Heliyon 6 (2020) e03917

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A User Interface (UI) and User eXperience (UX) evaluation framework for cyberlearning environments in computer science and software engineering education

Hakam W. Alomari [a,*], Vijayalakshmi Ramasamy [a], James D. Kiper [a], Geoff Potvin [b]

[a] Department of Computer Science and Software Engineering, Miami University, Oxford, OH, 45056, USA
[b] Department of Physics, Florida International University, Miami, FL, 33199, USA

ARTICLE INFO

Keywords:
Computer science
STEM education
User interface
User experience
Cyberlearning environment
Usability evaluation

ABSTRACT

Despite the widespread availability and increasing use of cyberlearning environments, there remains a need for more research about their usefulness in undergraduate education, particularly in STEM education. The process of evaluating the usefulness of a cyberlearning environment is an essential measure of its success and is useful in assisting the design process and ensuring user satisfaction. Unfortunately, there are relatively few empirical studies that provide a comprehensive test of the usefulness of cyberlearning in education. Additionally, there is a lack of standards upon whose usefulness evaluators agree.

In this research, we present multiple user studies that can be used to assess the usefulness of a cyberlearning environment used in Computer Science and Software Engineering courses through testing its usability and measuring its utility using user interface and user experience evaluations. Based on these assessments, we propose an evaluation framework to evaluate cyberlearning environments. To help illustrate the framework utility and usability evaluations, we explain them through an example SEP-CyLE (Software Engineering and Programming Cyberlearning Environment). The evaluation techniques used are cognitive walkthroughs with a think-aloud protocol and a heuristic evaluation survey. We further use a network-based analysis to find the statistically significant correlated responses in the heuristic evaluation survey with regard to the students' perceptions of using SEP-CyLE.

Our goal is to improve cyberlearning practice and to emphasize the need for evaluating cyberlearning environments with respect to its designated tasks and its users using UI/UX evaluations. Our experiments demonstrated participants were able to utilize SEP-CyLE efficiently to accomplish the tasks we posed to them and to enhance their software development concepts, specifically, software testing. We discovered areas of improvement in the visibility and navigation of SEP-CyLE's current design. We provide our recommendations for improving SEP-CyLE and provide guidance and possible directions for future research on designing cyberlearning environments for computer education.

1. Introduction

As defined by the National Science Foundation (NSF)[1], cyberlearning is: *"the use of networked computing and communications technologies to support learning"* [1]. Based on the current knowledge about how people learn, cyberlearning research can be defined as the study of how new technologies can be used to advance learning and facilitate learning experiences in ways that were never possible before. Of course, it is impossible to study cyberlearning without the use of technology itself [2]. The best way researchers have found to investigate potential advances is to design learning experiences and study them [3, 4, 5, 6, 7, 8]. This is our motivation in our work with SEP-CyLE.

The cyberlearning community report, the state of cyberlearning and the future of learning with technology [2], concludes that the major differences today from earlier research in the cyberlearning field is the usability, availability, and scalability of technologies used in cyberlearning. This report was organized by CIRCL (The Center for Innovative Research in Cyberlearning) and co-authored by 22 members of the U.S.

* Corresponding author.
E-mail address: alomarhw@miamioh.edu (H.W. Alomari).
[1] https://www.nsf.gov.

https://doi.org/10.1016/j.heliyon.2020.e03917
Received 11 November 2019; Received in revised form 24 January 2020; Accepted 30 April 2020

cyberlearning community, including both computer and learning scientists.

The ISO 9241 report [9] defines usability as: *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"*, where effectiveness is the accuracy and completeness with which users achieve specified goals; efficiency measures resources expended in relation to the accuracy and completeness of goals achieved; and satisfaction is the comfort and acceptability of the work system to its users and other people affected by its use [10]. While usability is essential to the success of any product design, the design's utility is also a major consideration in evaluating its quality. Usability and utility are closely related, however they are not identical. Based on Nielsen [11], utility is concerned with usefulness whereas usability includes not only utility, but also effectiveness, efficiency, and satisfaction.

To this end, this article focuses on usability and utility evaluations for cyberlearning environments used in computer science education. The selected cyberlearning environment for this research is a frequently used environment, called SEP-CyLE [12], currently used in several institutions in the USA by several researchers and learners [3, 4, 5, 7, 8, 13, 14, 15]. SEP-CyLE, Software Engineering and Programming Cyberlearning Environment, is an external web-based learning tool that helps instructors integrate software development concepts into their programming and software engineering courses. SEP-CyLE is used in this research since it provides a variety of collaborative learning settings containing learning materials of both programming and software testing concepts and tools. SEP-CyLE includes learning objects and tutorials on a variety of computer science and software engineering topics. In this experiment, we chose to use learning objects related to software testing. Our research aim is to investigate the utility of SEP-CyLE and evaluate the usability of its user interface based on the actual user experience.

1.1. Contributions

This article extends the previous work on integrating software testing concepts and tools into CS and SE courses to improve software testing knowledge of CS/SE students [4, 6, 7, 8, 16, 17]. This article builds upon that work by making the following contributions:

1. A new UI/UX evaluation framework that is considered more appropriate to evaluate cyberlearning designs. The framework extends the current practice by focusing not only on cognitive aspects but also usefulness considerations that may influence cyberlearning usability. The framework uses four user studies; pre/posttests, heuristic evaluation, cognitive walkthrough, and UX survey.
2. Using network analysis to measure the design accuracy of cyberlearning system from the usability and utility perspectives, and to identify students' interaction patterns and designs into STEM courses.
3. Usability and utility recommendations and comments for future design of cyberlearning educational environments in CS and SE courses.

1.2. Problem definition

The number of users (students and instructors) of cyberlearning environments continues to grow [2, 18]. For example, the substantial growth in using SEP-CyLE in recent years is an evidence of the increasing importance of employing cyberlearning tools in the learning process. Consider that, as of summer 2017, the NSF has made approximately 280 cyberlearning research grant awards [2, 18, 19]. This is apparently another evidence of the emergent need for integrating technological advances that allow more personalized learning experiences among those not served well by the current educational practices. However, the usability evaluation of cyberlearning environments and their effectiveness are still an open research questions [20].

The process of evaluating the usability and the utility, for any given design, is not an easy task [21, 22]. Usefulness and effectiveness present significant challenges in evaluating the cyberlearning environment. Factors such as the significant increase in the cyberlearning technologies, identifying the cyberlearning users, and the context of the task provided by the cyberlearning environment impose additional difficulties. As Nielsen observed [21], if a given system cannot fulfill the user's needs (i.e., usefulness), then it is not important that the system is easy to use (i.e., effectiveness). Similarly, if the user interface is too difficult to use, then it is not important if the system can do what the user requires since the user cannot make it happen.

While there are many methods for the user to inspect a design's usability, one of the most valuable methods is to test using different usability tools such as the think-aloud protocol [23]. The commonality among all these methods that are based on having evaluators inspect a UI in the goal of finding usability problems in the system design. As mentioned by Nielsen, thinking aloud may be one of the most valuable usability engineering methods [24]. Nielsen also suggested that to study a system's utility, it is possible to use the same user research methods that improve usability [11].

Our goal in UI/UX evaluations of the SEP-CyLE tool is to help decide if SEP-CyLE fulfills the requirements of a well-designed cyberlearning environment for fundamental programming and software engineering courses by covering all important features of usability and utility. To accomplish this, we begin by understanding the key elements of a successful cyberlearning design, then building an evaluation framework that uses these elements to better understand the purpose of learning and how users learn. We hope that this understanding will be helpful in the future design of new technologies (cyberlearning-related) for these purposes and in their integration into cyberlearning environments to make computer education more meaningful and effective for a broad audience.

1.3. Article organization

The rest of the article is organized as follows. Section 2 introducing some background information on cyberlearning environments and SEP-CyLE. Section 3 describes the proposed cyberlearning evaluation framework. Section 4 discusses the evaluation. Section 5 presents the evaluation results. Section 6 discusses our takeaways and highlights from the data. We identify threats to validity in Section 7, and conclude and outline future work in Section 8.

2. Background and related work

This section provides the necessary background information on cyberlearning and SEP-CyLE cyberlearning environment. A complete details of the SEP-CyLE design and usage are described elsewhere [3], here we only provide an overview of the tool, the current design, and its main features that are available for its two types of users, i.e., instructors and students.

2.1. Cyberlearning

The cyberlearning definition produced in 2008 by NSF [1], as mentioned above in Section 1, focused on technologies that can be networked, to support communications between users. Before that, in 2005, Zia [25] during a presentation on game-based learning at the National Academy of Sciences[2] defined cyberlearning as "Education + Cyber-infrastructure". In 2013, Montfort [26] defined cyberlearning as any form of learning that is facilitated by the use of technology in such a way that changes the learner's access to and interaction with information. Other works considered cyberlearning as a modern twist on e-learning [27, 28]. While e-learning focuses on the transmission of information via

[2] http://www.nasonline.org.

a digital platform, cyberlearning uses a digital platform to establish a comprehensive, encompassing, technology-based learning experience, where students derive their understanding, and thus learning [29].

Essentially, the primary goal of both cyberlearning and e-learning is to provide learning experiences via a technology-based platform. The major difference lies in how such learning experiences are provided. For example, cyberlearning can help learners in their learning activities in a way that is effective and efficient using advanced electronic technologies. Collaborative tools, gamification, and virtual environments all are examples of cyberlearning technologies that are transforming education. These technologies can be used effectively by delivering appropriate learning contents and services that fulfill user needs in a usable manner [20]. Thus, cyberlearning extends e-learning by providing effective learning initiatives. It is apparent that cyberlearning has alternative definitions in the literature, and each definition emphasizes separate aspects.

Usability is a necessary condition for online learning success. According to several experts in the field [21, 30], *"usability is often the most neglected aspect of websites, yet in many respects it is the most important"*. If the designed cyberlearning environment is difficult to use and fails to state what the environment offers and what users can do, then users simply leave. That is, users are not going to spend much time trying to figure out an interface. In order to improve usability, the most basic and useful method is to conduct usability testing that includes three major components: representative users, representative tasks, and user's observation [11].

Usability assessments in the literature have been conducted using different methods and for different purposes. Nevertheless, there is no universal or standard technique or method upon which usability evaluators agree [20]. However, various evaluation techniques are used to support different purposes or types. For example, in other work, the evaluators used analytical evaluation techniques (i.e., heuristic evaluation [21], cognitive walkthrough [31], keystroke-level analysis [32]) to determine usability problems during the design process. And they have used empirical evaluation techniques (i.e., formal usability testing and questionnaires) to determine actual measures of efficiency, effectiveness and user satisfaction [30].

2.2. SEP-CyLE cyberlearning environment

SEP-CyLE was developed by Clarke et al. [3] at Florida International University as a cyberlearning environment called Web-Based Repository of Testing Tutorials (WReSTT) [6, 16]. Its major goal was to improve the testing knowledge of CS/SE students by providing a set of learning objects (LOs) and tutorials to satisfy the learning objectives. These LOs and the corresponding tutorials are categorized sequentially based upon the difficulty level. Subsequently, WReSTT has evolved into a collaborative learning environment (now called SEP-CyLE) that includes social networking features such as the ability to award virtual points for student social interaction about testing. These are called *virtual points* since an instructor may choose not to use these as a part of students' grades, but only for the motivation that collaborative learning and gamification may provide.

SEP-CyLE, current version of WReSTT used in this article, is a configurable learning and engagement cyberlearning environment that contains a growing amount of digital learning content in many STEM areas. Currently, the learning content primarily focuses on software engineering and programming courses. Unlike other e-learning management systems, SEP-CyLE uses *embedded learning and engagement strategies* (ELESs) to involve students more deeply in the learning process. These ELESs are considered to be cyberlearning technologies and currently include collaborative learning, gamification, problem-based learning and social interaction. The learning content in SEP-CyLE is packaged in the form of LOs and Tutorials. LOs are chunks of multimedia learning content that should take the learners between five to fifteen minutes to complete and contain both a practice (formative) assessment and a second summative assessment. The collaborative learning features allow students to upload their user profile, gain virtual points after completing a learning object, post comments on the discussion board, and monitor the activities of peers.

The choice of learning objects used in a given course is based on the level of the course, the course learning outcomes, and instructor preferences. A variety of learning objects and tutorials are available in SEP-CyLE. For example, SEP-CyLE can be used by instructors in both undergraduate and graduate courses by assigning students the learning contents with appropriate levels of difficulties. Note that all the students involved in the experiment described in this article were undergraduate students.

2.2.1. Student and instructor views

SEP-CyLE provides two views: one for instructors and one for students, as shown in Figure 1. As we can see in Figure 1 (a), SEP-CyLE allows the capability for an instructor to create a course module by enrolling students into the course and providing students with access credentials for using SEP-CyLE. By creating a course and using the course management page, as shown in Figure 1 (c), the instructor can 1) upload the class roster; 2) create unique login credential for the students; 3) assign students to virtual teams; 4) describe the rubric for the allocation of virtual points for different student activities; 5) create student reports; and 6) browse and assign several learning objects and testing tool tutorials.

As shown in Figure 1 (b), students can create a user profile by uploading a profile picture (and gain some virtual points), browse the testing tutorials, complete assigned learning objects by passing with at least 80% on assigned quizzes (and gain virtual points), watch tutorial videos on the different testing tools (e.g., JUnit, JDepend, EMMA, CPPU, Cobertura), interact with other students in the class via testing based discussions (and gain virtual points for relevant discussions), and monitor the activity stream for whole class. These features are illustrated in Figure 1 (d).

2.2.2. SEP-CyLE UX survey

The SEP-CyLE survey we have used hade a total of 25 questions. These questions aimed at receiving students' reflections, and are divided into three categories. Table 1 shows 14 questions that are related to the *overall reaction to SEP-CyLE*, Table 2 shows 6 questions related to the *testing concepts*, and finally, Table 3 shows 5 questions that are related to the *collaborative learning*.

3. Proposed cyberlearning evaluation framework

This section presents our work in evaluating SEP-CyLE's usability and validates its effectiveness. We present our proposed framework to evaluate cyberlearning environments in general and SEP-CyLE environment, specifically.

We conduct our evaluation through four user studies using a pre/posttest instruments, a heuristic evaluation, cognitive walkthroughs with a think-aloud protocol, and SEP-CyLE's UX survey. We preferred not to use the heuristic method of evaluation alone, since it sometimes found usability problems without providing suggestions for how to solve them [21]. The heuristic evaluation is used as a starting point to troubleshoot, and the thinking-aloud allows us to identify the potential SEP-CyLE improvements and gain insights into how users interact with the cyberlearning environments, thus benefiting the field as a whole.

3.1. Research objectives and questions

The objective of our evaluation is twofold. First, we want to measure SEP-CyLE's utility by measuring whether it satisfies the user needs by studying how SEP-CyLE's supporting materials affect the software testing knowledge acquisition by students. The second objective is to measure

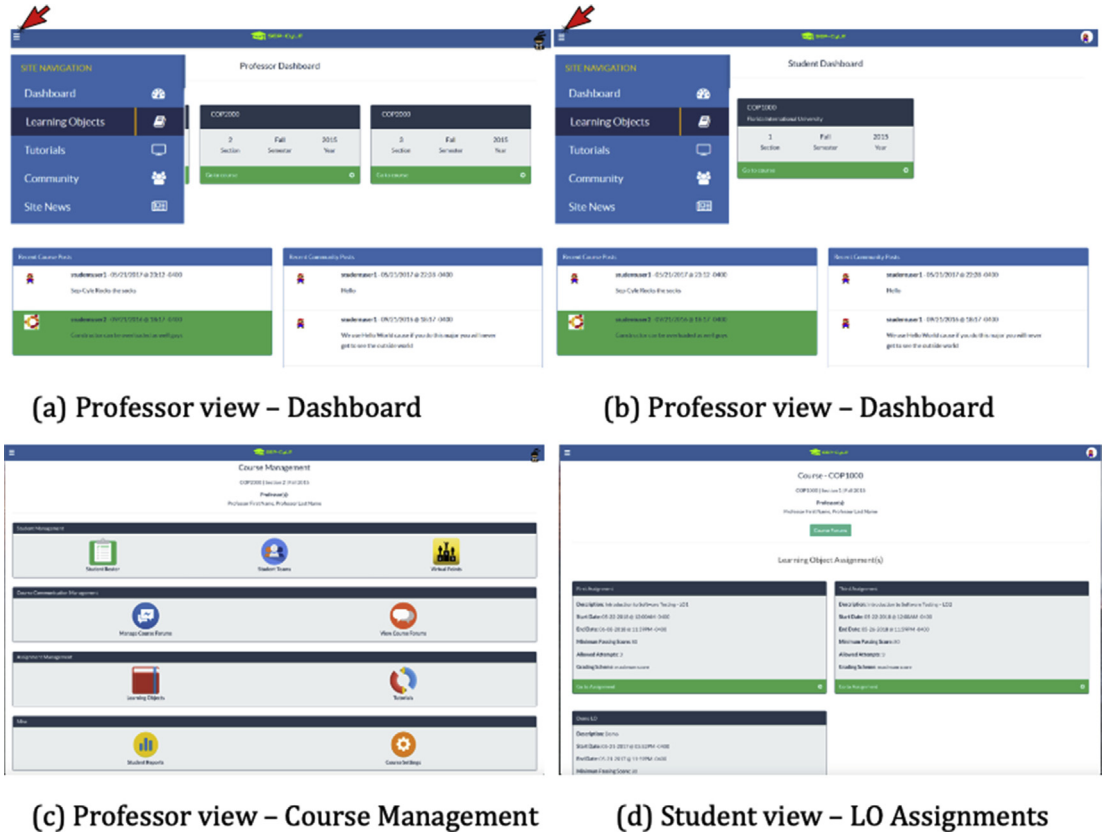


Figure 1. SEP-CyLE's Professor view and Student view.

Table 1. Overall reaction to SEP-CyLE.

1. Overall, I am satisfied with how easy it is to use the Web site.
2. It is simple to use the Web site.
3. I feel comfortable using the Web site.
4. It was easy to learn to use the Web site.
5. I believe I became productive quickly using the Web site.
6. The information provided with the Web site is clear.
7. It is easy to find the information I need.
8. The information is effective in helping me complete tasks and scenarios.
9. The interface of the Web site is pleasant.
10. I like using the interface of this Web site.
11. The Web site has all the functions and capabilities I expect it to have.
12. I believe that the Web site helped me earn a better grade.
13. I would recommend the Web site to fellow students.
14. Overall, I am satisfied with the Web site.

Table 2. Testing related questions.

1. Tutorials helped me to better understand testing concepts.
2. Tutorials helped me to better understand how to use unit testing tools.
3. Tutorials helped me to better understand how to use code coverage tools.
4. Tutorials helped me to better understand how to use functional testing.
5. The number of tutorials in SEP-CyLE is adequate.
6. I would have used testing tools in my project if SEP-CyLE did not exist.

SEP-CyLE's usability and its ease of use. Together, these objectives lead to two primary research questions that this study addresses:

Table 3. Collaborative learning-related questions.

1. The use of virtual points encouraged me to visit the site and complete the tasks.
2. The use of virtual points encouraged my team to visit the site & complete tasks.
3. The event stream encouraged me to complete my tasks in SEP-CyLE.
4. The event stream encouraged my team to complete my tasks in SEP-CyLE.
5. Our team devised a plan to get the max number of points in SEP-CyLE.

- ***RQ1****: Does SEP-CyLE meet utility requirements?* This question will help us measure the impact of using SEP-CyLE on the software testing knowledge gained by the students.
- ***RQ2****: Does SEP-CyLE meet usability requirements?* This question will help us measure the SEP-CyLE's ease of use and engaging features.

As shown in our proposed evaluation framework in Figure 2, a pre/posttest instrument, students' final scores, and SEP-CyLE UX survey were used to address the first research question (RQ1). We asked users to perform a heuristic evaluation and a think-aloud protocol to address the second research question (RQ2). Finally, we applied graph-theoretic analysis over the heuristic evaluation questions to study the insights into the students' perceptions of these questions.

3.2. Evaluation framework design

Assessing the UI/UX design quality of a cyberlearning environment first requires understanding the key elements of any cyberlearning design. There is some variability in the definition of cyberlearning. This disagreement reflects the underlying differences in understanding the purpose of learning and how people learn. Consequently, evaluating cyberlearning tools is still problematic for researchers [26]. We rely on

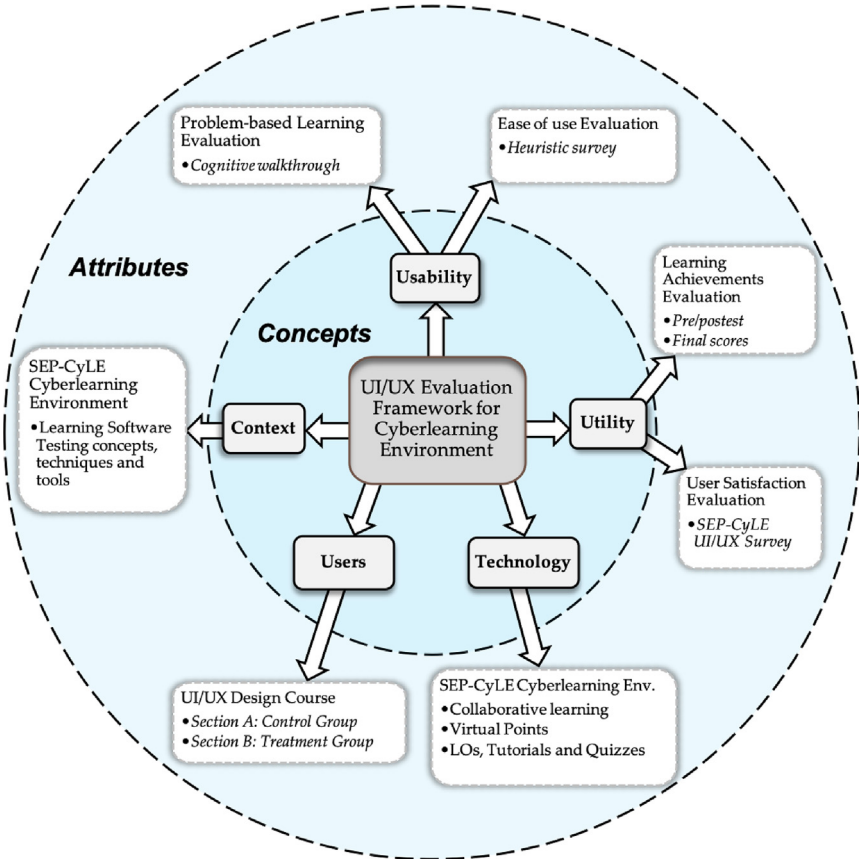


Figure 2. Concepts and attributes of our UI/UX evaluation framework of a cyberlearning enviroment - SEP-CyLE is used as a case study.

the following definitions of cyberlearning, usability, and utility. Cyberlearning: *"the use of networked computing and communications technologies to* support *learning"* [1]. Usability: *"the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments"*. And Utility of an object: *"how practical and useful it is"* [11]. We drew on the literature from several fields to develop a framework for evaluating the UI/UX of SEP-CyLE cyberlearning environment, and cyberlearning in general. Our framework identifies key concepts of cyberlearning and the important attributes associated with each concept. As shown in Figure 2, Technologies, for example, are considered one of the five important concepts in designing and evaluating a cyberlearning environment. Therefore, we chose those technologies implemented in SEP-CyLE to conduct our evaluation. The evaluation model has five primary concepts, as follows:

- Context: to evaluate the utility of a cyberlearning environment, the context in which it is used is important. For example, SEP-CyLE was designed to support the students' learning in the software development process, specifically, software testing concepts, techniques and tools.
- Users: a cyberlearning environment is valuable to the extent that there are users that value the cyberlearning technologies and services, thus providing the purpose for it to exist. In SEP-CyLE, students and instructors are the two user classes. In our work, both types of users are involved in the evaluation process; however students are the major source of data collection for evaluation.
- Technology: in our evaluation process, we try to understand how users learn with technology, and how technology can facilitate this learning. The data used for evaluation is collected in a way that is related to the specified technology used in the cyberlearning environment, such as collaborative learning and gamification. The collaborative learning component in SEP-CyLE is based mainly on students' involvement, cooperation, and teamwork factors as defined by Smith et al. [33].

SEP-CyLE achieved those factors by rewarding students with virtual points, requiring collaborative participation between team members, and providing social engagement opportunities. For more details about the collaborative learning strategies provided by SEP-CyLE, please refer to [5].

The other cyberlearning technology used in SEP-CyLE and measured by our evaluation is gamification. The game mechanics implemented and used in SEP-CyLE [5], are participant points, participation levels, and leader boards. These mechanics were adapted from the work established by Li et al. [34].

- Utility: in order to evaluate the usefulness of the used technology, we analyzed the collected data that is related to the technology itself. The analysis process considers the user satisfaction and the user needs of using that technology to understand the specified context. For example, we used the SEP-CyLE UX survey to measure the user satisfaction in the collaborative learning component and its effectiveness in the learning process. Another measure of the usefulness of SEP-CyLE in the learning achievements is the mapping of pre/posttest score to the corresponding student's final course grade.
- Usability: we measure how easy a particular technology is to use by specific class of users in a specified context to advance learning. For example, we used heuristic evaluation and cognitive walkthoughs to measure the subjective and the objective values of collaborative learning and problem-based (task-based) learning, respectively.

These concepts point to the activities and the processes in the evaluation model as shown in Figure 2. The concepts are interrelated, and we

Table 4. Participating subjects.

Class	Enrollment	Pre/Posttest		Cognitive Walkthrough		Heuristic & SEP-CyLE Surveys	
		Part.[a]	%[b]	Part.[a]	%[b]	Part.[a]	%[b]
212A*	37	27	73.0	20	54.1	–	–
212B**	32	24	75.0	–	–	25	78.1

- 51 participants in the pre/posttest study.
- 45 participants in the Cognitive walkthrough, heuristic evaluation, and SEP-CyLE's UX survey.

* Section A is considered as a control group from the pre/posttest perspective.
** Section B is considered as a treatment group from the pre/posttest perspective.
[a] Part. = Participation.
[b] % = Percentage.

believe that for a cyberlearning environment to be successful it should address all of them.

3.3. Subjects and courses

Within the context of the interactive educational environment, it is essential that feedback is elicited from real users [35]. Additionally, when the analytical evaluations are used, it is recommended that the evaluation is performed by heuristic evaluation and cognitive walkthrough experts [21], since this type of evaluation requires knowledge and experience to apply effectively. In our study, we chose students as the subject of our studies since the SEP-CyLE actual users are students. Although we do not have access to heuristic evaluation and cognitive walkthrough professionals, to carry out these assessments, we have used students in a UI/UX course who have been trained in these techniques as part of the course content.

As shown in Table 4, the course used for subjects and assessors in this study is a 200-level (sophomore-level) software engineering undergraduate course for UI/UX design. Students from two sections of this course in Fall 2017 were used to conduct these assessments. In order to reduce bias, the participants in the first section, Section A, were asked to perform cognitive walkthroughs, and the participants in the second section, Section B, were asked to complete both heuristic survey and SEP-CyLE UX survey. Students in Section B studied the usability principles during the UI/UX course before they completed the heuristic evaluation and the UX survey at the end of this course. They reviewed the SEP-CyLE website's interfaces and compared them against accepted usability principles that they had learned from the course content.

In addition to all these data, the pre/postest instrument was used to assess the SEP-CyLE's utility. That is, the software testing knowledge gained by the students after their exposure to SEP-CyLE. The pretest was administered to the students in both sections at week 1, and the posttest was administered for both sections at the conclusion of the same semester, week 14. Students in both sections were exposed to SEP-CyLE; however the only difference is that students in Section A were not explicitly assigned or instructed in learning any software testing concepts or assignments using SEP-CyLE.

The number of participants were chosen based on the recommendations provided in the literature, and summarized in Nielsen's work in [21, 36]. Nielsen outlines the number of participants needed for the study based on a number of case studies, as follows:

- The usability test using cognitive walkthroughs: you need at least 5 users to find almost as many usability problems as you'd find using more user participants.
- The heuristic evaluation: you need at least 20 users to get statistically significant numbers in your final results.

As shown in Table 4, a total of 51 students participated in the pre/posttest study. A 20 students participated in the cognitive walkthrough evaluation from Section A, and 25 students participated in the heuristic evaluation and the SEP-CyLE UX survey from Section B. Again, there were no substantive differences in terms of demographics and course preparation between the subjects in both sections. While both sections were exposed to the same lecture and other course contents and SEP-CyLE, the subjects in Section B (treatment group) were instructed explicitly to the software testing learning objects provided by SEP-CyLE. On the other hand, students in Section A (control group) just used SEP-CyLE as a cyberlearning environment to conduct their cognitive walkthrough usability inspection method without any explicit instructions to study the software testing LOs, tutorials, quizzes provided by SEP-CyLE.

4. Experimental design

The following subsections describe the experimental design aspects of evaluating the SEP-CyLE's utility and usability and ease of use for the proposed cyberlearning evaluation framework.

4.1. Utility evaluation

4.1.1. Pre/posttest instrument design

The pre/posttest instrument that we used and the SEP-CyLE survey are both available for download at https://stem-cyle.cis.fiu.edu/under the *publications* tab. The pre/posttest was designed to identify students' knowledge of software testing prior to being exposed to the learning objects in SEP-CyLE and after being exposed to the learning objects in SEP-CyLE (just in Section B). The eight questions in the pre/posttest focused on the topics listed below:

- Q1 — The objective of software testing.
- Q2 — Identification of different testing techniques.
- Q3 and Q4 — Use of testing tools related to unit testing, functional testing, and code coverage.
- Q5 and Q6 — Familiarity with other online testing resources.
- Q7 and Q8 — Importance of using testing tools in programming assignments.

4.1.2. SEP-CyLE UX survey design

To create our SEP-CyLE survey, we adapted and modified the original survey created by Clarke et al. [5] [3]. Clarke used this survey to evaluate the previous two versions of the SEP-CyLE: WReSTT v1 and WReSTT v2. The WReSTT survey consists originally of 30 questions divided into five groups, as follows:

- Group 1 (1 question) — focused on the use of testing resources other than WReSTT.
- Group 2 (14 questions) — focused on the overall reaction to the WReSTT website. These questions were created first by Albert et al. [37] in 2013, then adapted by Clarke et al. [5] to compare the overall reaction for both implementations of WReSTT (the initial version of

[3] Available for download at https://stem-cyle.cis.fiu.edu/.

SEP-CyLE). Here, we adapted the same questions, then modified these questions to evaluate SEP-CyLE.
- Group 3 (6 questions) — evaluate the impact of using WReSTT on learning software testing concepts and tools.
- Group 4 (5 questions) — focused on evaluate the impact of using collaborative learning components in WReSTT. We modified these questions to evaluate SEP-CyLE's collaborative learning component.
- Group 5 (4 questions) — open-ended questions to provide feedback on the different versions of WReSTT.

In our proposed SEP-CyLE survey, we just used Groups 2, 3, and 4, with a total of 25 questions.

4.2. Usability evaluation

4.2.1. Cognitive walkthrough design

Throughout the evaluation, participants were asked to verbalize their responses to all questions. Before beginning the cognitive walkthrough, we asked the participants to develop a strategy that they would expect to use in order to achieve the given goal and verbally explain the strategy they would follow to the evaluator. The evaluator then asked three questions before the participants took any action on each screen. The three questions were:

1. What is your immediate goal?
2. What are you looking for in order to accomplish that goal?
3. What action are you going to take?

Participants could respond to the first question based on the task they were assigned, but are more likely to respond with an intermediate goal. This could include responses such as “find the learning object" or “find a software testing tutorial". When participants respond to the second question, they would respond with the type of control they are looking for, such as a button or menu selection. The third question would be answered when the participants have decided what they intend to interact with, such as clicking a button or typing in input.

After these questions were answered, the participant completed the action and the evaluator asked the following two questions:

1. What happened on the screen?
2. Is your goal complete?

Participants responded to the first question based on what they perceived on the screen. Their response may be along the lines of “nothing", in such cases when they attempt an invalid action. However, a typical response is more likely to be “the screen is changed" or “an error message appears". Participants often answer the second question with a “yes" or “no" response. These steps were repeated until participants completed the primary tasks we provided them for evaluation. The evaluator noted observations of any usability problems during the cognitive walkthrough process. Additionally, we used video recordings to capture both audio and on-screen actions to facilitate a more detailed analysis and comparisons after completion of each cognitive walkthrough.

We gave the users the following software testing related tasks to be completed during the cognitive walkthroughs:

1. Identify names of the learning object (LO) assignments available
2. Navigate through one LO assignment
3. Identify the name of that assignment
4. Identify the number of pages provided in the content of this assignment
5. Take the quiz assigned by this learning assignment
6. Identify the number of questions provided in this quiz
7. Identify the final score you have after completing this quiz
8. Identify the number of tutorials provided by this LO
9. Navigate through one tutorial
10. Identify the name of this tutorial
11. Identify the number of videos provided with the tutorial
12. How many external tutorial links are available

For Tasks 4, 6, 8, 11, and 12, the participants were expected to respond with a number representing a count of “How many". Tasks 1, 3, and 10 required participants to respond with assignments and tutorials names. Task 7 required users to respond with the quiz final score. Participants were not asked to provide justification for their responses by the evaluator, although participants can explain their actions as long as these explanations are unsolicited.

4.2.2. Heuristic evaluation design

In order to reduce bias, we asked a group of participants different from those who completed the cognitive walkthrough to complete this evaluation. Our heuristic evaluation survey that was distributed digitally asked users to answer a series of standard questions regarding usability of SEP-CyLE using a Likert scale [38, 39]. The usability questions focused on three primary areas: visibility, affordability, and feedback. However, the survey also included secondary questions to gain more insights into the aspects of navigation, language, errors, and user support. The questions we used were from Nielsen's work [21], which we break into categories as listed below:

Visibility of system status

1. It is always clear what is happening on the site?
2. Whenever something is loading, a progress bar or indicator is visible?
3. It is easy to identify what the controls are used for?

Match between system and the real world

4. The system uses plain English?
5. The website follows a logical order?
6. Similar subjects/items are grouped?

User control and freedom

7. The user is able to return to the main page from every page?
8. The user is able to undo and redo any actions they may take?
9. Are there needless dialog prompts when the user is trying to leave a page?

Consistency and standards

10. The same words are used consistently for any actions the user may make?
11. The system follows usual website standards?

Error prevention

12. The system is error free?
13. There are no broken links or images on the site?
14. Errors are handled correctly, if they occur?

Flexibility and efficiency of use

15. Users may tailor their experience so they can see information relevant to them easily?

Help users recognize, diagnose, and recover from errors

16. Errors are in plain text?
17. The problem that caused an error is given to the user?
18. Suggestions for how to deal with an error are provided?

Help and documentation

19. Help documentation is provided to the user?
20. Live support is available to the user?
21. User can email for assistance?

Respondents answered these questions using a Likert scale consisting of five values (1–5) from strongly disagree (1) to strongly agree (5) [38]. This allowed for a more exploratory quantitative analysis of the data.

4.2.3. Network analysis design

Social Network Analysis, denoted by *SNA*, has emerged with a set of techniques to analyze social structures that highly depend on relational data in a broad range of applications. Many real-time data mining applications consist of a set of records/entities that solely emphasize the associations among the attributes. The graph-theoretic analysis of complex networks has gained importance to understand, model and gain more insights into the structural and functional aspects of the graph/network representation of such associations of data. It is possible to extract useful non-trivial information about the relationships between the transactional data by modeling the set of entities, their attributes and the relationships among these entities as networks. Han and Kamber [40] have studied the association rule mining. One of the most important techniques of data mining aims to extract important correlations, frequent patterns, and associations among sets of entities in the transaction databases. The measure of similarity among the transactions gives us a wealth of knowledge on identifying various communities of related transactions exhibiting common behavior in specific applications [41].

We use graph-theoretic analysis of the relational data collected to quantify the degree of relationship between pairs of variables to selectively measure the accuracy of the system under study. The 21 heuristic evaluation survey questions regarding the usability of SEP-CyLE representing the entities in the relational data is used for further analysis. We use Pearson Correlation Coefficient (r) that computes the linear relationship between two variables as a measure of the strength of correlation for graph construction [42]. Given two variables x and y, r is computed as a ratio of covariance of the variables to their standard deviations where n is the number of variables as in Eq. (1).

$$r = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

The r values calculated using x and y represent their degree of correlation in the range -1 to +1 inclusive. The results and discussion of the graph-based analyses are presented in Section 5.2.3.

5. Experimental results

The following subsections present the experimental results of evaluating the SEP-CyLE's utility and its usability.

5.1. Utility evaluation

5.1.1. Pre/posttest results

Only a few software testing topics and lectures were covered in the subject course of this study. Prior to these topics being taught in the class and at the end of the semester, participants in both sections were pretested and posttested. The pre/posttest instrument is used to assess the SEP-CyLE's utility or user satisfaction with the software testing knowledge provided. We used the same instrument in both pretest and posttest. The pre/posttest results of the control group (Section A) and the treatment group (Section B) are shown in Figure 3 and Figure 4, respectively. Please note that the participants' scores are converted into percentages.

The pre/posttest results of the treatment group are compared using the post-hoc t-test (two-tailed) and the results are shown in Table 5. The mean differences are computed with 95% confidence interval (CI) where

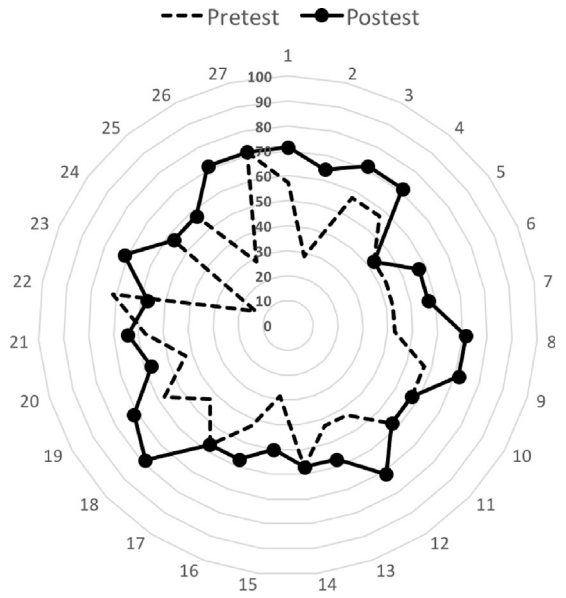


Figure 3. Pretest and Posttest scores for Control Group.

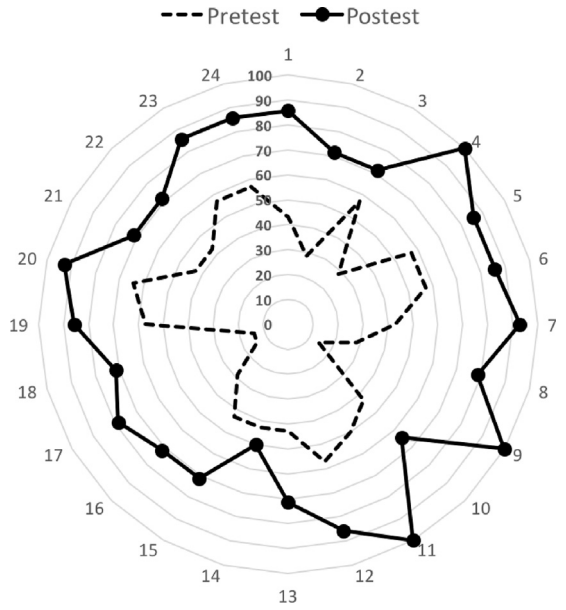


Figure 4. Pretest and Posttest scores for Treatment Group.

the null hypothesis H_0 (*that the mean values of pre/posttest scores of treatment group participants are the same*) is rejected against the alternative hypothesis H_1 (*that the mean values are significantly different between the two tests*).

5.1.2. SEP-CyLE UX survey results

Students in the treatment group were additionally asked to complete a SEP-CyLE user experience survey at the end of the semester. The same 24 students (1 students decided not to finish both surveys) were recruited again. As mentioned earlier, we modified this survey to focus on three main categories: overall reaction of SEP-CyLE, usefulness of SEP-CyLE, and usefulness of collaborative learning component.

As can be seen in Figure 5, students' overall reaction to SEP-CyLE was positive (Q1 through Q14), with a mean score equal to 4.10. However, one question (Q9) revealed some problems in using the interface. These problems could be explained using the results shown in the heuristic evaluation below.

Table 5. Statistical validation of mean differences between pre/posttest using Post-Hoc T-Test.

Number of Participants	Mean diff.	95% CI	p-value
24 (Treatment Group)	-1.2353	[-1.6423, -0.8283]*	1.82e-10

* Mean difference is significant at α* ˂ 0.05.

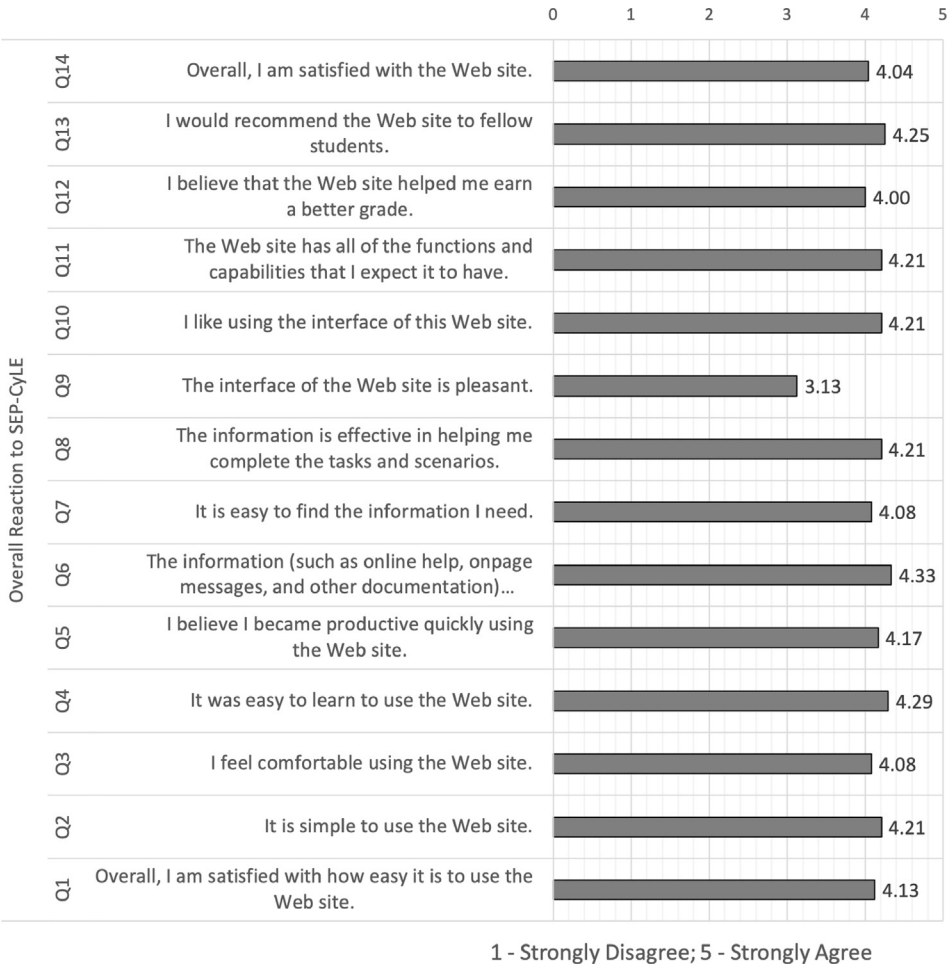


Figure 5. Students' mean scores measuring their overall reactions to SEP-CyLE.

As shown in Figure 6, students' perceptions regarding the usefulness of the testing tutorials were also positive. The mean score for this group of questions (Q15 - Q20) is 3.49. Additionally, it can be inferred that the collaborative learning component used in SEP-CyLE contributes to motivate students' learning. The mean score for these questions (Q21 - Q25) is equal to 3.83. As we can see, the virtual points and event streaming encouraged students and their team members to keep visiting the website and complete their assignments.

5.2. Usability evaluation

5.2.1. Cognitive walkthrough results

We completed a total of 20 cognitive walkthroughs with undergraduate computer science and software engineering students. Students were primarily third year students, although some second and fourth year students were included. We video-captured every cognitive walkthrough experiment. Upon review of the videos, 5 walkthroughs had unusable recordings; 2 due to technical problems and 3 due to failure to follow the protocol. These walkthroughs were excluded from our analysis. This resulted in 15 usable walkthroughs.

We evaluated each cognitive walkthrough via a rubric. The difficulty level of each task was evaluated on a scale of 0, 1, or 2. Participants received a 0 if they experienced extreme difficulty, failed to complete the task, or answered incorrectly. They received a 1 if they experienced moderate difficulty and a 2 if they finished the task correctly with no difficulty. The tasks correspond to the 12 tasks we presented earlier in Section 4.2.1.

Table 6 displays the raw scores from the cognitive walkthrough using think-aloud protocol along with the calculated means (avg) for these values once per task (t) (i.e., avg/t) and once per group (g) (i.e., avg/g). By dividing these values by the maximum possible average (which is 2 since the possible scores were 0, 1, or 2), we present the percentage of these values in Figure 7. As we can see, we broke these tasks into categories based on which portion of the system was being evaluated. The first grouping (Home) included Task 1 only. This task allows for the evaluation of navigation of the website upon initialization. The second group (LOs) includes Tasks 2 through 4 that evaluated LO use, including navigation and comprehension. The third group (Quizzes) includes Tasks 5 through 7. We used these tasks to evaluate the LO's corresponding quiz. The fourth grouping (Tutorials) includes the remaining tasks, Tasks 8

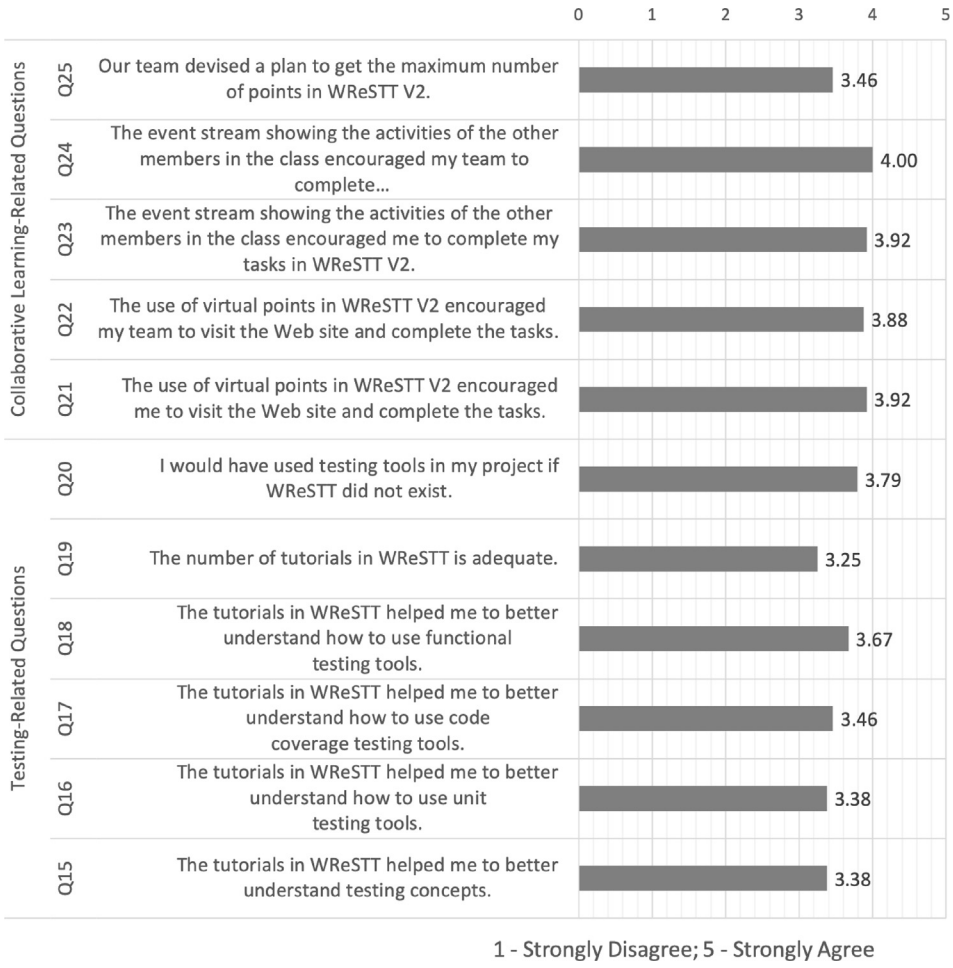


Figure 6. Students' mean scores measuring usefulness of testing tutorials and collaborative learning environment in SEP-CyLE.

Table 6. Students' Think-aloud Protocol data.

Ts*	Participants															avg_t**	avg_g**
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
T01	2	2	2	2	2	2	1	2	2	2	2	2	2	0	2	1.80	1.80
T02	2	2	0	0	2	2	1	1	2	2	2	2	2	0	2	1.47	1.51
T03	2	2	0	2	2	2	2	2	2	2	2	2	2	0	2	1.73	
T04	2	2	0	1	2	2	0	2	1	2	2	0	2	0	2	1.33	
T05	1	2	0	2	2	2	2	1	2	2	2	2	2	0	2	1.60	1.60
T06	2	2	0	2	2	2	2	2	2	2	2	2	2	0	2	1.73	
T07	2	2	0	2	2	2	2	2	2	2	2	0	2	0	0	1.47	
T08	0	1	2	0	0	2	2	1	1	2	2	2	1	2	0	1.20	1.40
T09	1	2	2	0	2	1	2	2	2	2	2	2	2	2	0	1.60	
T10	2	2	2	0	2	2	2	2	2	2	2	0	2	2	0	1.60	
T11	2	1	1	0	2	2	2	2	2	2	2	0	1	2	0	1.40	
T12	1	2	1	0	2	2	0	2	1	2	2	0	1	2	0	1.20	

* T01 - T12 = Task 01 to Task 12.
** avg_t = average per task, avg_g = average per group.

through 12. We used these tasks to evaluate the LO's corresponding tutorial(s).

From the above table and figure, it is difficult to predict with precision where the participants struggled significantly. Therefore, we calculated the two-way frequency table as shown in Table 7. In this table, the tasks are aggregated based on the frequencies of the three categories of difficulty levels (i.e., 0, 1, and 2). The entries corresponding to the frequencies of 3 levels of difficulty are called as joint frequencies and the sum of rows and columns are called as marginal frequencies. The histogram plot in Figure 8 represents three bins of categories 0, 1, and 2 (X-axis) and the frequencies computed in Table 7 (Y-axis).

The results of ANOVA (Analysis of Variance) using the data in Table 7, are shown in Table 8. The "Columns" header represents the between-group variation, and the "Error" header represents within-group

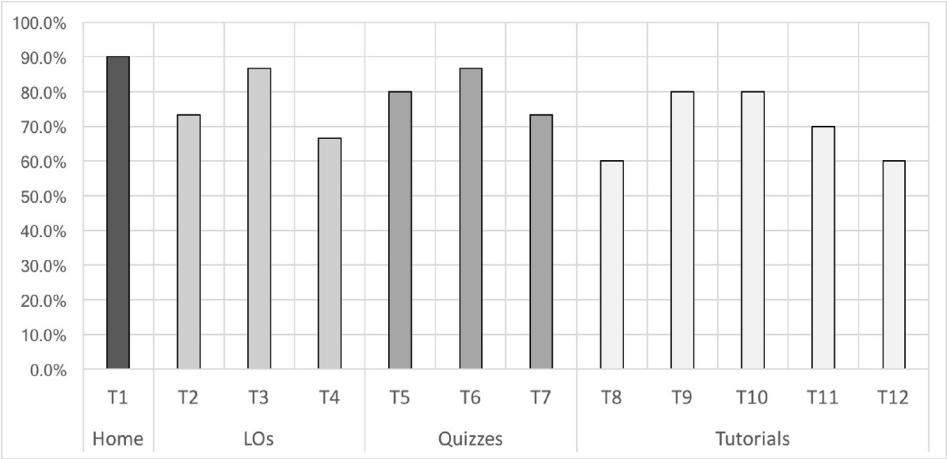


Figure 7. Rubric evaluation scores in % grouped by category name (Home, LOs, Quizzes, and Tutorials), and by task number (T1 - T12) being evaluated.

Table 7. Two-way frequency table.

Tasks	Categories			Sum
	0	1	2	
Task01	1	1	13	15
Task02	3	2	10	15
Task03	2	0	13	15
Task04	4	2	9	15
Task05	2	2	11	15
Task06	2	0	13	15
Task07	4	0	11	15
Task08	4	4	7	15
Task09	2	2	11	15
Task10	3	0	12	15
Task11	3	3	9	15
Task12	4	4	7	15
Sum	34	20	126	180

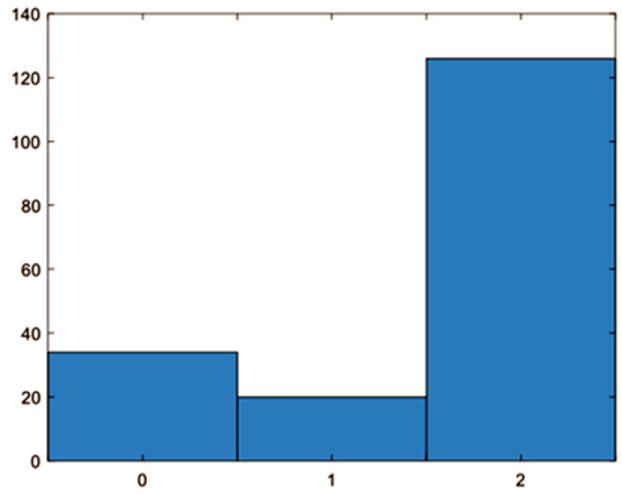


Figure 8. Histogram of difficulty levels of various categories of tasks.

variations. The total degrees of freedom **df** is 35 (total number of observations (36) minus one). The between-group degrees of freedom is 2 (number of groups (3) minus one). The within-group degrees of freedom is 33 (total degrees of freedom minus the between-group degrees of freedom). The **MS**, mean squared error is computed as **SS**/**df** for each source of variation. The **F-statistic** is the ratio of the mean squared error (293.0833/2.2374 = 130.9944) of the between-group variation to the within-group variation. The **p-value** (or probability value) is the probability that the test statistic can take a value greater than the value of the computed test statistic, i.e., P(F > 130.9944). The small p-value of 2.0123e-16 indicates that the differences between column means are significant.

The frequency values of the three levels of difficulty are compared using the post-hoc t-test (two-tailed). The t-test results are shown in Table 9. The mean differences are computed with 95% confidence interval. The null hypothesis H_0 (*that the mean values of the three groups of difficulty levels are the same*) is rejected for group (DL1 & DL2) and group (DL2 & DL0) against the alternative hypothesis H_1 (*that the mean values are significantly different between the groups*). The actual differences in means of the difficulty levels are shown in Figure 9.

The initial investigation of these categorical variables is performed to measure the association between the levels of difficulty quantitatively. One of the most common statistical methods to measure associations between categorical variables is the Chi-Square test. To setup the hypothesis for Chi-Square goodness of fit test, we assumed that the null hypothesis H_0 (*The distribution of the three categories of difficulty levels are the same while accessing the SEP-CyLE website*), and the alternative hypothesis H_1 (*There is a significant difference of the distribution of the three categories of difficulty levels while accessing the SEP-CyLE website*). The Chi-Square goodness of fit (χ^2) test for each task T is calculated as shown in Eq. (2).

Table 8. ANOVA table.

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Columns'	[586.1667]	[2]	[293.0833]	[130.9944]	[2.0123e-16]
'Error'	[73.8333]	[33]	[2.2374]	[]	[]
'Total'	[660]	[35]	[]	[]	[]

Table 9. Statistical validation of mean differences between the 3 groups of difficulty levels using Post-Hoc T-Test.

Compared Groups	Mean difference	95% Confidence Interval	p-value
DL0 & DL1	1.1667	[-0.3735, 2.7069]	0.1944
DL1 & DL2	-7.9167	[-9.4569, -6.3765]*	0.0000
DL2 & DL0	-9.0833	[-10.6235, -7.5431]*	0.0000

DLx – Difficulty level x.
* Mean difference is significant at α ˂ 0.05.

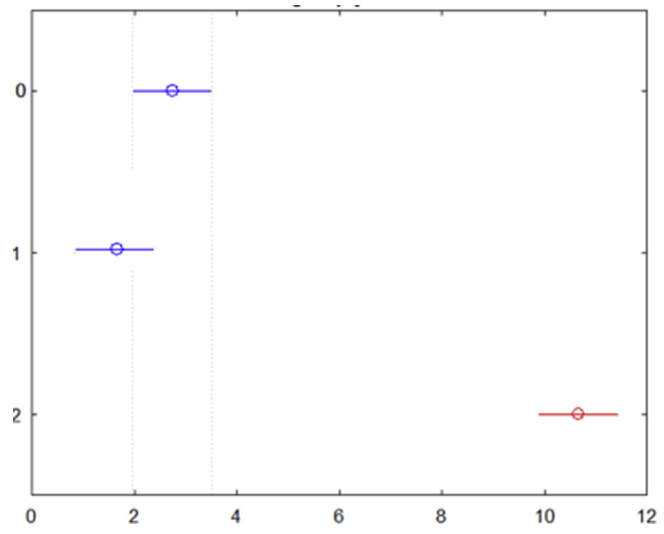


Figure 9. Significant difference between the means of (DL0 & DL1), (DL1 & DL2), and (DL0 & DL2) groups.

$$(\chi^2) = ((O - E)^2 / E) \quad (2)$$

where O is the observed frequency of each categorical value, and E is the expected frequency of a value (33.3%). The degrees of freedom is 2 and the alpha level is $\alpha = 0.05$. The p-value returned by the Chi-Square statistic represents the significance. Smaller p values mean greater significance.

As shown in Table 10, the tasks 1, 2, 3, 5, 6, 7, 9, and 10 reject the null hypothesis showing that there is a significant difference among the distribution of the three difficulty levels. The students found these tasks are relatively easy in terms of accessibility. On the other hand, the tasks 4, 8, 11 and 12 accept the null hypothesis, meaning that there is no significant differences among the distribution of three difficulty levels. These tasks were relatively difficult for the students while accessing the SEP-CyLE website.

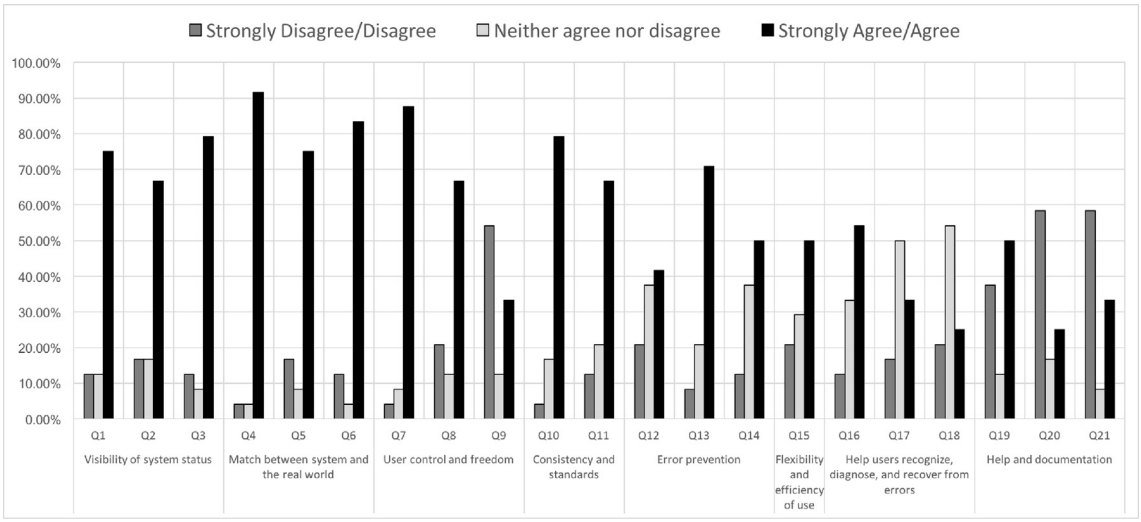
5.2.2. *Heuristic evaluation results*

25 students were recruited to participate in the heuristic evaluation study and were asked to complete the heuristic evaluation survey individually. Of the 25 participants, one (1) opted not to finish it and that response was dropped from the analysis. A Likert scale was used to represents students' responses, where strongly agree is represented by a 5, neutral is represented by a 3, and strongly disagree is represented by a 1. Thus, if a user responds with a lower value, they hold lower opinion of that particular statement.

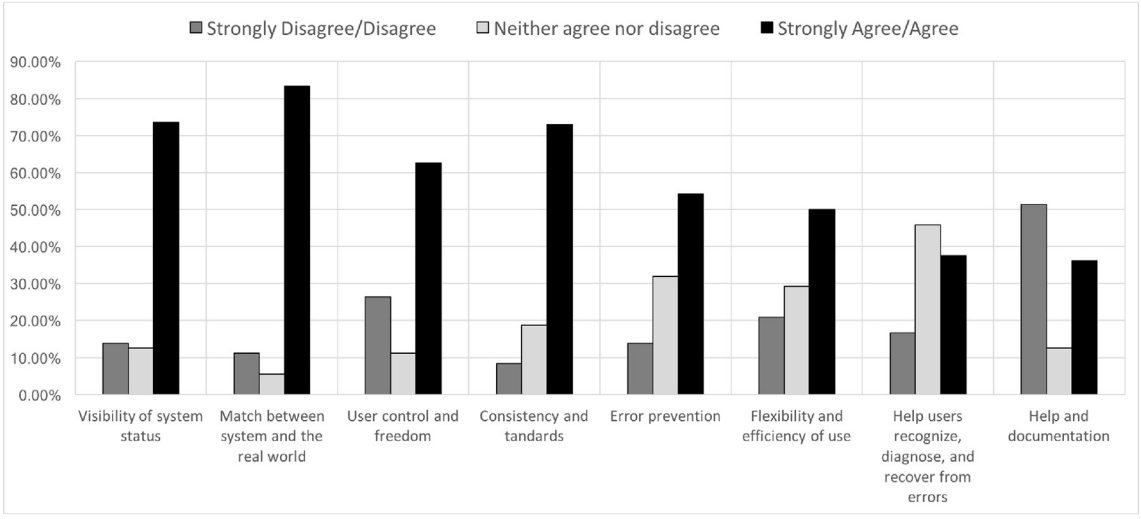
We counted participants' responses by combining strongly disagree responses with disagree responses and strongly agree responses with agree responses, then calculated the percentage of these values by dividing the summation by 24 (which is the number of participants). We present the percentage of these values per each question as in Figure 10 (a), per each grouping of question as in Figure 10 (b), and finally per each response as in Figure 10 (c).

Table 10. Chi-Square statistical results.

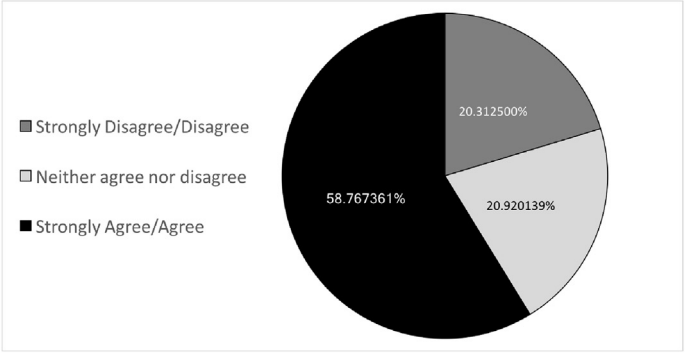
Category	Tasks	χ^2	p-value
Home	Task01	19.2	6.80E-05
LOs	Task02	7.6	0.02237
	Task03	19.6	5.50E-05
	Task04	**5.2**	**0.074274**
Quizzes	Task05	18.8	0.000611
	Task06	19.6	5.50E-05
	Task07	12.4	0.002029
Tutorials	**Task08**	**1.2**	**0.548812**
	Task09	10.8	0.004517
	Task10	15.6	0.00041
	Task11	**4.8**	**0.090718**
	Task12	**2.8**	**0.246597**
Overall		110.533	<0.00001



(a) Responses percentages per question



(b) Responses percentages per grouping of questions



(c) Responses percentages per response

Figure 10. Percentages of participants' responses.

As we can see in Figure 10 (c), participants' responses to all groups of questions are ≈ 20% strongly disagree/disagree, ≈ 21% neither agree nor disagree, and ≈ 59% strongly agree/agree. For the first five grouping of questions, as shown in Figure 10 (b); this includes, *visibility of system status*, *match between system and the real world*, *user control and freedom*, *consistency and standards*, and *error prevention* groups, the participants were more than 50% satisfied with these features.

In Figure 10 (a), we can see that participants are not happy with questions Q9, Q20, and Q21. Additionally, some participants neither agree nor disagree for questions Q12, Q17 and Q18 depicting uncertainty in these tasks. More details about these results with explanations are provided in Section 6.

We used medians to determine central tendencies [39]. These calculations can be seen in Table 11. The system was generally rated favorably, with the average scores of 15 questions (Q1 - Q8, Q10, Q11, Q13 - Q16, and Q19) rated higher than 3, or neutral. These questions represent the 8 categories established by Nielsen [36]. Three questions (Q12, Q17, and Q18) were rated as neutral as illustrated in Figure 11. We present the central tendencies for the heuristic evaluation by the questions' groupings in Table 11 which is represented as a dashed line in Figure 11.

To gain more insights into the perception of the students with regard to how the responses to the questions are correlated, we model the overall heuristic evaluation survey responses as a network. The results of identifying significant patterns of correlations between the responses to the 21 questions are presented in the following section.

5.2.3. Network analysis results

The 21 heuristic evaluation survey questions regarding the usability of SEP-CyLE represent the entities in the relational data. Each entity contains the aggregates of responses by the 25 students as shown in Figure 10 (a). We model the relational data as a network to study the insights into the students' perceptions about the system regarding various questions asked. Modeling this task-based survey network is achieved by considering the 21 heuristic evaluation survey questions as *nodes*(*vertices*), and the pair-wise associations between the aggregate response vectors against each question estimated using Pearson correlation measure as *edges*(*links*).

The heuristic evaluation survey network is constructed as a weighted undirected complete graph (each node, i.e., question, is connected to every other node in the network) where the weights on the edges between the nodes represent the varying levels of correlations between the responses for these questions. A statistically significant subnetwork is computed from this survey network by computing a matrix of p-values, *PVAL* for testing the hypothesis of no correlation between the responses to the questions (say, i and j) against the alternative that there is a statistically significant correlation. Each element of *PVAL* is the p value for the corresponding element of r. If $PVAL(i,j)$ is small, say less than 0.05, then the correlation $r(i,j)$ is statistically significant. For instance, if the value of r is -0.17576 and the two-tailed value of p is 0.62719, then by normal standards, the association between the responses to the two questions would not be considered statistically significant and the edge is removed from the network. On the other hand, if the value of r is 0.96997 and the two-tailed value of p is 0.04419, then by normal standards, the association between the responses would be considered statistically significant and the edge is retained. The resulting heuristic evaluation survey subnetwork contains clusters of questions representing the strongest statistical correlations among the questions as shown in Figure 12

A clique is a subset of nodes in a network where a node is connected to every other node in the subset representing similar traits among the nodes. A k-clique is a clique of k nodes in which every possible edge is present. The 5-clique (encircled) in Figure 12 (a) shows the collection of questions Q1, Q3, Q5, Q8, Q9 with more similar responses from the students. The students' responses to the question Q3 (with a maximum degree 6) in the subnetwork is more correlated with the questions Q1, Q4, Q5, Q8, Q9 and Q20. The responses to question Q10 (in consistency and standards group) is highly dissimilar with Q18 (in help users recognize, diagnose, and recover from errors group). This means that the sizeable proportion of students agree that the website follows consistency and standard and many students disagree that the system helps users to deal with an error. Also, the questions Q12, Q14, Q15 an Q16 are sparsely connected representing varying responses by the students as shown in Figure 12 (b). The students' responses to the 5 questions listed in Figure 12 (c) are not correlated meaning that the responses to these questions are diverse with no significant common perceptions.

6. Discussion

In this section, we discuss our takeaways and highlights from the exploratory analyses. This includes our comments and recommendations intended to address any shortcomings and opportunities for improvement. The pre/posttest results, students' final scores, and SEP-CyLE UX survey results are used to illustrate the first research question (RQ1). The cognitive walkthrough results, the heuristic evaluation results, and the network analysis results are used to address the second research question (RQ2).

6.1. RQ1 — SEP-CyLE's utility

The pre/posttest results indicate that using SEP-CyLE as a cyberlearning tool for students learning software testing can significantly improve their understanding and use of software testing techniques and tools. This is clearly obvious using the results shown in Table 5 and Figures 3 and 4. The posttest scores were much better than the pretest scores in the treatment group showing that SEP-CyLE would be one factor in helping students improve their conceptual knowledge on testing. ANOVA test returned p values of < 0.05 in all cases showing that posttest scores were significantly better than pretest scores.

The final scores of the treatment and control groups are compared to evaluate the SEP-CyLE utility and usefulness in enhancing the software testing skills of students. Figure 13 shows the final scores (in %) of the control group (Section A) plotted against the treatment group (Section B). Identical exams are provided for both groups. However, the average score of all students in the treatment section was found to be comparatively higher than that of students in the other section. The only substantial difference was in the application of their testing learning and experience in all the assignments and exams.

The proposed SEP-CyLE UX survey extends conventional web utility criteria and integrates them with criteria derived from cyberlearning design so that to address the utility of cyberlearning technologies (i.e., collaborative learning and gamification) and address the users as learners. The SEP-CyLE UX survey results, as shown in Figure 5 and Figure 6, confirm the usefulness of SEP-CyLE in its designed context (software testing) for the specified users (students). Also, it is obvious that the cyberlearning technologies used in the design of SEP-CyLE, both collaborative learning and gamifications adhere to the utility requirements and can help the students in their insightful learning. The detailed and more pragmatic experimental analysis help us understand how users learn with these types of learning mechanisms.

In conclusion, students do find SEP-CyLE an engaging and useful learning resource for learning software testing concepts, techniques and tools.

6.2. RQ2 — SEP-CyLE's usability

The cognitive walkthroughs results and the heuristic evaluation results using network analysis are both used to evaluate the SEP-CyLE usability and measure its effectiveness. Specifically, we used the cognitive walkthroughs results as an objective evaluation, since students used

Table 11. Calculated medians by question number and question grouping.

Question	A			B			C			D		E			F	G			H		
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
By Question	4	4	4	4.5	4	4	5	4	2	4	4	3	4	3.5	3.5	4	3	3	3.5	2	2
By Group	4			4			4			4		4			3.5	3			2		

A = Visibility of system status; B = Match between system and the real world; C = User control and freedom; D = Consistency and standards; E = Error prevention; F = Flexibility and efficiency of use; G = Help users recognize, diagnose, and recover from errors; H = Help and documentation.

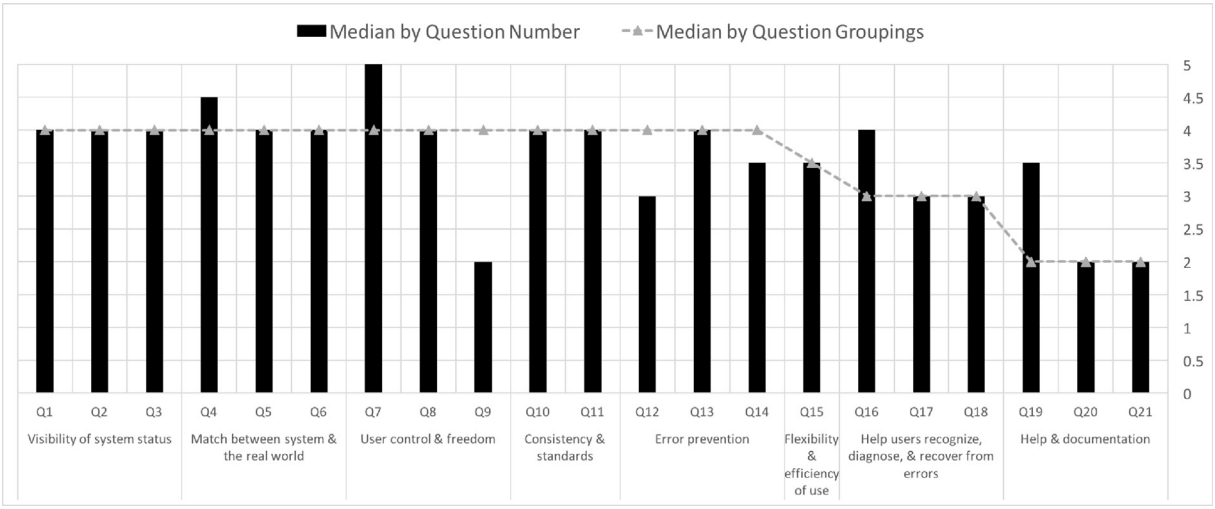


Figure 11. Medians from heuristic evaluation by question number (columns) and by question groupings (dashed line).

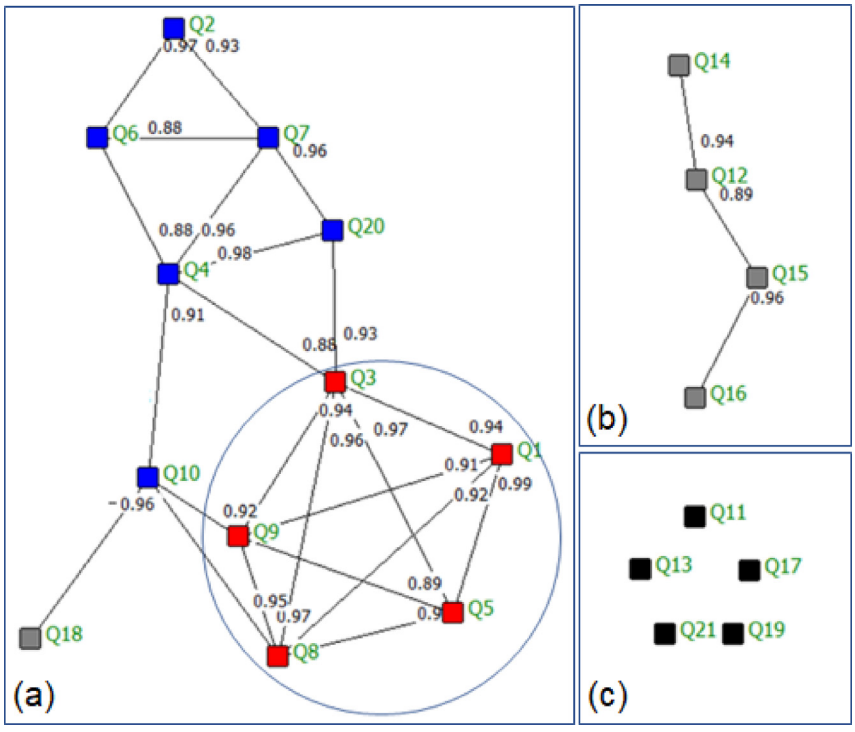


Figure 12. Statistically Significant Heuristic Survey Graph. 21 survey questions - nodes, pair-wise associations between the responses estimated using the statistical Pearson Correlation (r) - edges (a) Subnetwork represents significant correlations of responses between 12 questions; a 5-clique shown in circle with 5 questions of similar responses; High negative correlation of responses between Q10 and Q18 (b) Sparsely Connected Subnetwork of 4 questions showing varied responses (c) Disconnected subnetwork with very dissimilar responses to 5 questions.

thinking-aloud with specific task(s). And we used the heuristic evaluation results as a subjective evaluation, since each student is asked individually to evaluate the usability of the SEP-CyLE based on his experience after using the SEP-CyLE learning objectives and tutorials.

6.2.1. *Cognitive walkthrough analysis*

The tasks where participants generally performed well include Tasks 1–3, 5–7, 9, and 10. This indicates that controls for these tasks have good visibility, affordability, and feedback. It is apparent that several participants had significant difficulty on some tasks. For example, participants

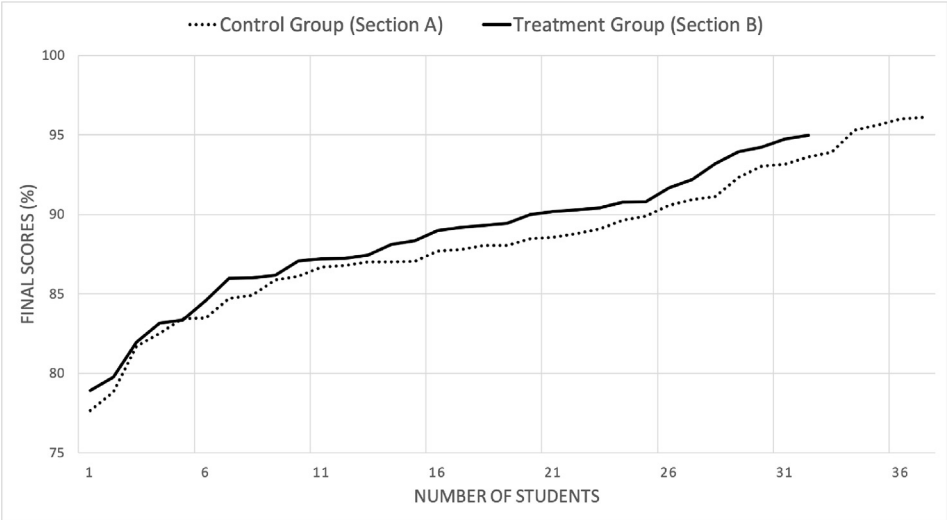


Figure 13. Final Scores comparison between treatment group and control group.

struggled with Tasks 4, 8, 11, and 12. From the grouping of tasks, task 4 is focused on the use of LOs, and Tasks 8, 11, and 12 are focused on the use of software testing tutorials. One common thing that all these tasks address is that the participants are required to identify numbers, as follows:

- T4: Identify the number of pages provided in the content of the LO assignment.
- T8: Identify the number of tutorials provided by the course.
- T11: Identify the number of videos provided with the tutorial.
- T12: How many external tutorial links are available?

Users likely struggled in these questions due to a lack of visibility and affordability with the SEP-CyLE's controls. For example, after a careful investigation of the recorder videos, we found that: (1) some LOs content pages have no 'next' button associated with it; (Students used the navigation bar to proceed to the next LO's content pages) (2) the 'next' button is not disabled while the student is at the last page of content, thus causing confusion that some students try clicking it; (3) some students thought that the LO is the same as tutorial since the number of content pages associated with the LO confused students.

Similarly, we noticed the following issues while a student is at the tutorial page; (1) the text titles are sometimes underlined that makes him think that these are hyperlinks, but he could not click them anyways, (2) some hyperlinks have no linking websites associated, (3) some students find it difficult to locate and navigate to tutorials from the home page, (4) some students think LO as a tutorial, and (5) some students tend to take the quiz straight away without reading the content in the LO or tutorial, therefore have no chance to count the number of pages in the LO assignment nor the number of tutorials provided.

We conclude that there needs to be better association of the intended behavior to the intended task. We recommend enabling the practice quiz control just after the student skims through or navigates through all the pages under the content. We would like users to be able to complete these types of tasks easily in the system, but students seem to have mixed results. While participants were easily able to see each page, tutorial, and video in the selected LO, they have to keep a count the "pages, tutorials, videos, external links" manually to answer these questions. There is a visibility issue in that some participants do not realize that clicking or hovering on controls yield a response to the question at hand.

Users' actions while using SEP-CyLE indicate that not all controls are visible to the user. Specifically, users failed to recognize some of the possible interactions with the LOs, resulting in longer response times, incorrect answers, and confusion about the system as a whole. Additionally, some controls did not afford the action required to utilize them. Users expected controls to exist that did not, such as going back from one assignment. Some users first tried to arrange the LO's assignments in an order to obtain information more quickly. However sorting the LOs is currently not a control available.

6.2.2. *Heuristic evaluation analysis*

As demonstrated in Figure 11, the only central tendency measurement for any grouping that fell below the neutral option dashed line is group number 8 (*Help and documentation*). Group number 7 (*Help users recognize, diagnose, and recover from errors*) was rated as neutral. The remaining six groupings of questions had consistently positive scores (> 3). The most poorly rated questions were Q9, Q20, and Q21. These questions cover two groupings: *user control and freedom* and *help and documentation*.

The grouping *help and documentation* was rated poorly overall when compared with other portions of the system. The grouping *help users recognize, diagnose, and recover from errors*, was rated as neutral by students. This is likely because few participants encountered issues that were interpreted as errors. However, this portion of the system could be improved through the addition of FAQ pages and system tutorials to reduce error rates and to provide information on what to do when errors are encountered.

After studying the participants' responses individually for Q9 (*Are there dialog prompts that are unnecessary when the user is trying to leave a page.*), we discovered that most of the responses were below the neutral value. We believe that respondents misunderstood this question in which a lower response leads to a positive outcome. Specifically, we stated the question poorly, resulting in a lower rating leading to a positive outcome for this particular usability question. However, if this is not the case, subject responses to Q9 could be improved through modification to controls on the tutorials, LOs, and quizzes interfaces and with the addition of text based tutorials regarding the system's intended usage as well as better feedback in the system overall. Furthermore, shortcomings identified in Q20 (*Live* support *is available to the user*) and Q21 (*User can email for assistance*) can also likely be improved by providing contact information and online support as tutorials.

In general, the heurisitc evaluation results showed that students thought that the SEP-CyLE interface design was effective and enabled them to explore all learning objectives, testing tutorials, and practice quizzes. They could find certain buttons and tabs on the interface. One item that required attention was "Help and documentation", as shown in Figure 10 (b). The overall mean for this grouping is below "neutral". A possible explanation of this result is that to complete any given task,

students had to go through two or three different pages and might have been looking for a quick help and/or documentation to answer that task. Issues found through the heuristic evaluation can be easily addressed through tutorials, additional visual cues, documentation, or controls visually separate from the inline contents.

7. Threats to validity

One major threat to validity in our utility evaluation is that our application of the framework is limited to only one cyberlearning environment at this time, SEP-CyLE, a tool used in programming and software engineering courses. However, the ultimate goal of our work is to present an evaluation framework for cyberlearning environments. We are the first to present such an evaluation since we contend current evaluation frameworks are not applicable to this problem.

There are two primary threats to our validation of the proposed cyberlearning evaluation framework. The first involves the structure of the think-aloud protocols. The student users had varied experiences, they performed think-aloud protocols in groups of 2 or 4, and conducted the protocol independently of a researcher. This may have impacted the quality of results. Student users may have been subjected to "group think". The second threat stems from the limitations of the questions in the heuristic evaluation. It is possible that the heuristic evaluation does not cover every relevant topic of design to properly evaluate the site. Additionally, there is some debate over whether a 5-point Likert scale is sufficient to determine individuals' attitudes toward statements and alternatives, compared to the use of fuzzy logic-based responses [43].

8. Conclusion and future work

We presented our efforts in evaluating SEP-CyLE, a cyberlearning, web-based, learning environment. Our experiments demonstrated participants were able to utilize SEP-CyLE efficiently to accomplish the tasks we posed to them and to enhance their software development concepts, specifically, software testing. We discovered areas of improvement in the visibility and navigation of SEP-CyLE's current design. Specifically, participants often failed to recognize all the available LOs, tutorials and quizzes, resulting in some incorrect results, and confused participants. Also of note, some struggled switching from one LO to another, resulting in some affordability, visibility, and feedback issues. While our study identified areas of improvement, these are solvable. Our recommendations include adding text-based tutorials to improve user understanding of the LOs, tutorials and quizzes. Improvements can be made to address specific user feedback regarding the system visibility and the types of controls available to users, such as sorting LOs in the assignments section. Additionally, comparison of SEP-CyLE with other cyberlearning environments using the same learning and comprehension tasks is an interesting evaluation venture we would consider.

Overall, SEP-CyLE demonstrates a strong ability in assisting users in software development concepts and software testing tools and techniques. SEP-CyLE successfully allows users to quickly and efficiently gain essential software testing knowledge and apply the UI/UX evaluations techniques they have learned in the class. It is our hope that our evaluation design and results could be used by those conducting research on UI/UX evaluation of cyberlearning environments.

Declarations

Author contribution statement

H. W. Alomari, V. Ramasamy, J. D. Kiper, G. Potvin: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work is supported in part by the NSF under grants DUE-1225742 and DUE-1525112. Any opinions, findings, and conclusions or recommendations expressed here are those of the author(s) and do not necessarily reflect the views of the NSF.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] C.L. Borgman, H. Abelson, L. Dirks, R. Johnson, K.R. Koedinger, M.C. Linn, C.A. Lynch, D.G. Oblinger, R.D. Pea, K. Salen, et al., Fostering Learning in the Networked World: the Cyberlearning Opportunity and challenge. A 21st century Agenda for the National Science Foundation, 2008.

[2] J. Roschelle, W. Martin, J. Ahn, P.E. Schank, Cyberlearning Community Eport: the State of Cyberlearning and the Future of Learning with Technology, 2017.

[3] P.J. Clarke, A.A. Allen, T.M. King, E.L. Jones, P. Natesan, Using a webbased repository to integrate testing tools into programming courses, in: Proceedings of the ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion, OOPSLA '10, ACM, New York, NY, USA, 2010, pp. 193–200.

[4] P.J. Clarke, D.L. Davis, R. Chang-Lau, T.M. King, Impact of using tools in an undergraduate software testing course supported by wrestt, ACM Trans. Comput. Educ. 17 (4) (2017) 18:1–18:28.

[5] P.J. Clarke, D. Davis, T.M. King, J. Pava, E.L. Jones, Integrating testing into software engineering courses supported by a collaborative learning environment, Trans. Comput. Educ. 14 (3) (2014) 18:1–18:33.

[6] R.C. L, J. K, Y. F, Peter J. Clarke, Debra L. Davis, G.S. Walia, Using Wrestt Cyberlearning Environment in the Classroom, Computers in Education Division, ASEE, 2017. https://www.asee.org/public/conferences/78/papers/20158/view.

[7] H.W. Alomari, J.D. Kiper, G.S. Walia, K. Zaback, Using web-based repository of testing tutorials (wrestt) with a cyber learning environment to improve testing knowledge of computer science students, in: 124th ASEE Annual Conference & Exposition, Computers in Education Division, ASEE, 2017. https://www.asee.org/public/conferences/78/papers/20158/view.

[8] V. Ramasamy, H.W. Alomari, J.D. Kiper, G. Potvin, A minimally disruptive approach of integrating testing into computer programming courses, in: 2nd IEEE/ACM International Workshop on Software Engineering Education for Millennials (SEEM), Computers in Education Division, SEEM, 2017. https://www.asee.org/public/conferences/78/papers/20158/view.

[9] I. Standardization, ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs): Part 11: Guidance on Usability, 1998. https://books.google.com/books?id=TzXYZwEACAAJ.

[10] D.V. Greunen, J. Wesson, Formal usability testing of interactive educational software: a case study, in: Proceedings of the IFIP 17th World Computer Congress - TC13 Stream on Usability: Gaining a Competitive Edge, Kluwer, B.V., Deventer, The Netherlands, The Netherlands, 2002, pp. 161–176. http://dl.acm.org/citation.cfm?id=646869.759751.

[11] J. Nielsen, Usability 101: Introduction to Usability, 2012. https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

[12] P.J. Clarke, Sep-cyle Cyberlearning Environment, 2010. https://stem-cyle.cis.fiu.edu/institutions/.

[13] Y. Fu, P.J. Clarke, Gamification based cyber enabled learning environment of software testing, in: 123rd ASEE Annual Conference & Exposition, Computing and Information Technologies (CIT), ASEE, 2016. https://www.asee.org/public/conferences/78/papers/20158/view.

[14] P.J. Clarke, J. Pava, Y. Wu, T.M. King, Collaborative web-based learning of testing tools in se courses, in: Proceedings of the 42Nd ACM Technical Symposium on Computer Science Education, SIGCSE '11, ACM, New York, NY, USA, 2011, pp. 147–152.

[15] A. Goswami, S. Abufardeh, Using a web-based testing tool repository in programming course : an empirical study, in: Proceedings of 2014 International Conference on Frontiers in Education: Computer Science and Computer Engineering, FECS, 2014.

[16] R. Chang-lau, P.J. Clarke, Web-based Repository of Software Testing Tutorials a Cyberlearning Environment (WReSTT-CyLE), Jun. 2017. http://wrestt.cis.fiu.edu/.

[17] V. Ramasamy, U. Desai, H.W. Alomari, J.D. Kiper, Tp-graphminer: a clustering framework for task-based information networks, in: International Conference on Systems, Computation, Automation and Networking, 2018.

[18] J.S. London, Exploring Cyberlearning through a Nsf Lens, 2012.

[19] J. Roschelle, New Cyberlearning Community Report: Five Great Reasons to Read it, 2017. https://digitalpromise.org/2017/10/02/new-cyberlearning-community-report-five-great-reasons-read/.

[20] K.M. Qureshi, M. Irfan, Usability Evaluation of E-Learning Applications, a Case Study of It's Learning from a Student's Perspective, Master's Thesis, School of Computing, Kashif Manzoor Qureshi 0046-700511015, Muhammad Irfan, 2009, 0046767142792.
[21] J. Nielsen, R. Molich, Heuristic evaluation of user interfaces, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '90, ACM, New York, NY, USA, 1990, pp. 249–256.
[22] E. Frkjr, M. Hertzum, K. Hornbk, Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated? CHI, 2000.
[23] J. Nielsen, User Testing: Why & How, 2012. https://www.nngroup.com/videos/user-testing-jakob-nielsen/.
[24] J. Nielsen, Thinking Aloud: the #1 Usability Tool, 2012. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/.
[25] L. Zia, Nsf Support for Research in Game-Based Learning, National Academy of Sciences, Washington, DC, 2005.
[26] D.B. Montfort, S. Brown, What do we mean by cyberlearning: characterizing a socially constructed definition with experts and practitioners, J. Sci. Educ. Technol. 22 (1) (2013) 90–102.
[27] J.L. Moore, C. Dickson-Deane, K. Galyen, e-learning, online learning, and distance learning environments: are they the same? Internet High Educ. 14 (2) (2011) 129–135, web mining and higher education: Introduction to the special issue, http://www.sciencedirect.com/science/article/pii/S1096751610000886.
[28] M. Nichols, E-learning in context, E-Primer series 1 (2008).
[29] A. Pereira, Cyberlearning vs. Elearning - Is There a Difference?, 2016. https://www.thetechedvocate.org/cyberlearning-vs-elearning-difference/.
[30] J. Wesson, Usability evaluation of web-based learning, in: TelE-Learning, Springer, 2002, pp. 357–363.
[31] C. Lewis, Using the "thinking Aloud" Method in Cognitive Interface Design, Research Report, IBM T.J. Watson Research Center. https://books.google.com/books?id=F5AKHQAACAAJ.
[32] L. Zeng, Designing the user interface: strategies for effective human computer interaction (5th edition) by b. shneiderman and c. plaisant, Int. J. Hum. Comput. Interact. 25 (7) (2009) 707–708.
[33] B.L. Smith, J.T. MacGregor, What Is Collaborative Learning, 1992.
[34] C. Li, Z. Dong, R.H. Untch, M. Chasteen, Engaging computer science students through gamification in an online social network based collaborative learning environment, Int. J. Inf. Educ. Technol. 3 (1) (2013) 72.
[35] M.A. Storey, B. Phillips, M. Maczewski, M. Wang, Evaluating the usability of web-based learning tools, Educ. Technol. Soc. 5 (3) (2002) 91–100.
[36] J. Nielsen, How many Test Users in a Usability Study?, 2012. https://www.nngroup.com/articles/how-many-test-users/.
[37] W. Albert, T. Tullis, Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics, Newnes, 2013.
[38] R. Likert, A technique for the measurement of attitudes, Arch. Psychol. 22 (140) (1932) 55.
[39] S. Jamieson, Likert scales: how to (ab) use them, Med. Educ. 38 (12) (2004) 1217–1218.
[40] J. Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.
[41] M. Newman, Networks: an Introduction, Oxford University Press, Oxford, UK, 2010.
[42] D. Kornbrot, Pearson Product Moment Correlation. Encyclopedia of Statistics in Behavioral Science, John Wiley and Sons, New York, 2005.
[43] Q. Li, A novel likert scale based on fuzzy sets theory, Expert Syst. Appl. 40 (5) (2013) 1609–1618. http://www.sciencedirect.com/science/article/pii/S095741741201069X.