	4.3000	4.6000	5.8600	5.7150	18.0700	17.3100
Median (Q2)	5.7100	4.8600	5.5000	4.8300	1.4000	2.0000	5.2000	5.4000	6.4300	6.4300	20.3500	19.6100
3rd quartile (Q3)	6.5000	5.5700	6.0850	5.5000	2.3000	2.5000	6.0000	6.8000	7.0000	6.7850	22.7850	21.8050
Mean	5.8564	4.9082	5.4709	5.0009	1.8364	2.3818	5.0364	5.5818	6.3382	6.2600	20.8655	19.3691
Variance (s^2)	0.5339	0.9930	0.7586	0.5058	1.3265	3.0676	1.8865	1.3156	0.3680	0.3223	10.0621	12.4245
Standard deviation (s)	0.7307	0.9965	0.8710	0.7112	1.1518	1.7515	1.3735	1.1470	0.6067	0.5677	3.1721	3.5248
Diff. between means	0.9482 (15.80%)		0.4700 (7.83%)		0.5454 (9.09%)		0.5454 (9.09%)		0.0782 (1.30%)		1.4964 (4.99%)	

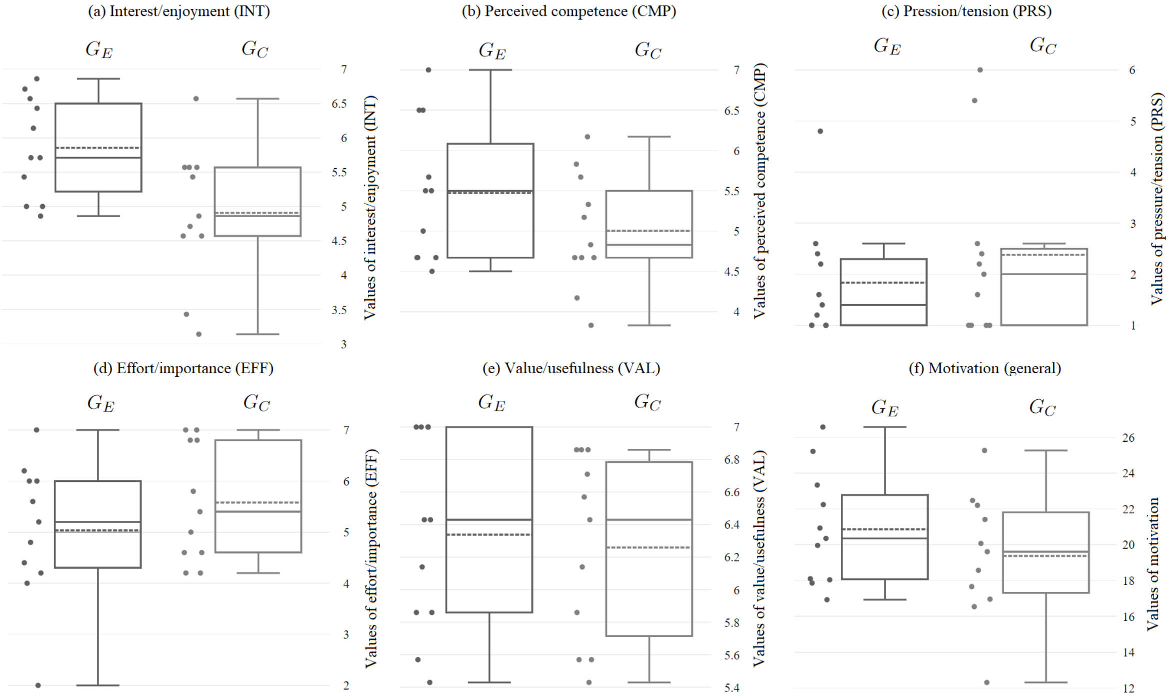
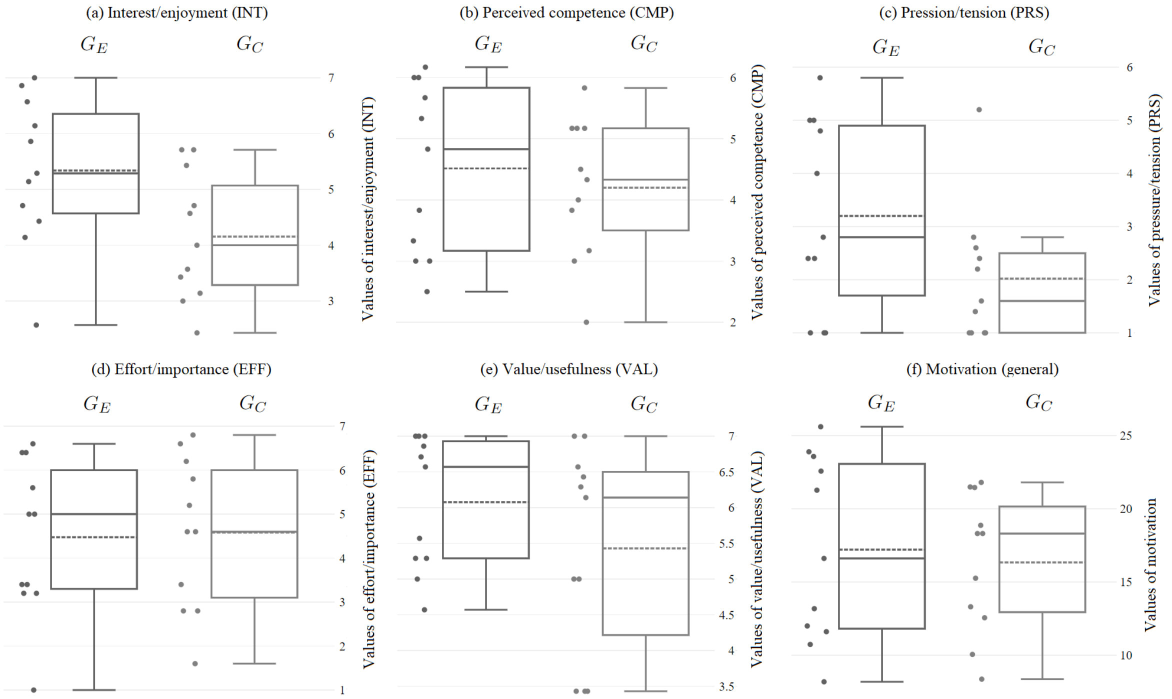


Figure 12. Distributions of motivation results (IMI subscales and general motivation) — 1st execution

Table 7. Motivation statistics — 2nd execution

Statistics	INT		CMP		PRS		EFF		VAL		Motivation	
	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C	G_E	G_C
Minimum value	2.57	2.43	2.50	2.00	1.00	1.00	1.00	1.60	4.57	3.43	8.19	8.37
Maximum value	7.00	5.71	6.17	5.83	5.80	5.20	6.60	6.80	7.00	7.00	25.60	21.80
1st quartile (Q1)	4.5700	3.2850	3.1650	3.5000	1.7000	1.0000	3.3000	3.1000	5.2900	4.2150	11.8050	12.9350
Median (Q2)	5.2900	4.0000	4.8300	4.3300	2.8000	1.6000	5.0000	4.6000	6.5700	6.1400	16.6100	18.3100
3rd quartile (Q3)	6.3550	5.0700	5.8350	5.1700	4.9000	2.5000	6.0000	6.0000	6.9300	6.5000	23.0700	20.1600
Mean	5.3373	4.1545	4.5145	4.1973	3.2000	2.0182	4.4727	4.5818	6.0782	5.4291	17.2027	16.3445
Variance (s^2)	1.7870	1.3174	1.9771	1.3104	3.2400	1.5956	3.1382	3.0116	0.8739	2.0843	39.9003	22.4204
Standard deviation (s)	1.3368	1.1478	1.4061	1.1447	1.8000	1.2632	1.7715	1.7354	0.9348	1.4437	6.3167	4.7350
Diff. between means	1.1828 (19.71%)		0.3172 (5.29%)		1.1818 (19.70%)		0.1091 (1.82%)		0.6491 (10.82%)		0.8582 (2.86%)	


(a) Interest/enjoyment (INT)

(b) Perceived competence (CMP)

(c) Pression/tension (PRS)

(d) Effort/importance (EFF)

(e) Value/usefulness (VAL)

(f) Motivation (general)

Figure 13. Distributions of motivation results (IMI subscales and general motivation) — 2nd execution

To test the hypotheses related to motivation ($3H_0$ and $3H_1$) and its subscales (see **Table 1**), data of each execution was first tested for normality with the Shapiro-Wilk test. From **Tables 8** and **9** it is possible to notice that, except for PRS (in both executions) and VAL (in the second execution), in all other samples data distribution was normal for the results of both groups, since p-values were greater than α (0.05).

Table 8. Shapiro-Wilk test results (motivation) — 1st execution

Variables	G_E		G_C	
	W	**p-value**	**W**	**p-value**
Motivation	0.9332	0.4578	0.9816	0.9656
INT	0.9187	0.3668	0.9365	0.4782
CMP	0.8898	0.1862	0.9709	0.8743
PRS	0.7597	0.0096	0.7707	0.0097
EFF	0.9443	0.5527	0.8600	0.0692
VAL	0.8746	0.0973	0.8668	0.0824

Table 9. Shapiro-Wilk test results (motivation) — 2nd execution

Variables	G_E		G_C	
	W	**p-value**	**W**	**p-value**
Motivation	0.8940	0.2122	0.9155	0.3469
INT	0.9503	0.6253	0.9328	0.4549
CMP	0.8781	0.1133	0.9578	0.7158
PRS	0.8909	0.1931	0.7971	0.0120
EFF	0.9108	0.3172	0.9416	0.5198
VAL	0.8429	0.0435	0.8410	0.0418

Thus, the Mann-Whitney test was applied to the samples where data distribution was not normal (results in **Tables 10** and **11**) and Student's t-test was applied to the other samples (results in **Tables 12** and **13**).

Considering the results from **Tables 10** to **13**:

- $3.1H_0$ was rejected (and consequently $3.1H_1$ was accepted), since in both executions p-value (two-sided) $<$ α for INT. Furthermore, the effect sizes (i.e., the practical significance of the results) were large, since Hedges'

Table 10. Mann-Whitney test results (PRS) — 1st execution

Statistics	G_E	G_C
Sample size	11	11
Rank sum (Ri)	118.0	135.0
Median	1.40	2.00
U	52.00	
Z(U)	0.5582	
p-value (two-sided)	0.5767	

Table 11. Mann-Whitney test results (PRS/VAL) — 2nd execution

Statistics	**PRS**		**VAL**	
	G_E	G_C	G_E	G_C
Sample size	11	11	11	11
Rank sum (Ri)	147.5	105.5	144.5	108.5
Median	2.80	1.60	6.57	6.14
U	39.50		42.50	
Z(U)	1.3790		1.1820	
p-value (two-sided)	0.1679		0.2372	

Table 12. Student's t-test results (motivation) — 1st execution

Statistics	**IMI subscale**				**Motiv.**
	INT	**CMP**	**EFF**	**VAL**	
Homosc.	Yes	Yes	Yes	Yes	Yes
s_a^2	0.7634	0.6322	1.6011	0.3452	11.2433
t	2.5450	1.3863	-1.0110	0.3121	1.0466
ν	20	20	20	20	20
p-value (two-sided)	0.0192	0.1808	0.3241	0.7582	0.3077
Power (0.05)	0.8159	0.3979	0.2629	0.0873	0.2747
CI (95%)	0.1710 to 1.7254	-0.2372 to 1.1772	-1.6709 to 0.5800	-0.4444 to 0.6008	-1.4861 to 4.4789

Table 13. Student's t-test results (motivation) — 2nd execution

Statistics	**IMI subscale**			**Motivation**
	INT	**CMP**	**EFF**	
Homosc.	Yes	Yes	Yes	Yes
s_a^2	1.5522	1.6438	3.0749	31.1603
t	2.2264	0.5804	-0.1459	0.3605
ν	20	20	20	20
p-value (two-sided)	0.0376	0.5681	0.8855	0.7222
Power (0.05)	0.7195	0.1422	0.0595	0.0962
CI (95%)	0.0746 to 2.2909	-0.8231 to 1.4577	-1.6688 to 1.4506	-4.1070 to 5.8234

$g_s = 1.04$ (in the first execution) and $g_s = 0.91$ (in the second execution).

- $3.2H_0$, $3.3H_0$, $3.4H_0$, $3.5H_0$, and $3H_0$ could not be rejected, since in both executions the results of p-value (two-sided) were greater than α for CMP, PRS, EFF, VAL, and motivation, respectively.

Therefore, in both executions of the experiment (a) it was evidenced a statistically significant difference between the treatments regarding interest/enjoyment, with results favoring ADoTe, and (b) it was not possible to state there was a statistically significant difference between the treatments concerning general motivation and the IMI subscales related to perceived competence, pressure/tension, effort/importance, and value/usefulness.

7.4 Retrospective results

Figure 14 illustrates the results of the thematic analysis conducted on the answers of the experimental group from the first execution of the experiment to the retrospective questionnaire.

In the diagram of **Figure 14**, participants are identified through the numbers presented before each answer. The theme named "positive aspects" groups the answers obtained for question 1, which aimed to identify what went well in the activity/favored the learning of participants. The answers to questions 2 and 3 are organized under "negative aspects" and "improvement suggestions", respectively. Through these questions, we aimed to identify what did not go well in the activity/hindered the learning of participants and collect improvement suggestions for ADoTe.

Nine codes were established based on the answers to question 1. All codes relate to $M2$ (the moment of ADoTe with a practical focus) and three can also be associated with $M1$ (the moment with a theoretical focus). Furthermore, the codes were grouped into three subthemes and, finally, into a general theme: positive aspects. Concerning the subthemes:

- "Favorable dojo elements" groups codes related to dojo elements (e.g., characteristics and dynamics) that were mentioned as favorable to learning. The "Dynamics" code refers to answers that express the approval of participants concerning a specific dojo format or the dynamics of this activity as a whole. The other codes indicate the approval of participants regarding practical activities, acting cooperatively and collaboratively (interactions), establishing a pressure-free environment, and pair rotation.
- "Support and facilitation resources" refers to materials and actions that support the teaching and learning process. From its codes it is evident that participants also considered as positive: (a) the provision of feedback about the correctness of their answers throughout the practical activity; (b) explanations and mediation of learning by the teacher/facilitator; and (c) the presentation of definitions and examples about the content (with the possibility of consulting them during the missions).
- "Integration of theory and practice" groups positive mentions regarding the integration of theory (exposed in $M1$) and practice ($M2$), validating the execution of both moments of ADoTe.

In turn, four codes were established for negative aspects (they relate to $M1$ and $M2$ and are grouped into three subthemes) and two for improvement suggestions (they refer only to $M2$ and could be grouped into a single subtheme). Concerning the subthemes:

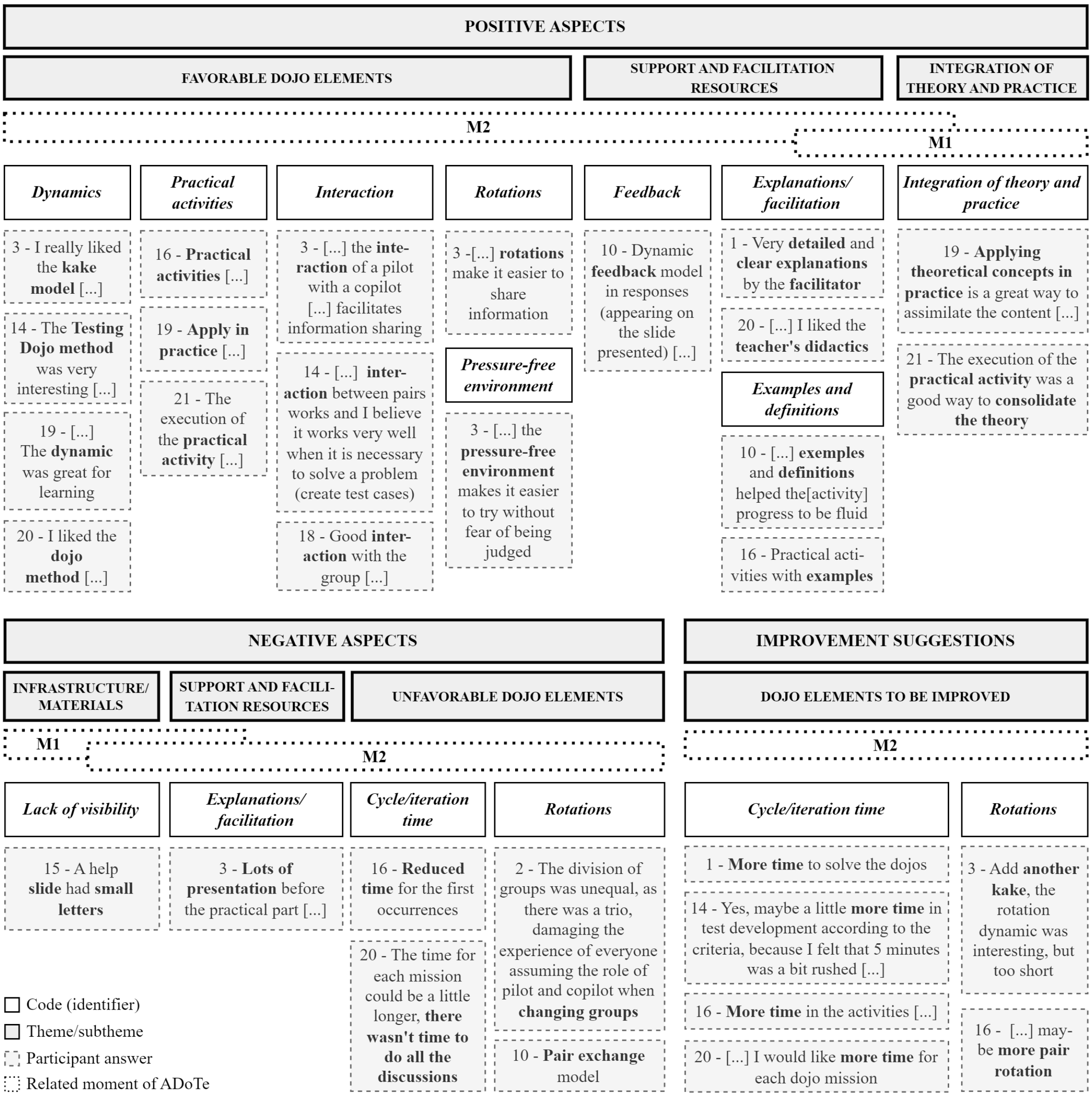

POSITIVE ASPECTS
FAVORABLE DOJO ELEMENTS
SUPPORT AND FACILITATION RESOURCES
INTEGRATION OF THEORY AND PRACTICE
M2
M1
Dynamics
Practical activities
Interaction
Rotations
Feedback
Explanations/ facilitation
Integration of theory and practice
3 - I really liked the kake model [...]
14 - The Testing Dojo method was very interesting [...]
19 - [...] The dynamic was great for learning
20 - I liked the dojo method [...]
16 - Practical activities [...]
19 - Apply in practice [...]
21 - The execution of the practical activity [...]
3 - [...] the interaction of a pilot with a copilot [...] facilitates information sharing
14 - [...] interaction between pairs works and I believe it works very well when it is necessary to solve a problem (create test cases)
18 - Good interaction with the group [...]
3 -[...] rotations make it easier to share information
Pressure-free environment
3 - [...] the pressure-free environment makes it easier to try without fear of being judged
10 - Dynamic feedback model in responses (appearing on the slide presented) [...]
1 - Very detailed and clear explanations by the facilitator
20 - [...] I liked the teacher's didactics
Examples and definitions
10 - [...] exemples and definitions helped the[activity] progress to be fluid
16 - Practical activities with examples
19 - Applying theoretical concepts in practice is a great way to assimilate the content [...]
21 - The execution of the practical activity was a good way to consolidate the theory
NEGATIVE ASPECTS
INFRASTRUCTURE/ MATERIALS
SUPPORT AND FACILITATION RESOURCES
UNFAVORABLE DOJO ELEMENTS
M1
M2
Lack of visibility
Explanations/ facilitation
Cycle/iteration time
Rotations
15 - A help slide had small letters
3 - Lots of presentation before the practical part [...]
16 - Reduced time for the first occurrences
20 - The time for each mission could be a little longer, there wasn't time to do all the discussions
2 - The division of groups was unequal, as there was a trio, damaging the experience of everyone assuming the role of pilot and copilot when changing groups
10 - Pair exchange model
IMPROVEMENT SUGGESTIONS
DOJO ELEMENTS TO BE IMPROVED
M2
Cycle/iteration time
Rotations
1 - More time to solve the dojos
14 - Yes, maybe a little more time in test development according to the criteria, because I felt that 5 minutes was a bit rushed [...]
16 - More time in the activities [...]
20 - [...] I would like more time for each dojo mission
3 - Add another kake, the rotation dynamic was interesting, but too short
16 - [...] maybe more pair rotation
□ Code (identifier)
□ Theme/subtheme
⬚ Participant answer
⬚ Related moment of ADoTe

Figure 14. Retrospective results — 1st execution

- "Infrastructure/materials" includes the code "Lack of visibility". In the answer classified with this code, it was mentioned that a slide (material) had small letters, which made it difficult to see his content when projected in the room. Aiming to prevent further occurrences of this situation, the slide was adjusted and the guideline about infrastructure ($G10$) was improved. Its new version (whose text is presented in Section 6) was considered in the second execution of the experiment.
- "Support and facilitation resources" includes the code "Explanations/facilitation". In the answer classified with this code, a participant mentioned that the amount of presentation (which consisted of explanations by the facilitator/teacher) before the practical part was excessive. On the other hand, as shown in the answers to question 1, other participants mentioned this same element as positive. Thus, we realized that varied perceptions can be obtained for this aspect due to differences in student profiles. To minimize problems related to the amount of facilitation, considerations on the topic were added to guideline $G4$.
- "Unfavorable dojo elements" groups codes related to dojo elements that, in the opinion of participants, did not go well or hindered their learning. "Dojo elements to be improved" corresponds to improvement suggestions for the dojo. The codes for these subthemes are: (a) "Rotations": in the first execution, only two rotations were performed when using Kake. As there was a trio, one participant did not play the role of pilot in this format, and this was reported as negative. In turn, another participant mentioned the pair exchange model as something that did not go well, which could be associated with the previous situation (an assumption, given that no further details were provided). Additionally, two participants suggested more rotations (which indicates the interest of students in this dynamic), one of them specifically referring to Kake. Aiming to avoid new occurrences of the problems reported and respond to the

improvement suggestions we (i) defined the $G5$ guideline and (ii) reorganized the third mission (in which the Kake format is used) so that it could be solved in the second execution through one more rotation; and (b) "Cycle/iteration time": some participants reported that the time to solve the missions was short, suggesting the allocation of more time to this. Therefore, we (i) revised the time and number of actions to be performed by pairs in each round (the time to solve the second mission was increased from 5 to 7 minutes; the third mission, solved in the first execution through a single 6-minute round, was reorganized to be solved in two rounds of 8 and 3 minutes, respectively) and (ii) added to $G7$ a suggestion of a process that can help to define mission times.

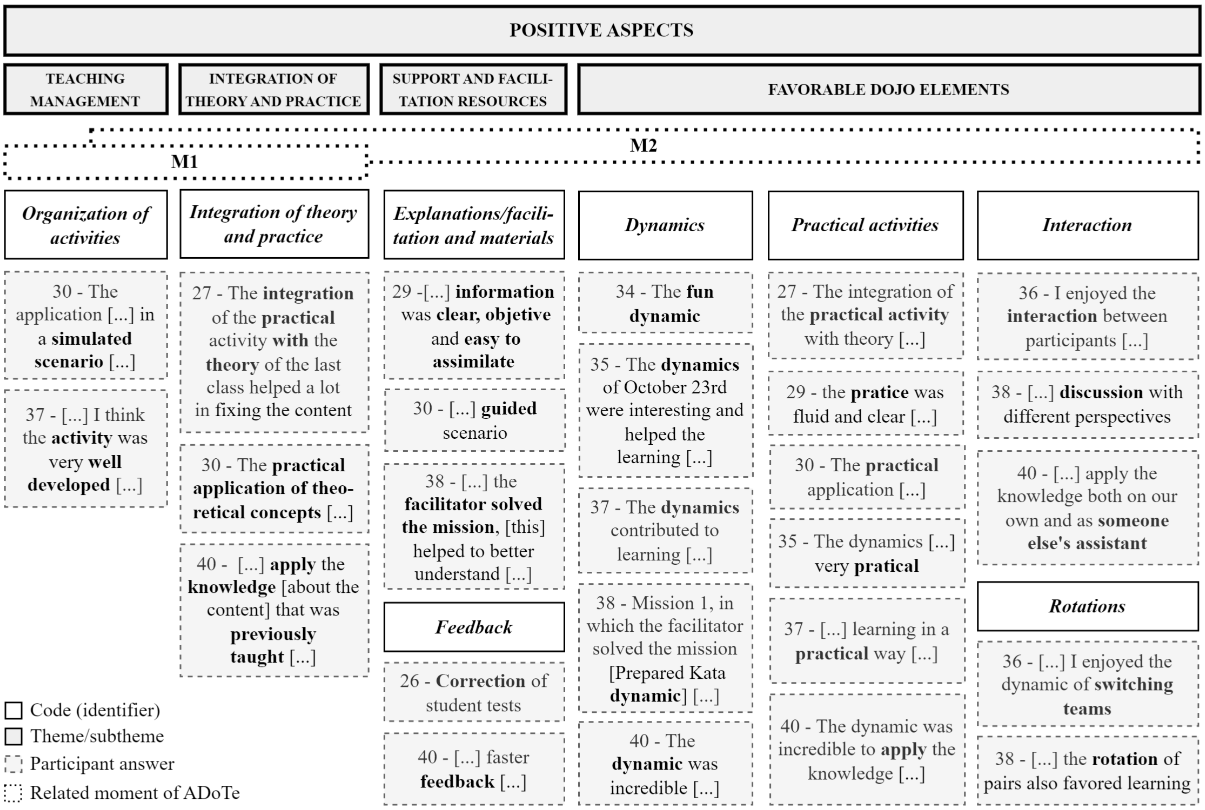
Figure 15 illustrates the results of the thematic analysis conducted on the answers of the experimental group from the second execution of the experiment to the first question of the retrospective questionnaire. Eight codes were established based on the answers to question 1. All codes relate to $M2$ (the moment of ADoTe with a practical focus) and two can also be associated with $M1$ (the moment with a theoretical focus). Furthermore, the codes were grouped into four subthemes and, finally, into a general theme: positive aspects. Concerning the subthemes:

- "Integration of theory and practice" groups positive mentions regarding the integration of theory (exposed in $M1$) and practice ($M2$), validating the execution of both moments of ADoTe.
- "Support and facilitation resources" refers to materials and actions that support the teaching and learning process. From its codes it is evident that participants also considered as positive: (a) explanations and mediation of learning by the teacher/facilitator; the information presented was also described as clear, objective, and easy to assimilate; and (b) the provision of feedback about the correctness of their answers. Regarding this aspect, one participant mentioned that feedback was provided more quickly (earlier than students would usually have in traditional teaching), which was also seen as positive. These considerations reinforce the importance of the guidelines $G2$ and $G9$ of ADoTe.
- "Favorable dojo elements" groups codes related to dojo elements that were favorable to learning. "Dynamics" answers express the approval of participants concerning a specific dojo format or the dynamics of this activity as a whole. The other codes indicate the approval of participants regarding practical activities, acting cooperatively and collaboratively (interactions), and pair rotation.
- "Teaching management" relates to aspects of management not included in the previous categories. In their responses, participants mentioned that the activity was very well developed and that some decisions (such as practice in a simulated scenario) benefited them. As the organization of activities results mainly from the planning stage of ADoTe, such considerations reinforce the relevance of this stage.

Concerning negative aspects and improvement suggestions, the answers collected in the second execution covered topics of $M1$ and $M2$ and could be grouped into four subthemes:

- "Unfavorable dojo elements" (for negative aspects) and "Dojo elements to be improved" (for improvement suggestions) groups the codes "Interaction" (participant ID 30 mentioned his unfamiliarity with other students as a negative aspect), "Cycle/iteration time" (considered short by participants with IDs 27, 29, 30, 34, and 40), and "Pressure" (ID 34 mentioned that little time to solve the activities results in pressure, which may be the reason why the PRS results were higher for G_E in the second execution). Concerning interactions, if necessary, the facilitator can conduct an integration activity before starting the dojo. In turn, by setting more assertive times for missions (following the guideline $G7$ of ADoTe) it is expected that participants feel less pressured.
- "Teaching management": some students mentioned that "Class time" was short (IDs 30, 36, 37, and 40) and thus suggested more class time (ID 36) and better time management (ID 38). This situation may have resulted from a failure from the teacher in the planning of the quantity of activities for a class. Due to the improvement cycle promoted by ADoTe, this situation can be adjusted in new executions of the approach.
- "Support and facilitation resources" groups the codes "Explanations/facilitation", "Materials", and "Examples". Some participants mentioned that explanations (IDs 29, 30, 35, and 38) and textual descriptions (IDs 35 and 41) were excessive, especially before the missions. Thus, they suggested more objectivity (IDs 30, 41; participant ID 35 suggested replacing instructional texts with more practical examples at the beginning of the dojo — i.e., missions in the Prepared Kata format —, which would be enough to understand how the next missions should be solved). Some participants also suggested more interactivity in $M1$ (ID 35), the execution of a mission in the Prepared Kata format with examples about boundary value analysis (ID 38), the provisioning of more details on how the solution of a mission should be documented (ID 40), and considered some explanations a little confusing (ID 38). However, as other participants mentioned the information was clear, objective, and easy to assimilate (ID 29), we realized that varied perceptions can be obtained for this aspect due to differences in student profiles (which reinforces the importance of the $G4$ guideline). Furthermore, as ADoTe promotes a PDCA cycle, situations of this type can be identified in the evaluation stage and subsequently considered in the planning of new uses of the approach.

Given the information presented, we can conclude that: (a) there was evidence of the acceptance of ADoTe by the students in both executions of the experiment, as mainly positive considerations were reported; (b) the answers of the first execution regarding negative aspects and improvement suggestions were useful for improving the guidelines of ADoTe; and (c) the answers of the second execution regarding negative aspects and improvement suggestions did not make ADoTe unfeasible and reinforced the importance of its stages and guidelines.



30 - The application [...] in a **simulated scenario** [...]

37 - [...] I think the **activity** was very **well developed** [...]

27 - The **integration** of the **practical** activity **with** the **theory** of the last class helped a lot in fixing the content

30 - The **practical application of theoretical concepts** [...]

40 - [...] **apply** the **knowledge** [about the content] that was **previously taught** [...]

29 -[...] **information** was **clear, objetive** and **easy to assimilate**

30 - [...] **guided** scenario

38 - [...] the **facilitator solved the mission**, [this] helped to better understand [...]

26 - **Correction** of student tests

40 - [...] faster **feedback** [...]

34 - The **fun dynamic**

35 - The **dynamics** of October 23rd were interesting and helped the learning [...]

37 - The **dynamics** contributed to learning [...]

38 - Mission 1, in which the facilitator solved the mission [Prepared Kata **dynamic**] [...]

40 - The **dynamic** was incredible [...]

27 - The integration of the **practical activity** with theory [...]

29 - the **pratice** was fluid and clear [...]

30 - The **practical** application [...]

35 - The dynamics [...] very **pratical**

37 - [...] learning in a **practical** way [...]

40 - The dynamic was incredible to **apply** the knowledge [...]

36 - I enjoyed the **interaction** between participants [...]

38 - [...] **discussion** with different perspectives

40 - [...] apply the knowledge both on our own and as **someone else's assistant**

36 - [...] I enjoyed the dynamic of **switching teams**

38 - [...] the **rotation** of pairs also favored learning

Figure 15. Second retrospective — Positive aspects

8 Discussion and Practical Implications

In this section, we discuss the study results. Subsection 8.1 presents a discussion of sample demographics. Subsection 8.2 details insights and accounts about the approach from the applicators of ADoTe. Finally, Subsection 8.3 answers the research questions.

8.1 Discussion of sample demographics

From the results presented in Subsection 7.1 it was possible to notice that the majority of participants (86.4%) in the experiment were male. The smaller number of participants with other genders is justified by the composition of the classes where recruitment took place, which had, for example, a low number of female students, a common scenario in science, technology, engineering, and mathematics (STEM) courses (Bowman et al., 2022).

Regarding academic background, most participants were in the final periods of an Information Systems course, since recruitment took place in classes of a subject offered (by the institution where the experiment was conducted) in the final periods of this course. These classes were chosen for convenience (as exposed in Subsection 5.3.1) and also to minimize the threat of a non-representative selection of the sample (see Section 9).

Furthermore, it was evidenced that most participants were young adults with little or no professional experience in software testing. In addition, regarding their level of prior knowledge about functional testing technique criteria (specifically equivalence partitioning, boundary value analysis, and systematic functional testing): (1) in the first execution several participants already had at least theoretical knowledge on the topic, because equivalence partitioning and boundary value analysis had already been taught (using a traditional teaching approach) in the class in which the experiment was conducted. Consequently, in this execution the knowledge of students about functional testing technique criteria was reinforced and complemented; and (2) in the second execution, for most participants their first contact with the aforementioned criteria occurred through the activities conducted in the experiment. Therefore, in this study it was possible to evaluate the impact of ADoTe as an approach to both primary and complementary teaching and learning of functional testing technique criteria.

The analysis of sample demographics allowed a better understanding of the context for which the results of this study were evidenced. Consequently, it also allowed us to identify possible threats to the validity of the results (see Section 9).

As already stated in Subsection 5.3, the experimental and control groups were balanced regarding the APC of their members and the number of participants. However, due to time constraints, other demographics were not considered in balancing and therefore might have impacted the results. These are non-exhaustive examples of possible impacts (which did not necessarily occur in this study): (1) participants with prior knowledge on functional testing technique criteria may have greater results in the pre-test and, consequently, smaller differences between their pre and post-test scores. Thus, an imbalance between groups regarding the prior knowledge of their members may influence learning results; and (2) participants without professional experience might have more difficulty in perceiving the value/usefulness of testing criteria in their careers. Thus, an imbalance regarding professional experience might influence the results of VAL (motivation subscale).

Finally, considering the discussion above, we visualize the opportunity (suggested as future work in Section 10) of conducting more evaluation experiments of ADoTe — to identify possible similarities or differences between scenarios — with sample demographics equal to or different from those of this study (e.g., varying and analyzing the correlation/influence of participants' professional experience, previous subject knowledge, gender, age, course, and course period on the results). In these experiments demographic variables should preferably be considered in balancing. Pre-study surveys (to assess participant demographics) and automation (to define the groups quickly) may be used as strategies to facilitate balancing even in scenarios with time constraints.

8.2 Insights from applicators

This subsection details insights and accounts about the approach from the authors/teachers who have applied ADoTe in the experiment conducted in this work.

First, regarding the need of preparation for using ADoTe, this is only necessary for the teacher/dojo facilitator. Before starting the planning stage activities, the person who will play this role must: (1) understand the concepts of kata and testing dojo, as well as the characteristic dojo elements (e.g., infrastructure, roles, moments, and formats). To this purpose, they can use the content of this work (such as the descriptions presented in Subsection 4.2 and in the experimental package) and, if necessary, complement their knowledge with the alternatives mentioned in Section 6 (supplementary material and participation in dojo sessions promoted by other professionals); (2) know the topic to be taught (i.e., functional testing technique criteria), seeking explanations in specialized literature (e.g., references of Subsection 4.1) if necessary; and (3) understand the base structure and guidelines of ADoTe, presented in Section 6.

In general, the application of ADoTe is not complex for teachers, since some activities in the approach (e.g., planning of the content to be taught, execution of a moment with theoretical focus, etc.) are already typical of this role in traditional teaching. However, a difficulty/challenge that can be faced by teachers is defining the ideal time to each iteration used to solve a kata/mission. As presented in Subsection 7.4, in the first execution of the experiment some students reported that the time to solve the missions was short. Thus, the mission times were revised and extended for the second execution. However, the results of the second execution indicated that several students still pointed out that the cycle/iteration times were short, resulting in pressure. Based on these results, we realized that when a new kata is used, it may take a few cycles of application of ADoTe to find the appropriate time for each iteration of this kata. Therefore, seeking to support the teacher, we suggest as future work (in Section 10) the creation of an instrument that helps to estimate or determine the ideal cycle/iteration times for any kata, as well as the development of a tool to support the approach.

Although both experimental groups were composed of eleven participants, ADoTe is scalable and can be used in classes with other numbers of students. However, as presented in guideline $G4$ of Section 6, it is important that the teacher considers the size of the group when defining the dojo formats to be used, i.e., in the planning stage of the approach. In scenarios with a larger number of students, for example, the Kake format is preferable, since it makes easier to meet the guideline $G5$ and avoids the idleness of the audience for long periods (as explained in Subsection 4.2.4).

As exposed in Subsection 7.1, the participants of the experiment were undergraduate students, most of them from the final periods of their courses. However, ADoTe can also be used in scenarios composed of participants with other levels of expertise (from beginners to advanced students). This is possible because, as presented in Section 6, ADoTe allows teachers to instantiate the approach according to their contexts (which includes students' level of expertise). Thus, for scenarios in which the students are less experienced, simpler katas and dojo formats with more mediation can be chosen in the ADoTe planning stage. In turn, in scenarios with advanced students (or with more knowledge about the content), katas that require more complex analyses/solutions can be used and the amount of mediation in the session can be reduced.

Finally, in Section 1 we presented the rationale why functional testing technique criteria was chosen as the focal topic of ADoTe. However, from the practical experiences obtained through this work, we realized that ADoTe can possibly be easily adapted for teaching and learning other topics (e.g., unit testing and exploratory testing, both mentioned in some studies as possible to be taught or learned through dojos, as presented in Section 3), since the content consists of an input artifact to the approach, which also presents several stages and guidelines not directly linked to the content to be taught, but rather to the activities to be conducted for teaching and learning. Thus, to further investigate this possibility, we also included in Section 10 a suggestion for future work to evaluate in details the feasibility, level of difficulty, and changes required to adapt ADoTe for teaching and learning other contents/subjects within or beyond the realms of software engineering or computing.

8.3 Answers to research questions

In this subsection, we discuss the study results, focusing on answering its research questions. To help the discussion, **Table 14** presents a summary of the quantitative results of the study. For each experiment execution and dependent variable are exposed: (a) the means of the results for the experimental (μ_{G_E}) and control (μ_{G_C}) groups; (b) the percentage difference between the means (Diff. μ)[14]; and (c) the p-value resulting from testing the hypotheses associated to the variable. Underlined μ_{G_E} and μ_{G_C} values correspond to the best results for each variable in the execution considered. Underlined p-values indicate statistical significance.

[14] Diff. $\mu = (|\mu_{G_E} - \mu_{G_C}| / \text{ amplitude}) * 100$. As amplitude value, we used: (a) for learning, 20 (since the results of this variable can vary from -10 to 10); (b) for motivation, 30 (given that results can range from -3 to 27); and (c) for the IMI subscales, 6 (since results can vary from 1 to 7).

Table 14. Summary of the quantitative results of the experiment

Variable	**1st execution**				**2nd execution**			
	μ_{G_E}	μ_{G_C}	Diff. μ	p-value	μ_{G_E}	μ_{G_C}	Diff. μ	p-value
Learning	2.0000	1.5455	2.27%	0.5823	4.6364	3.7273	4.54%	0.3714
Motivation	20.8655	19.3691	4.99%	0.3077	17.2027	16.3445	2.86%	0.7222
INT	5.8564	4.9082	15.80%	0.0192	5.3373	4.1545	19.71%	0.0376
CMP	5.4709	5.0009	7.83%	0.1808	4.5145	4.1973	5.29%	0.5681
PRS	1.8364	2.3818	9.09%	0.5767	3.2000	2.0182	19.70%	0.1679
EFF	5.0364	5.5818	9.09%	0.3241	4.4727	4.5818	1.82%	0.8855
VAL	6.3382	6.2600	1.30%	0.7582	6.0782	5.4291	10.82%	0.2372

Definition of the teaching and learning approach (RQ$_1$). RQ$_1$ referred to the definition of an approach supported by testing dojo to teaching and learning functional testing technique criteria in higher education.

As previously stated in sections 5.2 and 6, the definition of the approach (named ADoTe) was supported by literature data (including a SLM regarding the use of dojos in computing) and by the results of a controlled experiment conducted to evaluate ADoTe.

Data from the evaluation experiment indicated that (a) on average, in both executions participants submitted to ADoTe answered correctly 71.82% of the post-test questions (i.e., had a good performance after being submitted to the approach), (b) the learning and general motivation averages of the groups submitted to ADoTe were positive and greater than those of the groups submitted to traditional teaching (more details will be presented in the discussions of RQ$_2$ and RQ$_3$), and (c) the answers to the retrospective questionnaire showed that ADoTe was well accepted by students and reinforced the importance of its stages and guidelines.

Thus, it can be stated that ADoTe meets its objective, consisting of an approach (with positive results regarding student learning and motivation) that can be used to teach and learn functional testing technique criteria in higher education.

Impact of ADoTe on student learning (RQ$_2$). To answer RQ$_2$ we analyzed the impact of ADoTe on student learning, comparing its results to those obtained from a traditional teaching approach.

Data from the evaluation experiment indicated that participants answered more knowledge questions correctly (20.00% and 46.36% in the first and second executions, respectively) after being submitted to ADoTe, which indicates a positive impact of the approach on student learning. The first percentage is less expressive, because, as exposed in Subsection 7.1, in the first execution participants had already had contact with part of the content previously. Thus, they had greater results in the pre-test and, consequently, the differences between their pre and post-test scores were smaller.

Furthermore, from **Table 14** it is possible to notice that the learning averages of the experimental groups were 2.27% and 4.54% greater than those of the control groups. Despite this, the null hypothesis $2H_0$ could not be rejected, since p-value results were greater than α (0.05). Thus, (a) statistically, it was not possible to state that the learning of students who study functional testing technique criteria through an approach supported by testing dojo is greater than the learning of those who study the same content through a traditional teaching approach, and (b) in practice, there is no guarantee that, when using ADoTe, students' learning will be greater than that they would obtain from a traditional teaching approach.

Impact of ADoTe on student motivation (RQ$_3$). To answer RQ$_3$ we analyzed the impact of ADoTe on student motivation, comparing its results to those obtained from a traditional teaching approach.

Data from the evaluation experiment indicated that the motivation levels of the groups submitted to ADoTe were around 79.55% and 67.34% (i.e., higher than the percentage of neutrality, 50%), which indicates positive perceptions of the participants about the approach. Furthermore, **Table 14** shows that the motivation means of the experimental groups were 4.99% and 2.86% greater than those of the control groups. Despite this, the null hypothesis $3H_0$ could not be rejected, since p-value results were greater than α (0.05). Thus, (a) statistically, it was not possible to state that the general motivation of students who study functional testing technique criteria through an approach supported by testing dojo is greater than the motivation of those who study the same content through a traditional teaching approach, and (b) in practice, there is no guarantee that, when using ADoTe, students' general motivation will be greater than that they would feel if submitted to a traditional teaching approach.

Concerning the IMI subscales, **Table 14** shows that in both executions the means of G_E were higher than those of G_C for INT, CMP, and VAL (the opposite occurred for EFF). For PRS the best result (lowest average) was verified for G_E in the first execution and for G_C in the second (which may have resulted from the perceptions of participants about the time of the missions and the class, as explained in Subsection 7.4). Regarding the hypotheses, $3.2H_0$, $3.3H_0$, $3.4H_0$, and $3.5H_0$ could not be rejected, as p-value $> \alpha$ for CMP, PRS, ESF, and VAL. Therefore, in practice, students' perception of the aspects related to these subscales will not necessarily be greater when using ADoTe (compared to what they would feel if submitted to a traditional teaching approach).

In turn, $3.1H_1$ was accepted, since INT results were statistically significant (for this subscale p-value $< \alpha$ and, in executions, μ_{G_E} was 15.80% and 19.71% greater than μ_{G_C}). Thus, it can be stated that the interest/enjoyment (aspect related to intrinsic motivation) of students submitted to ADoTe is greater than that of students submitted to a traditional teaching approach. Additionally, the effect sizes (i.e., the practical significance of the results) were large, since the values of Hedges' g_s were equal to 1.04 and 0.91.

9 Threats to validity and limitations

Below we describe the threats to the validity of our study (mainly related to the experiment) according to the categories specified in Wohlin et al. (2012) and the actions taken to mitigate or neutralize them.

Conclusion validity. To mitigate threats related to statistical issues, the definition of statistical tests and the verification of their assumptions were based on specialized literature (Barbetta et al., 2010; Wohlin et al., 2012). Regarding the reliability of measures, it was defined in the planning stage of the experiment how the study variables would be evaluated. Additionally, objective measures were collected from instruments based on or already validated in other studies. The instruments were also discussed with two professors with experience in Software Engineering teaching and research. To minimize risks related to the reliability of treatment implementation, the process and experimental package were documented in detail, making it possible to conduct the study multiple times in a standardized way. Finally, to mitigate random irrelevancies in the experimental setting (e.g., noises and interruptions) participants were instructed to avoid inappropriate conversations. Furthermore, at the beginning of each execution, a few minutes were waited before starting the activities, minimizing interruptions due to participant delays.

Internal validity. In experiments, it is essential that the outcomes (effects) result from the treatment (cause) and not from other aspects (Wohlin et al., 2012). Regarding the threats that can impact this causal relationship, the following actions were taken: (a) to reduce the impact of APC on learning outcomes, the experimental and control groups were balanced according to this variable; (b) to mitigate biases related to the repetition of tests, we changed the statements or answer alternatives between knowledge tests. Furthermore, test answers were not made available to the participants, so that they would not influence their performance; (c) to mitigate instrumentation problems (e.g., poor formulation), the instruments were discussed/reviewed by more than one researcher; (d) participation in the study was voluntary and not mandatory. To mitigate the threat that the selected sample would not represent the population (as volunteers are usually more motivated), we highlighted the importance of participation when recruiting subjects. Furthermore, as recruitment took place in pre-defined groups and execution occurred during class time (without dismissal of non-participants, but rather compensation with other activities), we believe that students who would not be volunteers if the invitation and operation occurred in another way also participated in the experiment; (e) as participants were allowed to withdraw from the study at any time (due to ethical requirements), to mitigate mortality they were informed about the importance of not withdrawing; (f) to mitigate threats related to interactions of participants with researchers or secondary elements, we sought to limit differences between groups to the teaching and learning approaches, using the same exercises, as well as presenting equivalent explanations/answers to participants from both groups; (g) to meet ethical requirements, participants were informed in advance, at a high level, about what would happen in the experiment. However, to mitigate threats related to the knowledge of treatments, participants were not informed about the study hypotheses, the names or details of the treatments, nor which treatment they were submitted. Furthermore, participants were asked not to comment with students from the other group about the activities to which they were being submitted; (h) history threats were neutralized, since each participant was submitted to only one treatment and only once; (i) compensatory threats were neutralized, since students were not rewarded or punished for participating or not in the study; and (j) since demographic data and the choice of specific dojo formats/kata types together with the complexity of the kata itself may impact the dependent variables, we suggested (in Section 10) conducting more evaluation experiments of ADoTe (varying the katas, dojo formats, and sample demographics) to identify possible similarities or differences between scenarios. Furthermore, in Subsection 8.1 we also recommended considering demographic variables in balancing and proposed strategies for incorporating them into future experiments, even under time constraints.

External validity. Threats to external validity limit the ability to generalize results outside the experimental environment (Wohlin et al., 2012). The following are risks related to this category: (a) non-representative selection of the sample: to minimize this threat, the sample was composed of direct participants from the target population (students who were enrolled in a subject in which functional testing technique criteria are often taught). However, as recruitment took place in a single course and institution, this may be a threat to the generalization of results to different courses and institutions (due to student profile and location); (b) execution of the experiment in a scenario that does not represent reality: to minimize this threat, the experiment took place in a real environment (in the classroom, during Software Engineering classes). However, due to time limitations, the activities did not include higher levels of complexity (existing in industrial settings). Therefore, there may be threats to the generalization of the results to distinct scenarios from that investigated in this work (such as the industrial one); and (c) impact of time: participants were submitted to the treatments during a limited time. Furthermore, learning was measured immediately after applying the treatments. Therefore, the study results refer to immediate learning, and their generalization to other contexts is not guaranteed. However, as time limitation for the teaching of testing is common in academic contexts (Garousi et al., 2020b; Melo et al., 2020; Valle et al., 2015a), we believe that the execution of the experiment considering the time usually available for teaching, as carried out in this study, is important for the results to be realistic.

Construct validity. Threats to construct validity are typically of two types: (a) design threats: to mitigate them, the guidelines proposed by Wohlin et al. (2012) were followed. In addition, study decisions were validated and reviewed by two researchers with experience in Software Engineering research and knowledge about dojos and software testing teaching; and (b) social threats: in experiments, participants may act differently than they normally would (e.g., for fear of evaluations, they may try to appear as if they are performing better than reality; when presupposing the study hypotheses, they may try to manipulate the results to favor or disfavor a treatment). Thus, to mitigate these threats, participants were informed about the importance of their ethical participation and that the data collected would be used to evaluate the treatments, not their performances individually (data would not be used to favor or harm participants and information that could identify them would not be published).

Regarding **research limitations**, the limited time available for the application and evaluation of ADoTe resulted in the following restrictions in the experiment (some already mentioned above): recruitment of participants in two classes of a single course and institution; measurement of learning right after applying the treatments; adaptation of the activities to the teaching context, but without considering higher levels of complexity (existing in industrial settings); and approach of only three testing criteria (equivalence partitioning, boundary value analysis, and systematic functional testing). Despite this, ADoTe is not limited to these three criteria, as it allows the organization of activities related to any criteria of the functional technique. Furthermore, aspects that suffered limitations in this study may be overcome in future work.

10 Conclusions and future work

This paper contributes to the improvement of software testing education by defining and evaluating ADoTe, an approach supported by testing dojo to teaching and learning functional testing technique criteria in higher education.

The approach definition was supported by literature data (including a SLM regarding the use of dojos in computing) and by the results of the evaluation of ADoTe. The analysis of such information allowed the definition of an approach that can be adapted to different teaching contexts and enables the continuous improvement of the teaching and learning process.

To evaluate ADoTe, we conducted two executions of a controlled experiment. From the analysis of data collected from 44 participants we noticed that, although the averages for learning and general motivation of the groups submitted to ADoTe were positive and greater than those of the groups submitted to a traditional teaching approach, it was not possible to state that, for these variables, there was a statistically significant difference between the approaches evaluated. In turn, the results of the IMI subscale related to intrinsic motivation were statistically significant, indicating that students feel more interest/enjoyment in learning functional testing technique criteria through ADoTe (the averages of the groups submitted to ADoTe were 15.80% and 19.71% higher than those of the control groups). Furthermore, the results of thematic analyses conducted on the answers to a retrospective questionnaire showed that ADoTe was well accepted by students and reinforced the importance of its elements.

Therefore, considering the evaluation context of this study, we conclude that ADoTe positively impacts the learning and motivation of students and, compared to traditional teaching, tends to lead to greater levels of interest/enjoyment in learning functional testing technique criteria.

As future work, we suggest (i) conducting more evaluation experiments of ADoTe (to identify possible similarities or differences between scenarios, correlations and influence factors): (a) in more classes, institutions, and with sample demographics equal to or different from those of this study; (b) with different activities (e.g., varying the katas and dojo formats, including katas related to other functional testing criteria and with different levels of complexity); and (c) in industrial settings, (ii) verifying the effects of ADoTe regarding long-term learning, (iii) creating an instrument (e.g., automated tool, Artificial Intelligence model, or formula) that helps to estimate or determine the ideal cycle/iteration times for any kata, (iv) evaluating the feasibility, level of difficulty, and changes required to adapt ADoTe for teaching and learning other contents/subjects within or beyond the realms of software engineering or computing, (v) conducting longitudinal studies to evaluate the sustained impact of the approach on learning and motivation over multiple semesters, (vi) investigating the effectiveness of ADoTe in online and hybrid learning environments, since these are becoming increasingly important in higher education, and (vii) developing a tool to support the approach. This tool can guide the research throughout the dojo testing process, provide an environment to organize the dojo sessions, collect, summarize, and share data, facilitating future replications, comparisons, and results analysis. Its development can be iterative: in the initial iterations, resources may be implemented for storing and retrieving katas, participant demographics, feedback from retrospectives, and general data about the sessions (e.g., participants, katas solved, and time allocated to each step/cycle); in later iterations, features can be implemented for the analysis and sharing of data collected, recommendation of katas and the size of its steps (based on the teaching and learning context, integrating resources such as the suggested in (iii)), and provision of additional functionalities that are perceived as useful over time.

References

Aniche, M., Hermans, F., and van Deursen, A. (2019). Pragmatic software testing education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pages 414–420, New York. ACM.

Aniche, M. F. and Silveira, G. d. A. (2011). Increasing learning in an agile environment: Lessons learned in an agile team. In *2011 Agile Conference*, pages 289–295.

Bache, E. (2013). *The Coding Dojo Handbook: a practical guide to creating a space where good programmers can become great programmers*. Leanpub, 1 edition.

Barbetta, P. A., Reis, M. M., and Bornia, A. C. (2010). *Es-*

tatística Para Cursos de Engenharia e Informática. Atlas, São Paulo, 3 edition.

Bossavit, L. and Gaillot, E. (2005). The coder's dojo – A different way to teach and learn programming. In Baumeister, H., Marchesi, M., and Holcombe, M., editors, *Extreme Programming and Agile Processes in Software Engineering*, pages 290–291, Berlin, Heidelberg. Springer.

Bowman, N. A., Logel, C., LaCosse, J., Jarratt, L., Canning, E. A., Emerson, K. T. U., and Murphy, M. C. (2022). Gender representation and academic achievement among STEM-interested students in college STEM courses. *Journal of Research in Science Teaching*, 59(10):1876–1900.

Bravo, M. and Goldman, A. (2010). Reinforcing the learning of agile practices using coding dojos. In Sillitti, A., Martin, A., Wang, X., and Whitworth, E., editors, *Agile Processes in Software Engineering and Extreme Programming*, pages 379–380, Berlin, Heidelberg. Springer.

Carpinetti, L. C. R. and Gerolamo, M. C. (2016). *Gestão da qualidade ISO 9001: 2015: requisitos e integração com a ISO 14001:2015*. Atlas, Rio de Janeiro, 1 edition.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, New York, 2 edition.

Costa, I. E. F. and Oliveira, S. R. B. (2019). A systematic strategy to teaching of exploratory testing using gamification. In *Proceedings of the 14th International Conference on Evaluation of Novel Approaches to Software Engineering*, pages 307–314. SciTePress.

Costa, I. E. F. and Oliveira, S. R. B. (2022). Development of a teaching plan to support learning activities of exploratory test design and execution. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.

Costa, I. E. F., Oliveira, S. R. B., Elgrably, I. S., Guerra, A. d. S., Soares, E. M., and Costa, I. V. F. (2023). Using active methodologies for teaching and learning of exploratory test design and execution. *Education Sciences*, 13(2).

CSDT (2023). Intrinsic Motivation Inventory (IMI). Center for Self-Determination Theory (CSDT).

de Oliveira, C. M. C., Canedo, E. D., Faria, H., Amaral, L. H. V., and Bonifácio, R. (2018). Improving student's learning and cooperation skills using coding dojos (in the wild!). In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.

Delamaro, M. E., Vincenzi, A. M. R., and Maldonado, J. C. (2006). A strategy to perform coverage testing of mobile applications. In *Proceedings of the 2006 International Workshop on Automation of Software Test*, pages 118–124. ACM.

Dybå, T., Dingsoyr, T., and Hanssen, G. K. (2007). Applying systematic reviews to diverse study types: An experience report. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 225–234.

Elgrably, I. S. and de Oliveira, S. R. B. (2021). A diagnosis on software testing education in the brazilian universities. In *2021 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE Press.

Elgrably, I. S. and Oliveira, S. R. B. (2018). Gamification and evaluation of the use the agile tests in software quality subjects: The application of case studies. In *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*, pages 416—423. SciTePress.

Elgrably, I. S. and Oliveira, S. R. B. (2020). Model for teaching and training software testing in an agile context. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE Press.

Elgrably, I. S. and Oliveira, S. R. B. (2022a). Perception from the professors' point of view in the remote teaching of software testing using active methodologies during the covid-19 pandemic. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.

Elgrably, I. S. and Oliveira, S. R. B. (2022b). A quasi-experimental evaluation of teaching software testing in software quality assurance subject during a post-graduate computer science course. *International Journal of Emerging Technologies in Learning*, 17(5):57–86.

Estácio, B., Kroll, J., and Prikladnicki, R. (2016a). Distributed coding dojo randori: An overview and research opportunities. In *Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems*.

Estácio, B., Oliveira, R., Marczak, S., Kalinowski, M., Garcia, A., Prikladnicki, R., and Lucena, C. (2015a). Evaluating collaborative practices in acquiring programming skills: Findings of a controlled experiment. In *2015 29th Brazilian Symposium on Software Engineering*, pages 150–159.

Estácio, B., Valentim, N., Rivero, L., Conte, T., and Prikladnicki, R. (2015b). Evaluating the use of pair programming and coding dojo in teaching mockups development: An empirical study. In *2015 48th Hawaii International Conference on System Sciences*, pages 5084–5093.

Estácio, B., Zieris, F., Prechelt, L., and Prikladnicki, R. (2016b). On the randori training dynamics. In *Proceedings of the 9th International Workshop on Cooperative and Human Aspects of Software Engineering*, pages 44—47, New York, NY, USA. ACM.

Felizardo, K. R., Nakagawa, E. Y., Fabbri, S. C. P. F., and Ferrari, F. C. (2017). *Revisão sistemática da literatura em engenharia de software: teoria e prática*. Elsevier, Rio de Janeiro, 1 edition.

Filho, A. F. G. and de Toledo, R. (2015). Visual management and blind software developers. In *2015 Agile Conference*, pages 31–39.

Fonseca, F. M., Silva, E. B. d., and Mendonça, D. S. (2019). Designing dojo: A collaborative method for teaching design patterns. In *2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 39–40.

Furtado, L. S. and Oliveira, S. R. B. (2020). A teaching proposal for the software measurement process using gamification: an experimental study. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.

Gaertner, M. (2010). Testing dojos. *Methods & Tools*, 18(4):8–14.

Garcia, F. W. d. S. G., Oliveira, S. R. B., and Carvalho, E. d. C. (2022). Application of a teaching plan for algorithm subjects using active methodologies: An experimental report. *International Journal of Emerging Technologies in Learning*, 17(7):175–207.

Garcia, F. W. d. S. G., Oliveira, S. R. B., and Carvalho, E. d. C. (2023). A second experimental study the application of a teaching plan for the algorithms subject in an undergraduate course in computing using active methodologies. *Informatics in Education*, 22(2):233–255.

Garousi, V., Felderer, M., Kuhrmann, M., Herkiloğlu, K., and Eldh, S. (2020a). Exploring the industry's challenges in software testing: An empirical study. *Journal of Software Evolution and Process*, 32(8).

Garousi, V., Rainer, A., Lauvås, P., and Arcuri, A. (2020b). Software-testing education: A systematic literature mapping. *Journal of Systems and Software*, 165.

Gehringer, E. F. (2007). Active and collaborative learning strategies for teaching computing. In *Proceedings of the 2007 ASEE Annual Conference & Exposition*, pages 12.167.1–12.167.13.

Gomes, R. F. and Lelli, V. (2021). GAMUT: GAMe-Based Learning Approach for Teaching Unit Testing. In *XX Brazilian Symposium on Software Quality*, pages 1–11, New York. ACM.

Gregory, J. and Crispin, L. (2014). *More Agile Testing: Learning Journeys For The Whole Team*. Addison-Wesley Professional, New Jersey, 1 edition.

Heinonen, K., Hirvikoski, K., Luukkainen, M., and Vihavainen, A. (2013). Learning agile software engineering practices using coding dojo. In *Proceedings of the 14th Annual ACM SIGITE Conference on Information Technology Education*, pages 97—102, New York, NY, USA. ACM.

Jesus, G. M., Paschoal, L. N., Ferrari, F. C., and Souza, S. R. S. (2019). Is it worth using gamification on software testing education? An experience report. In *Proceedings of the XVIII Brazilian Symposium on Software Quality*, pages 178–187, New York. ACM.

Kassab, M., DeFranco, J. F., and Laplante, P. A. (2017). Software testing: The state of the practice. *IEEE Software*, 34(5):46–52.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report version 2.3, Keele University and University of Durham Joint Technical Report, UK. EBSE Technical Report EBSE-2007-01.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(1):1–12.

Lee, Y., Marepalli, D. B., and Yang, J. (2017). Teaching test-drive development using dojo. *Journal of Computing Sciences in Colleges*, 32(4):106—112.

Linkman, S., Vincenzi, A. M. R., and Maldonado, J. C. (2003). An evaluation of systematic functional testing using mutation testing. In *Proceedings of 7th InternationalConference on Empirical Assessment in Software Engineering*, pages 1–15.

Luz, R. B. d., Neto, A. G. S. S., and Noronha, R. V. (2013). Teaching TDD, the coding dojo style. In *2013 IEEE 13th International Conference on Advanced Learning Technologies*, pages 371–375.

McAuley, E., Duncan, T., and Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1):48–58.

McConnell, J. J. (2005). Active and cooperative learning. *SIGCSE Bull.*, 37(2):27–30.

Meireles, M., de Souza, C., de Barros, F. C., Chaves, L., de Castro, R., and Giuntini, F. (2022a). The employment of testing dojo as a collaborative learning methodology for teaching failure analysis: An experience report. In *2022 4th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 47–54.

Meireles, M. A. C., Filho, A. R. L., Lima, K. R. d. S., Ferraz, L. G. d. C., Batista, F. A., Barros, F. C. P. d., Chaves, L. C., Souza, C. d., and Roque, L. F. N. d. M. (2022b). Use of testing dojo as a methodology of collaborative learning in teaching testcase writing: An experience report. In *Proceedings of the 13th International Conference on Society and Information Technologies (ICSIT 2022)*, pages 41–44.

Melo, S. M., Moreira, V. X. S., Paschoal, L. N., and Souza, S. R. S. (2020). Testing education: a survey on a global scale. In *Proceedings of the 34th Brazilian Symposium on Software Engineering*, pages 554–563, New York. ACM.

Melo, S. M., Santos, I., Souza, P. S. L., and Souza, S. R. S. (2022). A survey on the practices of software testing: a look into brazilian companies. *Journal of Software Engineering Research and Development*, 10(8):11:1–11:15.

Myers, G. J., Badgett, T., and Sandler, C. (2012). *The Art of Software Testing*. John Wiley & Sons, Inc, New Jersey, 3 edition.

Oliveira, R., Estácio, B., Garcia, A., Marczak, S., Prikladnicki, R., Kalinowski, M., and Lucena, C. (2016). Identifying code smells with collaborative practices: A controlled experiment. In *2016 X Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS)*, pages 61–70.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*, 93(3):223–231.

Robson, C. and McCartan, K. (2016). *Real World Research*. John Wiley & Sons Ltd, 4 edition.

Rodrigues, P. L. R., Franz, L. P., Cheiran, J. F. P., da Silva, J. P. S., and Bordin, A. S. (2017). Coding dojo as a transforming practice in collaborative learning of programming: An experience report. In *Proceedings of the XXXI Brazilian Symposium on Software Engineering (SBES)*, pages 348––357, New York. ACM.

Rooksby, J., Hunt, J., and Wang, X. (2014). The theory and practice of randori coding dojos. In Cantone, G. and Marchesi, M., editors, *Agile Processes in Software Engineering and Extreme Programming*, pages 251–259, Cham. Springer International Publishing.

Santos, E. a. and Oliveira, S. (2019). The use of game elements and scenarios for teaching and learning the function point analysis technique: A experimental study. In *Proceedings of the 14th International Conference on Software Technologies (ICSOFT)*, pages 162—169. SciTePress.

Santos, V., Goldman, A., and de Souza, C. R. B. (2015). Fostering effective inter-team knowledge sharing in agile software development. *Empirical Software Engineering*, 20:1006–1051.

Santos, V. A., Goldman, A., and Santos, C. D. (2012). Uncovering steady advances for an extreme programming course. *CLEI Electronic Journal*, 15.

Sato, D. T., Corbucci, H., and Bravo, M. V. (2008). Coding dojo: An environment for learning and sharing agile practices. In *Agile 2008 Conference*, pages 459–464. IEEE Press.

Scatalon, L. P., Fioravanti, M. L., Prates, J. M., Garcia, R. E., and Barbosa, E. F. (2018). A survey on graduates' curriculum-based knowledge gaps in software testing. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE Press.

Seaman, C. B. (2008). *Qualitative Methods*, chapter 2, pages 35–62. Springer-Verlag, London.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sherif, E., Liu, A., Nguyen, B., Lerner, S., and Griswold, W. G. (2020). Gamification to aid the learning of test coverage concepts. In *2020 IEEE 32nd Conference on Software Engineering Education and Training*, pages 1–5. IEEE Press.

Soomlek, C. (2015). Applying randori-style kata and agile practices to an undergraduate-level programming class. In *XP 2015: 16th International Conference on Agile Processes in Software Engineering and Extreme Programming*, pages 369–370.

Valle, P. H. D., Barbosa, E. F., and Maldonado, J. C. (2015a). CS curricula of the most relevant universities in Brazil and abroad: Perspective of software testing education. In *2015 International Symposium on Computers in Education (SIIE)*, pages 62–68. IEEE Press.

Valle, P. H. D., Barbosa, E. F., and Maldonado, J. C. (2015b). Um mapeamento sistemático sobre ensino de teste de software. In *Anais do XXVI Simpósio Brasileiro de Informática na Educação (SBIE 2015)*, pages 71–80.

Weyuker, E. and Jeng, B. (1991). Analyzing partition testing strategies. *IEEE Transactions on Software Engineering*, 17(7):703–711.

Weyuker, E. and Ostrand, T. (1980). Theories of program testing and the application of revealing subdomains. *IEEE Transactions on Software Engineering*, SE-6(3):236–2461.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *EASE'14: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, pages 1–10, New York. ACM.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer-Verlag, Berlin, 1 edition.

Appendix A Supplementary material to the SLM

This appendix presents supplementary material to the SLM detailed in Section 2: the search strings used in each research source (in Section A.1) and the set of studies considered in the SLM (in Section A.2).

A.1 Search strings

Below are listed the search strings used for automatic searches in each research source:

- **IEEE Xplore (ieeexplore.ieee.org):** (“Document Title”:teaching OR “Abstract”:teaching OR “Author Keywords”:teaching OR “Document Title”:training OR “Abstract”:training OR “Author Keywords”:training OR “Document Title”:learning OR “Abstract”:learning OR “Author Keywords”:learning OR “Document Title”:education OR “Abstract”:education OR “Author Keywords”:education) AND dojo NOT (“dojo toolkit”)
- **ACM Digital Library (dl.acm.org):** (Title:(teach*) OR Abstract:(teach*) OR Keyword:(teach*) OR Title:(learn*) OR Abstract:(learn*) OR Keyword:(learn*) OR Title:(train*) OR Abstract:(train*) OR Keyword:(train*) OR Title:(educat*) OR Abstract:(educat*) OR Keyword:(educat*)) AND dojo NOT (“dojo toolkit”)
- **Scopus (www.scopus.com):** TITLE-ABS-KEY ((teach* OR learn* OR train* OR educat*)) AND ALL (dojo OR dojos) AND NOT (“dojo toolkit”) AND (LIMIT-TO (SUBJAREA,“COMP”))
- **Compendex (www.engineeringvillage.com):** ((((teach* OR learn* OR train* OR educat*) WN KY) AND ((dojo) WN ALL)) NOT ((dojo toolkit) WN ALL))
- **Science Direct (www.sciencedirect.com):** “Find articles with these terms” = dojo NOT “dojo toolkit” + “Title, abstract or author-specified keywords” = teaching OR learning OR training OR education + filters: article type = review articles OR research articles; subject areas = computer science
- **Web of Science (webofknowledge.com):** (TI=(teach*) OR AB=(teach*) OR AK=(teach*) OR TI=(learn*) OR AB=(learn*) OR AK=(learn*) OR TI=(train*) OR AB=(train*) OR AK=(train*) OR TI=(educat*) OR AB=(educat*) OR AK=(educat*)) AND ALL=(dojo OR dojos NOT “dojo toolkit”) + filtering by Web of Science categories = Computer Science Theory Methods, Computer Science Software Engineering, Computer Science Information Systems, Computer Science Interdisciplinary Applications, Education Educational Research, and Education Scientific Disciplines

A.2 Set of studies considered in the SLM

Below are listed the studies considered in the SLM for data extraction and synthesis:

1. Bossavit, L. and Gaillot, E. (2005). The coder’s dojo – A different way to teach and learn programming. In Baumeister, H., Marchesi, M., and Holcombe, M., editors, *Extreme Programming and Agile Processes in Software Engineering*, pages 290–291, Berlin, Heidelberg. Springer.
2. Sato, D. T., Corbucci, H., and Bravo, M. V. (2008). Coding dojo: An environment for learning and sharing agile practices. In *Agile 2008 Conference*, pages 459–464. IEEE Press.
3. Bravo, M. and Goldman, A. (2010). Reinforcing the learning of agile practices using coding dojos. In Sillitti, A., Martin, A., Wang, X., and Whitworth, E., editors, *Agile Processes in Software Engineering and Extreme Programming*, pages 379–380, Berlin, Heidelberg. Springer.
4. Santos, V. A., Goldman, A., and Santos, C. D. (2012). Uncovering steady advances for an extreme programming course. *CLEI Electronic Journal*, 15.
5. Santos, V., Goldman, A., and de Souza, C. R. B. (2015). Fostering effective inter-team knowledge sharing in agile software development. *Empirical Software Engineering*, 20:1006–1051.
6. Gaertner, M. (2010). Testing dojos. *Methods & Tools*, 18(4):8–14.
7. Aniche, M. F. and Silveira, G. d. A. (2011). Increasing learning in an agile environment: Lessons learned in an agile team. In *2011 Agile Conference*, pages 289–295.
8. Luz, R. B. d., Neto, A. G. S. S., and Noronha, R. V. (2013). Teaching TDD, the coding dojo style. In *2013 IEEE 13th International Conference on Advanced Learning Technologies*, pages 371–375.
9. Heinonen, K., Hirvikoski, K., Luukkainen, M., and Vihavainen, A. (2013). Learning agile software engineering practices using coding dojo. In *Proceedings of the 14th Annual ACM SIGITE Conference on Information Technology Education*, pages 97—102, New York, NY, USA. ACM.
10. Rooksby, J., Hunt, J., and Wang, X. (2014). The theory and practice of randori coding dojos. In Cantone, G. and Marchesi, M., editors, *Agile Processes in Software Engineering and Extreme Programming*, pages 251–259, Cham. Springer International Publishing.
11. Estácio, B., Valentim, N., Rivero, L., Conte, T., and Prikladnicki, R. (2015). Evaluating the use of pair programming and coding dojo in teaching mockups development: An empirical study. In *2015 48th Hawaii International Conference on System Sciences*, pages 5084–5093.
12. Estácio, B., Oliveira, R., Marczak, S., Kalinowski, M., Garcia, A., Prikladnicki, R., and Lucena, C. (2015). Evaluating collaborative practices in acquiring programming skills: Findings of a controlled experiment. In *2015 29th Brazilian Symposium on Software Engineering*, pages 150–159.
13. Estácio, B., Kroll, J., and Prikladnicki, R. (2016). Distributed coding dojo randori: An overview and research opportunities. In *Workshop on Distributed Software Development*, Software Ecosystems and Systems-of-Systems.
14. Estácio, B., Zieris, F., Prechelt, L., and Prikladnicki, R. (2016). On the randori training dynamics. In *Proceed-*

ings of the 9th International Workshop on Cooperative and Human Aspects of Software Engineering, pages 44––47, New York, NY, USA. ACM.

15. Oliveira, R., Estácio, B., Garcia, A., Marczak, S., Prikladnicki, R., Kalinowski, M., and Lucena, C. (2016). Identifying code smells with collaborative practices: A controlled experiment. In *2016 X Brazilian Symposium on Software Components, Architectures and Reuse (SBCARS)*, pages 61–70.
16. Filho, A. F. G. and Toledo, R. (2015). Visual management and blind software developers. In 2015 *Agile Conference*, pages 31–39.
17. Soomlek, C. (2015). Applying randori-style kata and agile practices to an undergraduate-level programming class. In *XP 2015: 16th International Conference on Agile Processes in Software Engineering and Extreme Programming*, pages 369–370.
18. Lee, Y., Marepalli, D. B., and Yang, J. (2017). Teaching test-drive development using dojo. *Journal of Computing Sciences in Colleges*, 32(4):106—112.
19. Rodrigues, P. L. R., Franz, L. P., Cheiran, J. F. P., da Silva, J. P. S., and Bordin, A. S. (2017). Coding dojo as a transforming practice in collaborative learning of programming: An experience report. In *Proceedings of the XXXI Brazilian Symposium on Software Engineering (SBES)*, pages 348—357, New York. ACM.
20. de Oliveira, C. M. C., Canedo, E. D., Faria, H., Amaral, L. H. V., and Bonifácio, R. (2018). Improving student's learning and cooperation skills using coding dojos (in the wild!). In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.
21. Elgrably, I. S. and Oliveira, S. R. B. (2018). Gamification and evaluation of the use the agile tests in software quality subjects: The application of case studies. In *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE)*, pages 416—423. SciTePress.
22. Santos, E. a. and Oliveira, S. (2019). The use of game elements and scenarios for teaching and learning the function point analysis technique: A experimental study. In *Proceedings of the 14th International Conference on Software Technologies (ICSOFT)*, pages 162—169. SciTePress.
23. Furtado, L. S. and Oliveira, S. R. B. (2020). A teaching proposal for the software measurement process using gamification: an experimental study. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.
24. Elgrably, I. S. and Oliveira, S. R. B. (2020). Model for teaching and training software testing in an agile context. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE Press.
25. Elgrably, I. S. and Oliveira, S. R. B. (2022). A quasi-experimental evaluation of teaching software testing in software quality assurance subject during a post-graduate computer science course. *International Journal of Emerging Technologies in Learning*, 17(5):57–86.
26. Garcia, F. W. d. S. G., Oliveira, S. R. B., and Carvalho, E. d. C. (2022). Application of a teaching plan for algorithm subjects using active methodologies: An experimental report. *International Journal of Emerging Technologies in Learning*, 17(7):175–207.
27. Fonseca, F. M., Silva, E. B. d., and Mendonça, D. S. (2019). Designing dojo: A collaborative method for teaching design patterns. In *2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 39–40.
28. Meireles, M. A. C., Filho, A. R. L., Lima, K. R. d. S., Ferraz, L. G. d. C., Batista, F. A., Barros, F. C. P. d., Chaves, L. C., Souza, C. d., and Roque, L. F. N. d. M. (2022). Use of testing dojo as a methodology of collaborative learning in teaching testcase writing: An experience report. In *Proceedings of the 13th International Conference on Society and Information Technologies (ICSIT 2022)*, pages 41–44.
29. Costa, I. E. F. and Oliveira, S. R. B. (2022). Development of a teaching plan to support learning activities of exploratory test design and execution. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.
30. Elgrably, I. S. and Oliveira, S. R. B. (2022). Perception from the professors' point of view in the remote teaching of software testing using active methodologies during the covid 19 pandemic. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9.
31. Javadi, E., Gebauer, J., and Tanner, S. (2022). Flipped classrooms & project dojos for enhancing peer-learning in classrooms. *AMCIS 2022 TREOs*, 57(1):1.
32. Meireles, M., de Souza, C., de Barros, F. C., Chaves, L., de Castro, R., and Giuntini, F. (2022). The employment of testing dojo as a collaborative learning methodology for teaching failure analysis: An experience report. In *2022 4th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 47–54.
33. Garcia, F. W. d. S. G., Oliveira, S. R. B., and Carvalho, E. d. C. (2023). A second experimental study the application of a teaching plan for the algorithms subject in an undergraduate course in computing using active methodologies. *Informatics in Education*, 22(2):233–255.
34. Costa, I. E. F., Oliveira, S. R. B., Elgrably, I. S., Guerra, A. d. S., Soares, E. M., and Costa, I. V. F. (2023). Using active methodologies for teaching and learning of exploratory test design and execution. *Education Sciences*, 13(2).
35. Ferreira, D. J., Campos, D. S., and Gonçalves, A. C. (2024). A framework of contextualized social regulation strategies in introductory programming. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, pages 5124–5133.

Appendix B Means of IMI items

This appendix presents the means of the answers obtained for each IMI item. In **Figure 16** the IMI items are organized by subscale (INT, CMP, EFF, VAL, and PRS) and the means are presented by experiment execution and group (experimental and control). Reverse items are indicated by a "(R)" at the end of their descriptions and their values were already adjusted/reversed.

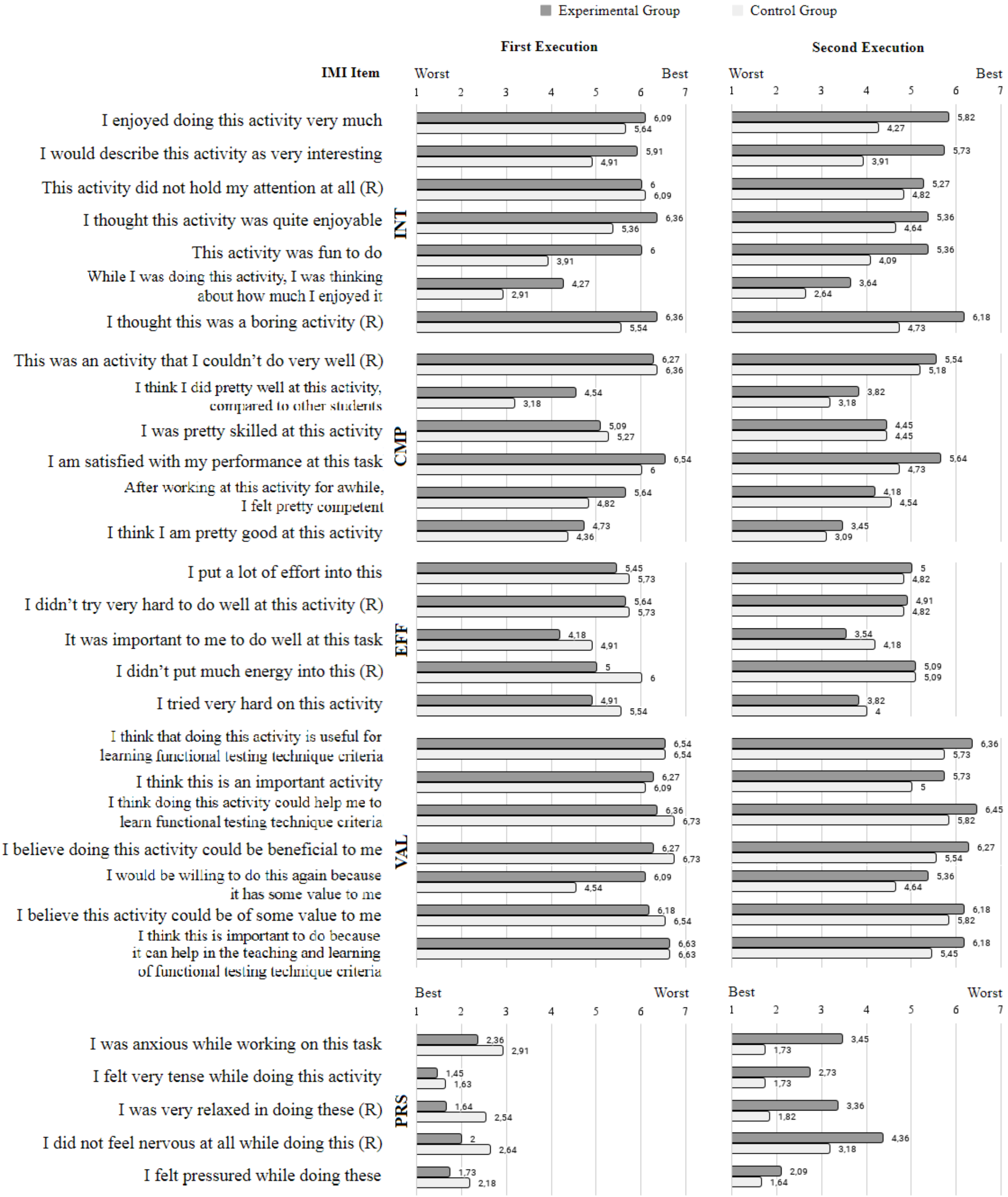


Figure 16. Means obtained for each IMI item

Appendix C Visual Diagnostics for Normality Assumptions

This appendix presents some visual diagnostics that can be used along with the Shapiro-Wilk test to assess the normality of the experimental results. To give readers a visual reference for the normality assumptions, **Figure 17** exposes sample histograms and Q-Q plots for the results of the INT subscale. For the other variables, analogous graphs can be plotted using the data available in Subsections 7.2 and 7.3.

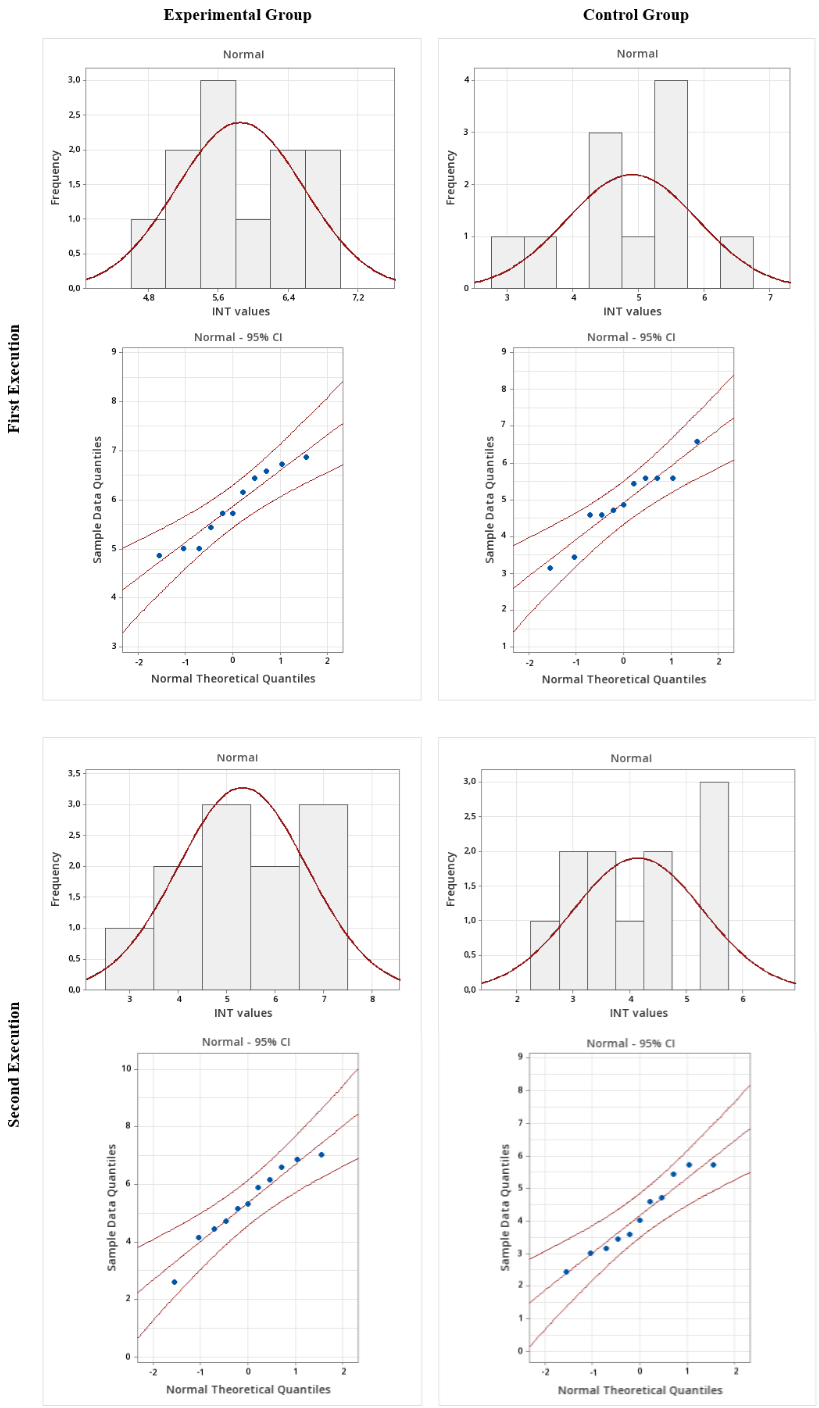


Figure 17. Sample histograms and Q-Q plots for INT results