Received: 25 September 2023 | Revised: 24 January 2024 | Accepted: 28 January 2024

DOI: 10.1002/cae.22725

RESEARCH ARTICLE

WILEY

Facilitating and automating usability testing of educational technologies

Mikel Villamañe[1] | **Ainhoa Alvarez**[2]

[1]Languages and Computer Systems, University of the Basque Country UPV/EHU, Bilbao, Spain

[2]Languages and Computer Systems, University of the Basque Country UPV/EHU, Donostia, Spain

Correspondence
Ainhoa Alvarez, Languages and Computer Systems, University of the Basque Country UPV/EHU. Manuel Lardizabal pasealekua, 1, 20018 Donostia-San Sebastian, Gipuzkoa.
Email: ainhoa.alvarez@ehu.eus

Funding information
Department of Education, Universities and Research of the Basque Government, Grant/Award Number: ADIAN, IT-1437-2; University of the Basque Country UPV/EHU

Abstract

Usability evaluation is a key element to ensure a positive user experience with any software and it is especially important in educational software tools where there are many different actors involved (lecturers, students, administrators, etc.). However, evaluating usability is not an easy task for nonexpert evaluators. To facilitate this evaluation task, this article presents a Methodology for Usability Testing (MUT) and a system (CALMUT) that assists nonexpert evaluators in the application of the methodology by automatizing the calculations and facilitating their interpretation. This can be very useful for learning and instructional designers but also to people involved in the decision of introducing or not a new educational software. To develop the proposal, a literature review of different usability metrics, methods, and systems was carried out first, followed by a selection and adaptation for novice usability evaluators. This article also presents a case study where lecturers tested the usability of an educational software following the proposal and shows that using MUT and CALMUT helps people without previous experience detect the main usability problems of educational systems before deciding whether to use them or not.

KEYWORDS
educational technologies, usability testing

1 | INTRODUCTION

Educational or learning technologies are included in all education levels. However, just using them does not imply any improvement in the educational process and several aspects such as usability must be taken into account [41].

Usability, along with other attributes such as utility, robustness, privacy, or desirability, is essential to provide a positive user experience (UX) [5, 29] and guarantee the quality and success of a software [6]. This is especially the case for a successful technology adoption [41]. In the case of educational software tools, different studies have detected that their usability is also of great relevance, as it influences student motivation, retention, success, and achievement [7, 10, 18].

Usability evaluation analyzes how satisfied are users with the system and allows to assess how intuitive the system is, the problems that could arise during its use and the formation required to be able to use it correctly [42]. There are some aspects of usability that can be measured, but it has been noted in the literature that

finding and selecting valid indicators for these aspects is not easy [1, 24]. There are many usability models that propose the use of different indicators to evaluate usability, but according to Hasan & Al-Sarayreh [12] most of them use the indicators suggested in the ISO 9241-11:2018 [16] which is the definition with the broadest perspective of usability. The ISO 9241-11:2018 defines usability as "the extent to which a product or a service can be used by users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specific context of use". Effectiveness is "the accuracy and completeness with which users achieve the specified goals." Efficiency is "the resources expended in relation to the accuracy and completeness" and satisfaction "is the freedom from discomfort, and positive attitudes towards the user of the product."

Evaluating usability is not an easy task, even more for nonexpert evaluators. This is often the case for learning and instructional designers as usability evaluation is not considered one of the core competencies of educational technology researchers [24, 41]. Moreover, due to the increase in the availability of educational systems, often teachers become the ones who select the tools to be used what makes important to provide them with tools to facilitate carrying out the usability tests of those systems [7].

According to Cayola and Macías [6], there are many methods to evaluate usability, but it is difficult to select and apply them due to the limited reporting and explanations available. Moreover, most of the information on the application of the methods is found in web repositories without any systemization [28, 38], what makes even harder to evaluate usability. Therefore, even with its relevance, in general, any person (i.e., developer, designer) without previous experience involved in software development have problems to correctly apply any of the existing models as there are few clear guidelines about how to apply the many individual existing methods for evaluating usability [4, 32].

To help learning designers to establish which data to be gathered and which techniques to use when evaluating usability, we present the Methodology for Usability Testing (MUT). MUT provides a guide that assists educational technology researchers at the time of measuring a system's usability and when interpreting the results. MUT measures usability using the effectiveness, efficiency, and satisfaction indicators and proposes a set of metrics for each of them. As applying the required mathematical formulae can be complex and tedious, we also present CALMUT (a usability calculator for the MUT methodology) to help in this task. In comparison to other proposals presented in the following section, the combination of this proposal will help nonexpert users to evaluate the usability of online and offline systems or prototypes and will produce a detailed report with explanations of the results and indications of the elements to be improved.

This article is structured as follows. First, the background and some related works are presented. Next, the MUT methodology and the CALMUT system are presented. After, an example of use in the usability evaluation of an educational system is depicted. Finally, some conclusions are drawn.

2 | BACKGROUND AND RELATED WORK

First, we present the results of the study carried out to determine the usability metrics for each indicator. Next, we analyze some existing systems intended to help evaluate usability.

2.1 | Usability metrics

To determine the metrics, we conducted a literature review. We searched the terms "usability," "measuring," "testing," "methods," and "metrics" linked with logical connectors in recognized databases in the field of computer systems' usability: Google Scholar, IEEE Xplore, and ACM Digital Library. The results were filtered to exclude papers published more than 20 years ago and ordered by their number of cites.

Then, we carried out the screening with the following inclusion criterion: the article should be generic (or oriented to educational systems) and address one or more of the proposed indicators. During the screening process, the authors detected that there were only a few papers that met the criterion, being [14, 36] the most relevant. The study of metrics presented in those papers were taken as the basis of this work. A new search about the most relevant metrics mentioned in the papers was conducted to analyze whether there were more recent studies that went against the results and conclusions of those papers and found none relevant. The review results are depicted next.

The most common metric to evaluate effectiveness is the tasks' completion rate [14, 36] which is calculated from the number of users that are able to complete each task.

Satisfaction can not be automatically determined, so it is always measured using questionnaires that users fill in after working with the system. There are some standard questionnaires whose validity and reliability have been tested and which are widely used [14, 37]. Among the most used standard questionnaires the system usability scale (SUS), System Usability Scale [3, 33], the CSUQ, Computer System Usability

Questionnaire [21], or the Usability Metric for User Experience (UMUX) can be found [9]. The three of them provide similar results [21] and among those, the SUS questionnaire is the most used because it is an effective and efficient tool [3, 21].

Finally, execution time and use patterns (how the system's interface is used) are the main metrics used to analyze efficiency [14], which can be enriched with metrics such as lostness [34]. Some authors measure the number of keystrokes [26] or mouse clicks [8] used to solve each task. Other authors measure the user's deviation from the optimal solution to the task [25, 35].

To calculate those metrics, it is required to run the test and compile the required data. In general, the methods that automate data gathering are the most advantageous [17]. Moreover, systems that can automatically collect the required information facilitate the deployment of *instrumented remote evaluations* that facilitate testing to be carried out with more participants, reduce the budget and the time needed to perform the test [43].

2.2 | Systems to help evaluate usability

We carried out an analysis of systems that help evaluate usability, identifying that none of them is supported by an established methodology.

It has been found that many enterprises provide the possibility of running usability tests. Most of them help find people to conduct usability tests and some of them also allow to record user actions when running the test and make annotations. In general, they provide the possibility of generating some kind of simple analysis report and are oriented to the evaluation of web applications.

One of the most complete and powerful systems is Loop11 [22], it allows creating usability tests by defining the tasks to be carried out and compiling the execution results. Once the information is collected it generates a report with information regarding the execution of the task. Figure 1 shows a screen with some of the results provided by the system.

However, it has some important drawbacks. On the one hand, it only allows evaluating online applications or online available prototypes. On the other hand, and what is more important, it does not provide information regarding acceptance criteria for the obtained results.

There is another kind of systems with less functionalities, we have denoted them calculators because all they do is calculate different metrics. An example of such system is MeasuringU [27], which allows (see Figure 2) different calculus to be executed.

MeasuringU allows making different calculus but it does not offer an overall view or guide of the study and

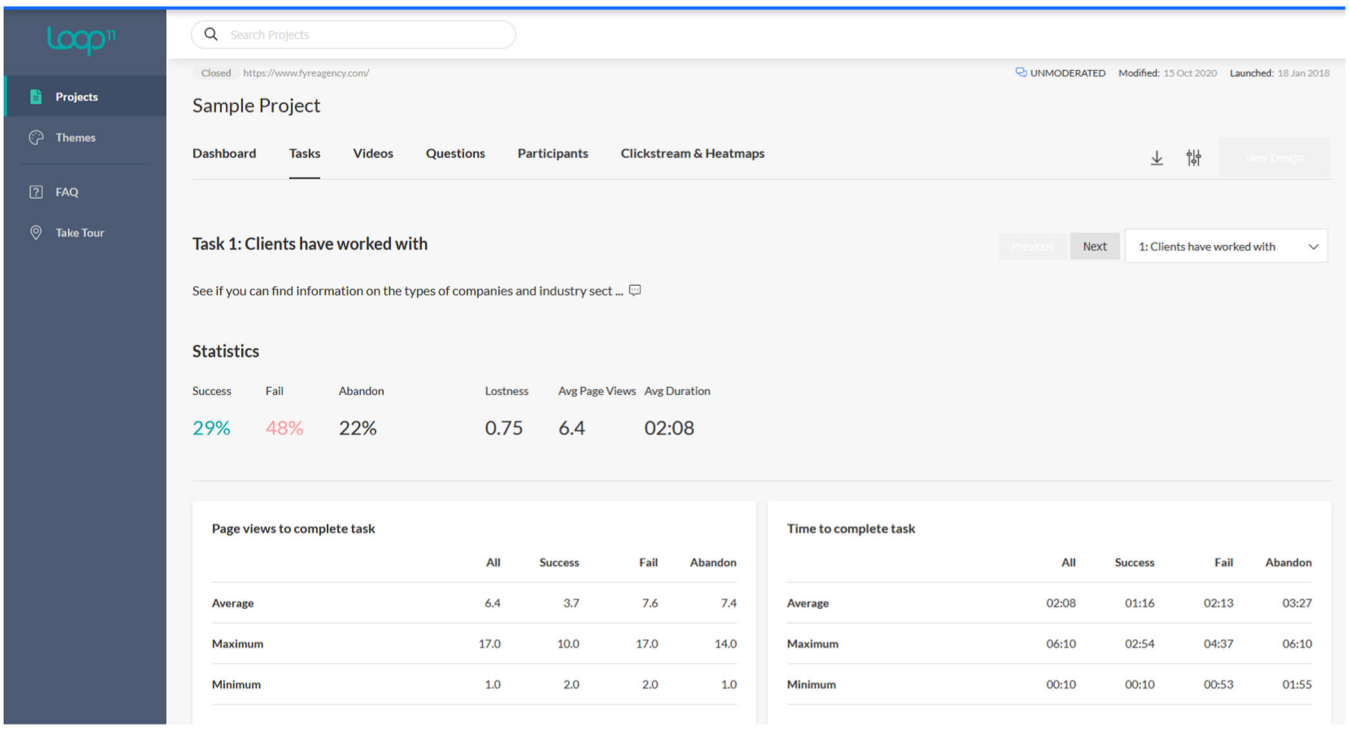


FIGURE 1 Loop11 interface.

does not help to select which calculus and why to be used.

There are also available in the web several spreadsheets to make similar calculus, but again without any help or assistance. Finally, many applications can be found to automate the calculus of the SUS questionnaire [39].

The first-mentioned tools are quite expensive and often quite complex. Moreover, they are usually thought to evaluate the usability of web pages. The calculators can be an interesting set of tools, but in general, it is not easy to use them as they have not any guide or description. Therefore, for novice evaluators other type of tool is required.

3 | MUT & CALMUT

The different calculators and tools presented in the previous section can help users to carry out some parts of usability evaluation. However, for novice users, this is not enough and they will continue having problems applying them with rigor. In this context, we present the MUT methodology and CALMUT, a tool to help apply the MUT workflow and metrics that have been defined to help nonexpert usability evaluators (which is often the case with educational technology developers).

MUT measures usability using the effectiveness, efficiency, and satisfaction indicators according to the ISO 9241-11:2018 definition and proposes a set of metrics for each of them. It uses the test method, in which representative users solve concrete tasks and when the execution is finished, the results are analyzed [11, 20].

CALMUT facilitates this data analysis using graphs to visualize data and help in the search for patterns and potential outliers which is a better way of doing it than sifting through textual data [2].

The MUT methodology stablishes three stages (see Figure 3) that provide a guide to assist usability testers. These stages are described in Section 3.2.

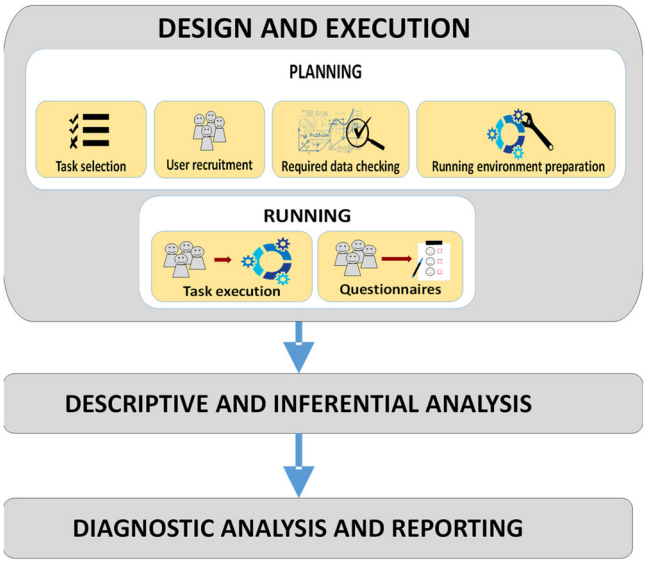


FIGURE 3 Stages of Methodology for Usability Testing (MUT).

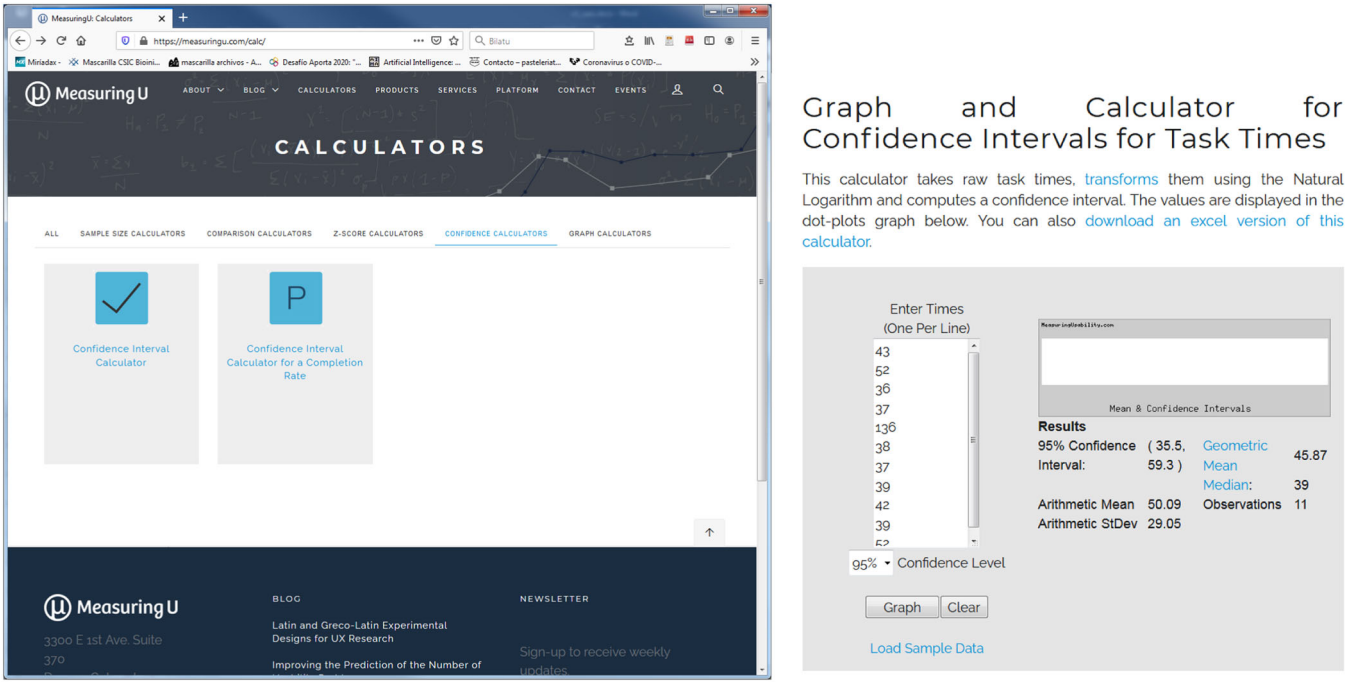


FIGURE 2 MeasuringU interface.

3.1 | CALMUT development

CALMUT has been developed following the prototype-driven development approach defining a prototype for each of the three main menu options (see Figure 4). This has facilitated the incremental inclusion of requirements in the system.

The three menu options are related to the different phases and stages of MUT. The option "Usability test structure definition" (Figure 4a) supports the planning phase; the option "load test data" supports the collection of data obtained in the running phase (Figure 4b); and the "Calculate results" option (c) supports the remaining two stages.

CALMUT has a Node.js developed back-end that follows a representational state transfer architecture where the business logic of the project is included. It is divided into different modules, and it includes a middleware that works as a security filter. The front end, developed using Angular, is in charge of the system graphical interface. Figure 5 shows the system's general architecture.

3.2 | Description of the MUT stages and its support by CALMUT

This section describes the different MUT stages and how the CALMUT system gives support to each of them.

3.2.1 | Design and execution stage—planning phase

The planning phase is divided into four steps and the phase result is the test structure stablished with four elements: nodes, testers, questionnaires, and tasks.

The task selection step involves determining the tasks to be tested, that must be selected among the ones foreseen to be executed frequently and those that represent better the system's functionality. For each task, its reasonable execution time must be estimated and its optimal and alternative paths identified. CALMUT interfaces allow defining the tasks (see Figure 6a) and its paths, which include the nodes (interface sections) to be visited, the order in which they must be visited and the number of clicks required on each (see Figure 6b).

For each task, the order in which each tester will execute it must be determined. MUT proposes to randomly obtain the execution order to decrease the influence that tiredness produced during the execution of the first tasks could have on the remaining [23, 36].

Regarding the user recruitment step, MUT proposes to first determine which are the characteristics that testers should have to represent the target or potential users of the system using the guidelines in Hinderer & Nielsen [13]. As there is no consensus regarding the optimal size for the sample, MUT proposes to use the rule of 10 ± 2 users according to the carried out bibliographical review [15].

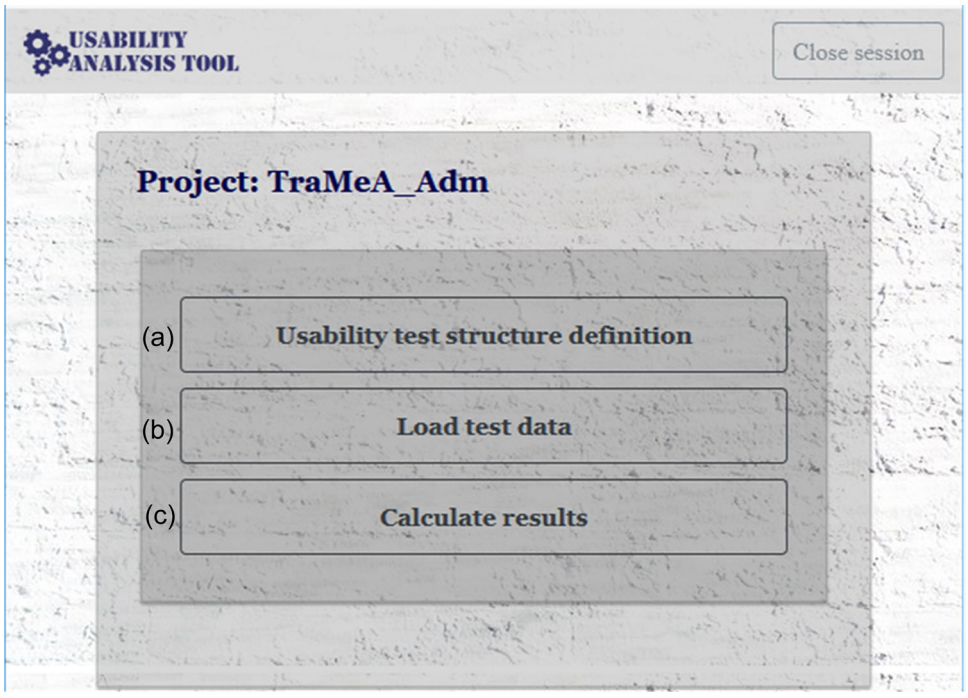


FIGURE 4 Main CALMUT screen.

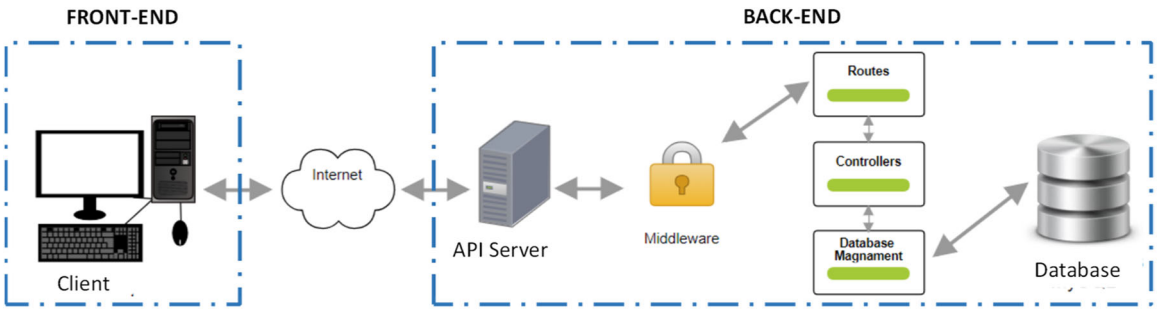


FIGURE 5 General architecture of the CALMUT system.

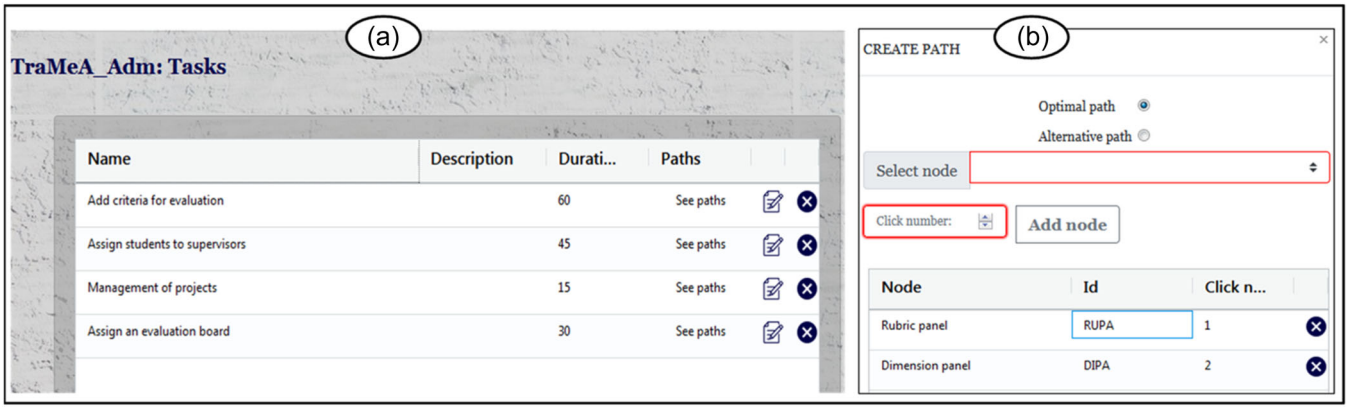


FIGURE 6 (a) Task definition and (b) path definition interfaces.

CALMUT gives an automatic identifier to each tester, to provide anonymity, and allows the creation of testers one by one or imported from a comma-separated values formatted file with basic information (name, gender...).

In the data-checking step, the information needed to carry out the usability test (see resume in Table 1) is analyzed. As usability tests of educational technologies should be done at the early stages of the system and not only at the end [24], the deployment state of the system must be considered to consider which information is already being gathered and which not.

In the running environment preparation step, the mechanism to collect the data that is not already being gathered is determined. Different possibilities include the modification of the system to gather the missing information, to have a human observer manually gather data or to generate a prototype. However, it must be taken into account that the collection of data should be automatized to the maximum possible [17].

3.2.2 | Design and execution stage—running phase

In the running phase, aspects such as the arrangement of meetings with the users when a human observer is required, or the establishment of a deadline to perform the tasks when an instrumented remote evaluation is selected are considered.

After that, testers are provided with access to the system, the description of the tasks to be carried out, the execution order, and the satisfaction questionnaire to answer when finished.

Once the usability test is run, data must be stored in CALMUT to be analyzed. CALMUT allows to import the data from comma-separated value files or to load it directly via its web services. It also provides validated language versions of the SUS questionnaire [3, 19] to allow the tester to select which one to use.

3.2.3 | Descriptive and inferential analysis

In this stage, both descriptive and inferential analysis of collected data is carried out. The first one is in charge of summarizing the information regarding the sample. The second one centers on generalizing the results for the population. The analysis is carried out for the three indicators previously identified and shown in the "Calculate results" menu: effectiveness, satisfaction, and efficiency (see Figure 7). The CALMUT interfaces for each indicator will be described in the next section.

TABLE 1 Data to be collected.

MUT metric	Data to be collected for each tester
Completion rate (Effectiveness)	Success or failure finishing each task
System Usability Scale (Satisfaction)	Answers to SUS questionnaire
Execution time (Efficiency)	Execution time for each task
Lostness (Efficiency)	Interface sections visited to carry out each task
Use patterns (Efficiency)	Interface sections visited to carry out each task and their visiting order. Number of clicks in each interface section

Abbreviation: MUT, Methodology for Usability Testing.

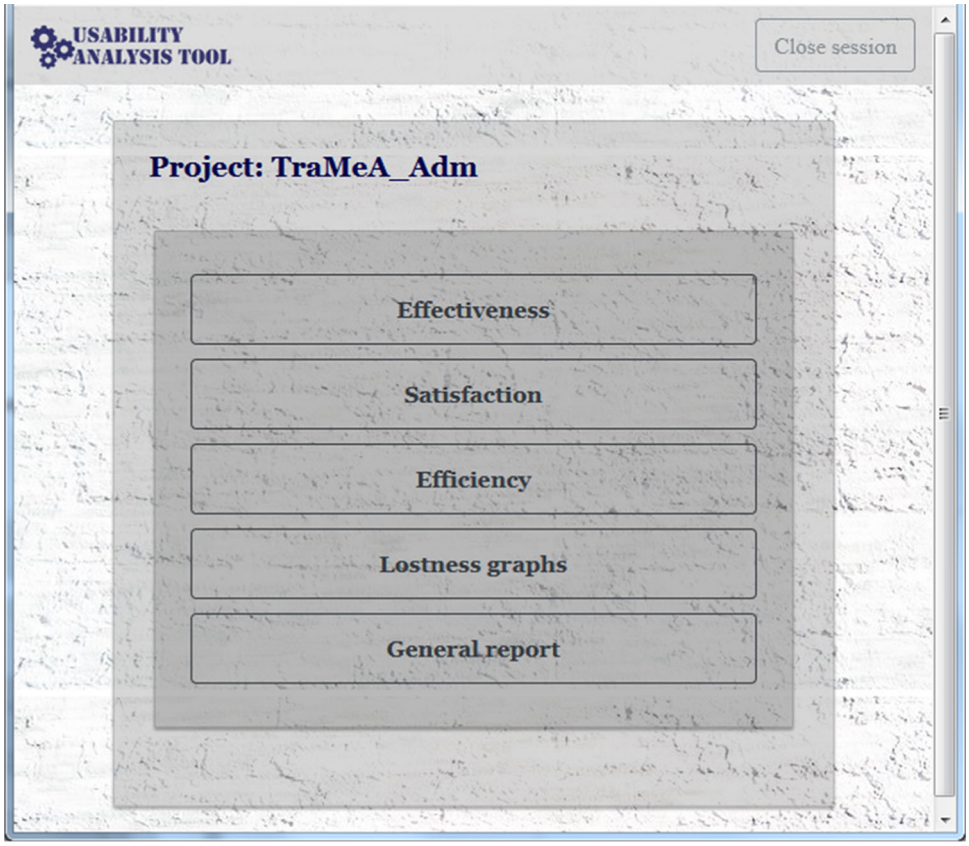


FIGURE 7 General analysis interface.

The metric selection for each indicator has been done considering the study previously presented in the background and related work section.

For effectiveness, the completion rate metric [14, 36] is proposed. This metric indicates, in a range between 0 and 1, the rate of users that have completed the task. When the evaluation of a preliminary version of the system is carried out, the effectiveness is considered acceptable when the completion rate is equal or greater to 0.7 (meaning 70% of users would finish the task). When a final version of the system is being evaluated, a value of 0.95 or higher should be reached [30].

Satisfaction is measured using the SUS questionnaires fulfilled by users after performing tasks with the system.

Finally, for efficiency, the execution time metric [14] that measures the time needed by the users to accomplish each of the defined tasks is used.

To enrich this information lostness is also used in MUT. This value is given in a [0,1] scale and it indicates to which extent the user has not followed the best possible path to finish a task [34]. The ideal value for the lostness level is 0, but values lower that 0.4 are also acceptable as they only indicate that the user has not followed the optimum path [34].

MUT also proposes the use of a particular graph type to visualize the testers' use pattern with aggregated data related to the user's number of mouse clicks and deviation from the optimal solution when solving a specific task that can help to see at which point the users have been lost or which are the nodes that generate more confusion.

This article does not present the used formulae as they are described in the bibliography and the CALMUT system is in charge of automatically making all the needed calculus.

3.2.4 | Diagnostic analysis and reporting

The objective of this stage is to identify the problems and strengths of the system. The results of the previous stage are analyzed with the aim of diagnosing the reasons behind the obtained results.

CALMUT generates a report that compiles the data obtained and the following process. The first part of the document includes the general introduction, information regarding the context, and the methodology used. Then it includes the gathered data and test results. The report finishes with the findings of the study and a description of the system aspects that need to be improved if any problem has been detected.

4 | CARRYING OUT THE USABILITY TEST OF ADESMUS

This section presents the usability test of adaptable evaluation system using multiple sources (AdESMuS) using MUT and CALMUT. AdESMuS is an educational system that combines aspects of project management, communication tools, and evaluation tools to provide solutions to the requirements generated during the process of development, supervision, and evaluation of Final Year Projects in Computer Engineering degrees [40]. Different user types can use AdESMuS: students, teachers, and administrative staff. In this section, we present the usability test for the perspective of the system related to the configuration aspects: defining evaluation criteria, assigning students to projects, and similar. Those tasks can be carried out either by teachers or by members of the administrative staff.

4.1 | Study design

The objective of the carried out study was to *confirm that users with no previous expertize in usability testing are able to conduct a usability test*. This main objective was divided into two research questions:

- RQ1: Does the use of CALMUT allow people with no previous background on usability to detect where the main usability problems of a system are?
- RQ2: Does the use of CALMUT help the usability testers interpret the statistical results?

Two lecturers without knowledge in usability testing carried out the system's usability test following the MUT methodology with the support of CALMUT.

To answer the research questions, the authors observed the usability testers during the study and conducted a semistructured interview (see Table 2) with them at the end [31].

Next, the AdESMuS system's usability test is described according to the MUT stages and the study results presented.

4.2 | Design and execution stage

A task was defined for each of the main four activities administrative staff can execute in AdESMuS (see Table 3).

The task execution order for each user was randomly defined. The estimated time together with the task optimal and alternative paths identified were defined in CALMUT (see Figure 6).

This section presents the usability test of the configuration tasks usually performed by teachers. As the usability test should be carried out by whom will be the final user of the system, 12 teachers were selected to take part in the test. As during the study, personal data would be collected, processed, stored, and approval of the Ethics Commission for Research and Teaching

TABLE 2 Interview questions.

Question
Did you detect any usability problem?
Does the interface of CALMUT facilitate the understanding of the statistical results obtained?
Was it easy to understand whether the results were correct or not?
How would you improve the reporting?
Did you miss something?
Would you use MUT and CALMUT to carry out other usability tests?

Abbreviation: MUT, Methodology for Usability Testing.

TABLE 3 Defined tasks.

Task code	Task description
Task0	Add criteria for evaluation
Task1	Assign students to supervisors
Task2	Management of projects
Task3	Assign an evaluation board

(CEID/IIEB) of the University of the Basque Country UPV/EHU was requested before recruiting the users. The approval was given with code M10-2016-181. The final sample was formed by a balanced set of males and females of different age ranges and with diverse experience levels in project management and evaluation who provided appropriate informed consent signing a document supervised also by the Ethics Commission.

At the time of the evaluation, AdESMuS was at an early stage of development and none of the data (Table 1) for the analysis was available. Therefore, a prototype was developed using Justinmind.[1] The prototype was prepared to allow instrumented remote evaluation and to register automatically the data required for the usability test in files with comma-separated values format that were after imported in CALMUT. One person involved in the development of AdESMuS but not in the MUT design created the prototype and provided it to the testers.

Once the planning phase was finished, the running phase began. The two main elements provided to the recruited users were the prototype's URL, a list indicating the particular order in which he or she should execute the tasks and the SUS questionnaire.

The prototype showed the user an interface with a list containing the four tasks to be executed, its description, and some general instructions. When the person carrying out the task considered that he or she understood the objective of the task, the execution could begin. From that moment, the user had total access to the prototype system and could freely navigate and interact with all the elements of the system she or he considered necessary to fulfill the proposed task.

Once the users finished the execution of all the tasks, they filled in the validated Spanish version of the SUS questionnaire [33].

4.3 | Descriptive and inferential analysis stage

CALMUT calculates all the metrics proposed by MUT (see Figure 7) and shows the results with indications to understand them in different screens so that testers can analyze data.

Figure 8 shows the tables obtained for the effectiveness results. In this case, CALMUT indicates (with a darker background) that the first task should be analyzed. Others have no marks, as its values are acceptable according to MUT.

For satisfaction, CALMUT shows the data obtained for the SUS questionnaire together with a guide indicating whether the values obtained are acceptable or not, as shown in Figure 9. If some of the values were not acceptable they would be marked on the screen.

Figure 10 shows the confidence intervals (calculated taking as a base the execution time of all participants) and maximum admissible time for each task. In this case, all the upper limits of the confidence intervals are below the maximum admissible execution time for each task.

Lostness is also used to test efficiency in MUT. CALMUT calculates the lostness metric for the sample and for the population and indicates whether those values are acceptable or not (see Figure 11).

In CALMUT, this information is accompanied by use patterns that help to see at which point the users have been lost or which are the interfaces that generate more confusion (see Figure 12). All the system graphic interfaces were alphabetically coded and each graph shows them as the nodes used to solve the task.

As shown in Figure 12 the navigation problems (red dashed arrows) to solve the task begin in the start node (denoted *C Start*) and continue until users reach the *L End* denoted node, through which they had to pass twice. Once it reached the *L End* node for the first time, users had no subsequent problems.

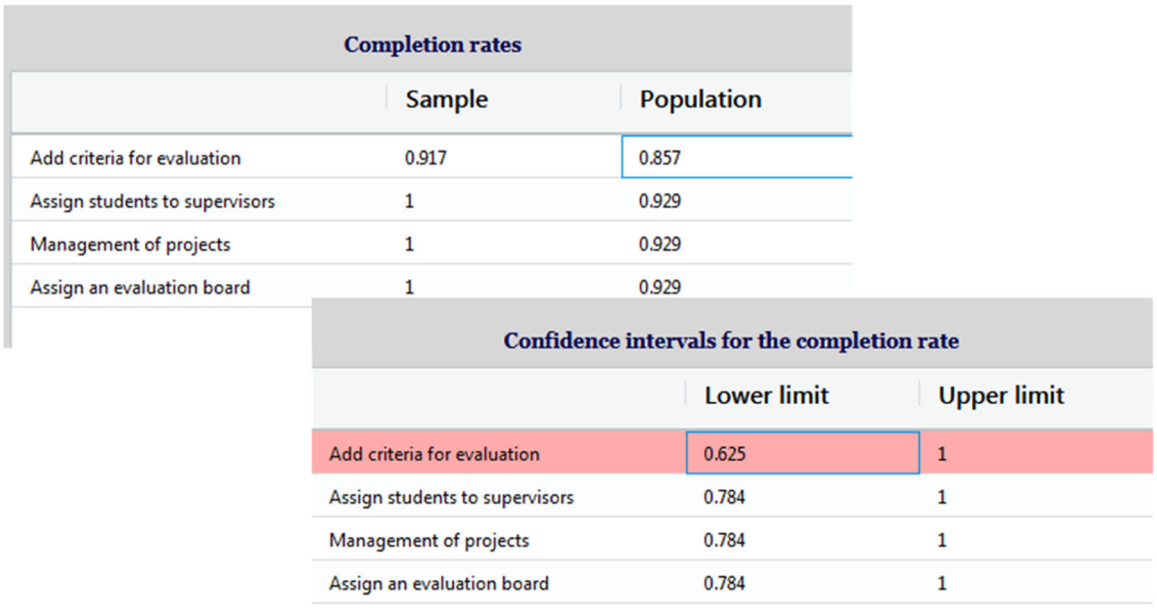
4.4 | Diagnostic analysis and reporting stage

The data and results to be included in the report are the contents of the CALMUT screens previously shown which are compiled in a final report when "General report" option (see Figure 7) is selected in CALMUT.

In the example presented, the results of the usability test have been quite good and each metric's acceptance criteria are met in most of the cases. Task0 is the only one that has thrown up several problems and that should be analyzed more in depth.

Testers indicated that seeing the use pattern graph for this task allowed them to see in more detail what was happening. From this graph, they could deduce where the problem with the interface was for this task.

[1] http://www.justinmind.com



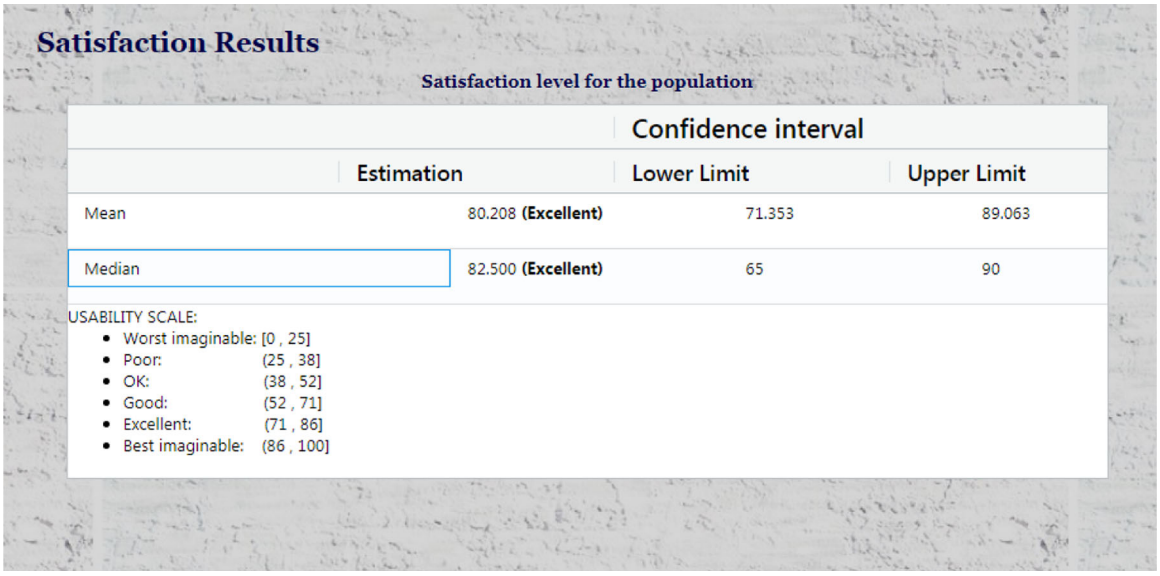
Completion rates

	Sample	Population
Add criteria for evaluation	0.917	0.857
Assign students to supervisors	1	0.929
Management of projects	1	0.929
Assign an evaluation board	1	0.929

Confidence intervals for the completion rate

	Lower limit	Upper limit
Add criteria for evaluation	0.625	1
Assign students to supervisors	0.784	1
Management of projects	0.784	1
Assign an evaluation board	0.784	1

FIGURE 8 Effectiveness data results interface.



Satisfaction Results

Satisfaction level for the population

	Estimation	Confidence interval	
		Lower Limit	Upper Limit
Mean	80.208 **(Excellent)**	71.353	89.063
Median	82.500 **(Excellent)**	65	90

USABILITY SCALE:
- Worst imaginable: [0 , 25]
- Poor: (25 , 38]
- OK: (38 , 52]
- Good: (52 , 71]
- Excellent: (71 , 86]
- Best imaginable: (86 , 100]

FIGURE 9 Satisfaction data results interface.

4.5 | Discussion

Lecturers with no previous experience in usability testing have been able to follow the MUT stages without a problem. They have correctly designed and executed the usability test and have been able to carry out the data analysis and the diagnostic analysis. The only moment where they needed support was in the running environment preparation step. As AdESMuS was in its early development stages, a functional prototype was needed to perform the whole usability test. This prototype was developed by the AdESMuS developers and provided to the testers to continue with the study.

According to the information gathered, testers were able to interpret the statistical results (RQ2). They indicated that the interpretation of the statistical results was easy because in the screens for each metric CALMUT shows which are the acceptance criteria and what they mean what makes possible to understand the problem without having to understand the statistics behind. CALMUT enriches this information showing for each task whether the obtained values meet the acceptance criteria or not. This facilitates identifying

Confidence intervals and maximum admissible time for each task

Task	Maximum time	Lower limit	Upper limit
Add criteria for evaluation	187.25	119.269	183.702
Assign students to supervisors	242.20	86.059	147.936
Management of projects	63.30	27.483	55.729
Assign an evaluation board	772.75	44.409	159.407

FIGURE 10 Confidence intervals and maximum admissible time for each task.

Confidence interval for the lostness

	Mean		Median	
	Lower Limit	Upper Limit	Lower Limit	Upper Limit
Add criteria for evaluation	0.038	0.202	0	0.270
Assign students to supervisors	0.017	0.198	0	0.290
Management of projects	0*	0.195	0	0
Assign an evaluation board	0*	0.176	0	0.350

*Negative values

ACCEPTANCE CRITERIA: Lostness value <= 0.4

FIGURE 11 Confidence intervals for the lostness.

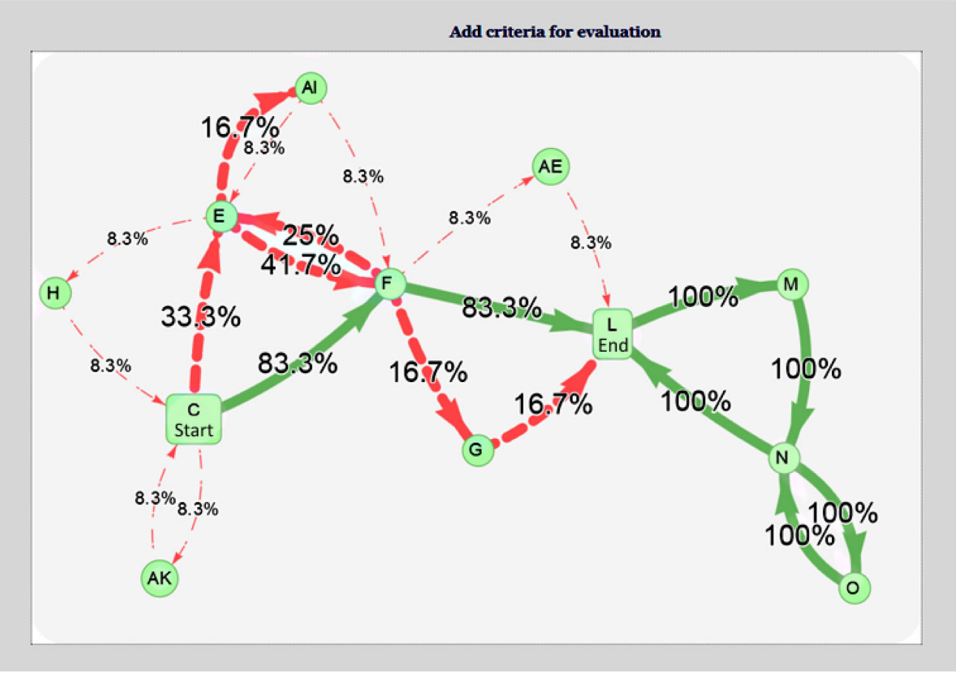


FIGURE 12 Lostness graphs.

those aspects that are generating problems and therefore detecting where the main usability problems of the system are (RQ1).

Knowing where the problem is, helps testers to center on the aspect that is generating usability problems and try to improve it. In this specific case, testers detected that in the interface there was a menu that was not clear enough and produced user lostness, so the menu was reorganized to reduce it. When the problem is not very clear, some kind of expert supervision will be required, however, in developers groups often this can be easily analyzed.

Before the lecturers executed the usability test, the authors of this article, as experts in usability testing, carried out an expert review of the system where some usability troubles were detected, but not solved. After, the novice-users used MUT and CALMUT to carry out a usability test. The results obtained by the lecturers using CALMUT were compared with the ones previously obtained by the experts. This comparison allowed us to analyze whether the results obtained by novice-users using MUT and CALMUT were similar to the ones obtained by a set of experts.

The comparison confirmed that the problems detected in the expert review and by the testers using MUT and CALMUT were the same.

Therefore, the carried out study allowed us to confirm the initial research questions. Users without experience in usability testing were able to interpret the statistical results (RQ2) and the use of MUT and CALMUT facilitates them detecting where the main usability problems of the system are and their probable causes (RQ1).

5 | CONCLUSIONS

This article has presented the MUT usability testing methodology and the CALMUT supporting tool, defined to help educational technology researchers, people who has to decide about using a software or not and learning designers to test the usability of systems. In comparison to systems analyzed in the related work, the proposed system and methodology combination allow the evaluation of both online and offline systems. MUT proposes three stages to tackle a usability test and uses the three usability indicators defined in the ISO 9241-11:2018: effectiveness, efficiency and satisfaction.

MUT guides analysts throughout the whole process of evaluating usability. The first stage is centered on the design and execution of the test.

The second stage is devoted to the descriptive and inferential analysis of the data collected during the running phase. This stage is the main contribution of the methodology, as it states the metrics to use, the data that must be collected to calculate them and a guide to interpret the obtained results. This guidance is especially useful for novice usability practitioners. The metrics have been selected carrying out a bibliographical study. MUT also uses graphs for the use patterns which makes it easier to detect the location of the more problematic sections of the system being tested, and to take corrective measures accordingly.

Finally, the last stage of MUT addresses the diagnostic analysis to determine the strengths and weaknesses of the analyzed system and the reporting of the study.

However, making the required calculus and understanding what the results mean is not easy and we have presented CALMUT, a system that facilitates carrying out the usability test following MUT. CALMUT guides the testers during the usability test, automates the calculus, and reports the results including remarks about the acceptance criteria to help understand the test results and facilitate the diagnosis.

This step-by-step guide and the reports including information about the acceptance criteria for the results make CALMUT more adequate for novice users than the systems and calculators presented in Section 2.

The proposed methodology and its calculator have successfully allowed lecturers with no background in usability testing to carry out a usability test for the prototype of the AdESMuS educational system. The obtained results and the performed analysis allowed analysts to detect the main usability problems of the system's prototype and to take remediation actions for the final version of the system.

The use of MUT and CALMUT also facilitates lecturers or academic administrators deciding about the use of an educational software tool via the analysis of its usability and the detection of the problems its use could generate.

ACKNOWLEDGMENTS

This work was supported by the Department of Education, Universities and Research of the Basque Government under Grant ADIAN, IT-1437-22. Open Access funding provided by University of the Basque Country UPV/EHU.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

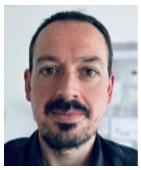
Mikel Villamañe http://orcid.org/0000-0002-4450-1056
Ainhoa Alvarez http://orcid.org/0000-0003-0735-5958

REFERENCES

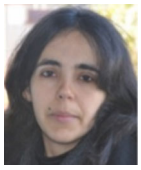
1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, *Standards for educational and psychological testing*, American Educational Research Association, Washington, DC, 2014.
2. I. D. Bleecker and R. Okoroji, *Remote Usability Testing: Actionable insights in user behavior across geographies and time zones*, Packt Publishing, Birmingham, UK, 2018.
3. J. Brooke, *SUS: A retrospective*, J. Usability Stud. **8** (2013), 29–40.
4. A. Bruun and J. Stage, *New approaches to usability evaluation in software development: Barefoot and crowdsourcing*, J. Syst. Softw. **105** (2015), 40–53. https://doi.org/10.1016/j.jss.2015.03.043
5. P. Buono, D. Caivano, M. F. Costabile, G. Desolda, and R. Lanzilotti, *Towards the detection of UX smells: The support of visualizations*, IEEE Access **8** (2020), 6901–6914. https://doi.org/10.1109/ACCESS.2019.2961768
6. L. Cayola and J. A. Macías, *Systematic guidance on usability methods in user-centered software development*, Inf. Softw. Technol. **97** (2018), 163–175. https://doi.org/10.1016/j.infsof.2018.01.010
7. S. Dimitrijevic and V. Devedzic, Usability evaluation in selecting educational technology, *proceedings of 11th Conference of Information Technology and Development of Education (ITRO 2020)*, University of Novi Sad, Serbia, 2020.
8. S. M. Drucker, A. Glatzer, S. De Mar, and C. Wong, *SmartSkip: consumer level browsing and skipping of digital video content*, *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, ACM Press, Minneapolis, Minnesota, USA, 2002, p. 219.
9. K. Finstad, *The usability metric for user experience*, Interact. Comput. **22** (2010), 323–327. https://doi.org/10.1016/j.intcom.2010.04.004
10. C. Gonzalez, *Student usability in educational software and games: improving experiences*, Information Science Reference, Hershey PA, 2013.
11. T. Granollers and J. Lorés, *Incorporation of users in the Evaluation of Usability by Cognitive Walkthrough*, *HCI related papers of Interacción 2004* (R. Navarro-Prieto and J. L. Vidal, eds) Kluwer Academic Publishers, Dordrecht, 2006, pp. 243–255.
12. L. A. Hasan and K. T. Al-Sarayreh (2015) An Integrated Measurement Model for Evaluating Usability Attributes. In: Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication. ACM, New York, NY, USA, pp. 94:1-94:6
13. D. Hinderer and J. Nielsen, *How to recruit participants for usability studies*, Fremont, USA, 2003.
14. K. Hornbæk, *Current practice in measuring usability: Challenges to usability studies and research*, Int. J. Hum. Comput. Stud. **64** (2006), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002
15. W. Hwang and G. Salvendy, *Number of people required for usability evaluation: The 10±2 rule*, Commun. ACM **53** (2010), 130–133. https://doi.org/10.1145/1735223.1735255
16. ISO, *Ergonomics of human-system interaction—Part 11: Usability: Definitions and concepts*, International Organization for Standardization, Geneva, Switzerland, 2018.
17. M. Y. Ivory and M. A. Hearst, *The state of the art in automating usability evaluation of user interfaces*, ACM Comput. Surv. **33** (2001), 470–516.
18. D. Karahoca, *Meta-Cognitive tool development for history teaching: investigating how software usability affects student achievements*, J. Univers. Comput. Sci. **19** (2013), 619–638. https://doi.org/10.3217/jucs-019-05-0619
19. P. Kortum, C. Z. Acemyan, and F. L. Oswald, *Is it time to go positive? Assessing the positively worded system usability scale (SUS)*, Hum. Factors J. Hum. Factors Ergon. Soc. **63** (2020), 987–998. https://doi.org/10.1177/0018720819881556
20. J. R. Lewis, *Sample sizes for usability tests: Mostly math, not magic*, Interactions **13** (2006), 29–33. https://doi.org/10.1145/1167948.1167973
21. J. R. Lewis, *Measuring perceived usability: The CSUQ, SUS, and UMUX*, Int. J. Human–Comput. Interact. **34** (2018), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805
22. Loop11. https://www.loop11.com/, 2022.
23. B. Losada, M. Urretavizcaya, J-M López-Gil, and I. Fernández-Castro (2012) Combining InterMod agile methodology with usability engineering in a mobile application development. In: Actas de International Conference on Interacción Persona-Ordenador. ACM Press, Elche, p 39:1-39:8.
24. J. Lu, M. Schmidt, M. Lee, and R. Huang, *Usability research in educational technology: A state-of-the-art systematic review*, Educ. Technol. Res. Dev. **70** (2022), 1951–1992. https://doi.org/10.1007/s11423-022-10152-6
25. I. S. MacKenzie, T. Kauppinen, and M. Silfverberg (2001) Accuracy measures for evaluating computer pointing devices. In: Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01. ACM Press, Seattle, Washington, United States, pp 9–16.
26. D. Marshall, J. C. Foster, and M. A. Jack, *User performance and attitude towards schemes for alphanumeric data entry using restricted input devices*, Behav. Inf. Technol. **20** (2001), 167–188. https://doi.org/10.1080/01449290110048007
27. MeasuringU, https://measuringu.com/calc/, 2022.
28. Nielsen Norman Group Nielsen Norman Group: UX Training, Consulting, & Research. In: Nielsen Norman Group. https://www.nngroup.com/, 2020.
29. D. Quiñones, C. Rusu, and V. Rusu, *A methodology to develop usability/user experience heuristics*, Comput. Stand. Interfaces **59** (2018), 109–129. https://doi.org/10.1016/j.csi.2018.03.002
30. J. Rubin and D. Chisnell, *Handbook of usability testing: How to plan, design, and conduct effective tests*, 2nd ed., Wiley Pub, Indianapolis, IN, 2008.
31. P. Runeson and M. Höst, *Guidelines for conducting and reporting case study research in software engineering*, Empir. Softw. Eng. **14** (2008), 131–164. https://doi.org/10.1007/s10664-008-9102-8
32. A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda, *Usability measurement and metrics: A consolidated model*, Softw. Qual. J. **14** (2006), 159–178. https://doi.org/10.1007/s11219-006-7600-8
33. M. D. R. Sevilla-Gonzalez, L. Moreno Loaeza, L. S. Lazaro-Carrera, B. Bourguet Ramirez, A. Vázquez Rodríguez, M. L. Peralta-Pedrero, and P. Almeda-Valdes, *Spanish version of the system usability scale for the assessment of electronic tools: Development and validation*, JMIR Hum. Factors **7** (2020), e21161. https://doi.org/10.2196/21161
34. P. A. Smith, *Towards a practical measure of hypertext usability*, Interact. Comput. **8** (1996), 365–381.
35. D. S. Tan, G. G. Robertson, and M. Czerwinski (2001) Exploring 3D navigation: combining speed-coupled flying with orbiting. In: Proceedings of the SIGCHI conference on Human factors in computing systems—CHI '01. ACM Press, Seattle, Washington, United States, pp 418–425.

36. T. Tullis and B. Albert, *Measuring the user experience: collecting, analyzing, and presenting usability metrics*, Second edition., Elsevier/Morgan Kaufmann, Amsterdam; Boston, 2013.
37. T. Tullis and J. N. Stetson (2004) A comparison of questionnaires for assessing website usability. In: Proceedings of the 13th UPA Conference. Minneapolis. https://api.semanticscholar.org/CorpusID:9670323
38. usabiliTEST, Usability testing & Information Architecture. https://www.usabilitest.com/, 2020.
39. Usabilityscale, https://usabilityscale.com/, 2022.
40. M. Villamañe, B. Ferrero, A. Álvarez, M. Larrañaga, A. Arruarte, and J. A. Elorriaga (2014) Dealing with common problems in engineering degrees' Final Year Projects. In: Actas de IEEE Frontiers in Education Conference. IEEE Computer Society, Madrid, pp 2663–2670.
41. P. Vlachogianni and N. Tselios, *Perceived usability evaluation of educational technology using the system usability scale (SUS): A systematic review*, J. Res. Technol. Educ. **54** (2022), 392–409. https://doi.org/10.1080/15391523.2020.1867938
42. K. E. Wiegers and J. Beatty, *Software requirements*, Third edition, Microsoft Press, s division of Microsoft Corporation, Redmond, Washington, 2013.
43. P. Zaphiris and S. Kurniawan, *Human computer interaction research in web design and evaluation*, Idea Group Inc (IGI), Hershey, PA. USA, 2007.

AUTHOR BIOGRAPHIES



Mikel Villamañe obtained the PhD degree in computer science from the University of the Basque Country UPV/EHU in 2017 where he has been a faculty member since 2003. His research, developed within the GaLan and Adian research groups, focuses on the field of educational computing, particularly on the use of technology to enhance learning and assessment processes. He has published over 30 works on topics such as learning analytics, complex assessment scenarios or active and cooperative methodologies.



Ainhoa Alvarez received her PhD degree in computer science from the University of the Basque Country UPV/EHU in 2010. Since 2001, she has been a faculty member in the Department of Computer Languages and Systems at the University of the Basque Country UPV/EHU. She develops her research activities within the GaLan and Adian research groups, focusing on the use of technology to improve learning environments, especially assessment processes. She is the author of more than 50 publications in areas such as computer-aided education, learning analytics, and intelligent tutoring systems.

How to cite this article: M. Villamañe and A. Alvarez, *Facilitating and automating usability westing of educational technologies*, Comput. Appl. Eng. Educ. **32**, (2024), e22725. https://doi.org/10.1002/cae.22725