COLLABORATIVE HEURISTIC EVALUATION CONDUCTED BY A GROUP OF EXPERT AND NOVICE EVALUATORS: A case study for a start-up company at Brazil

Data de entrega dos originais à redação em: 14/03/2016, e recebido para diagramação em: 30/11/2016.

Francisco Fabiano Neves[1]
Bruno Felipe Leal Delfino[2]
André de Lima Salgado[3]
Ana Elisa de Oliveira Siena[4]
Silvana Maria Affonso de Lara[5]

Neste estudo avaliamos a validade da estratégia adotada por uma empresa start-up para a avaliação de usabilidade, adaptada ao contexto econômico da empresa, no desenvolvimento de um aplicativo móvel que possibilita a melhoria da interação em tempo real de alunos/ouvintes com o professor/palestrante dentro do contexto de uma apresentação expositiva. Os resultados obtidos em cada uma das avaliações conduzidas pela empresa foram comparados a partir de métricas identificadas na literatura, e mostraram a validade da estratégia adotada pela empresa.
Palavras-chave: Avaliação Heurística Colaborativa, Usabilidade, Aplicação Móvel, Avaliador Novato, Avaliador Especialista.

This study analyses the validity of the strategy assumed by a start-up company, considering their economical conditions, to conduct periodic usability evaluations in the development of a mobile application designed to enable real-time interaction students/listeners with the teacher / lecturer within the context of an exhibition presentation. The results obtained from each of the evaluations conducted by the company were analyzed using metrics from the literature, and showed the validity of the strategy assumed by the company.
Keywords: Collaborative Heuristic Evaluation, Usability, Mobile Application, Novice Evaluator, Expert Evaluator.

1 INTRODUCTION

The development of educational software to support the learning process involves the definition of the pedagogical conception of those who are involved in its development and implementation. We purpose the development of *Painel Educativo* (Pedagogical Panel). The *Painel Educativo* consists on a set of web and mobile application capable of supporting the interaction between student and teachers in real time, independently of the pedagogical strategy assumed by those who will use it. In a wider view, our goal is to develop a platform for communication between audience

[1]Graduando em Análise e Desenvolvimento de Sistemas no IFSP – Campus São Carlos. E-mail: <chicofab@gmail.com>

[2]Graduando em Análise e Desenvolvimento de Sistemas no IFSP – Campus São Carlos. E-mail: <bruno.delfino1995@gmail.com>

[3]Mestrando em Ciências de Computação e Matemática Computacional pela Universidade de São Paulo. E-mail: <andrelima.salgado@gmail.com>

[4]Graduada em Engenharia de Computação pela Universidade de São Paulo. E-mail: <ana.siena@sienaidea.com.br>

[5]Doutora em Ciências de Computação e Matemática Computacional pela Universidade de São Paulo Professora da Área de Informática no Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – IFSP – Campus São Carlos. E-mail: silaffonso@ifsp.edu.br

and instructor, enabling a new dynamic of instant feedback in real time. This way, participants in the audience that could have any type of shyness can now expose your doubt through an anonymous manner. This possibility will enhance their learning.
Painel Educativo consist in a combination of a mobile application, which can be installed in the mobile devices of the participants in the audience, and a monitor display that will be located at the place of the event. The audience will be able to send any kind of feedback, as questions and doubts, using the app at their mobile device.
That *Painel Educativo* will make possible to send feedback using an anonymous feature, according to the users' preference. All feedback will be shown at a monitor display in a sort of priority, according to the popularity of the feedback. Participants will have an option at the application to support any specific feedback, in case they want to increase the popularity of it.
This project belongs to a partnership between IFSP (Instituto Federal de São Paulo, at São Carlos - Brazil) and Siena Idea, a Start-Up company at São Carlos – Brazil. It receives financial support from the Brazilian Government (CNPq). The *Painel Educativo* will be available to any institution aiming to receive better feedback from their audiences and to improve their process of teaching.
The following sections of this paper present an introduction for the case study, the description of methods and materials used, the results and discussion, and the conclusions obtained.

2 THE CASE STUDY

This case study is aimed to investigate the better strategy to support the development of *Painel Educativo* as a usable application. Usability is recognized as being an important support for software quality (ISO 25010). To develop a usable application, it is important to periodically apply Usability Evaluation Method (UEM) in order to be aware of the current usability of the product (DIX et al., 2003; ROGERS et al., 2011).
Test with users are capable of finding the problems users really care about (PETRIE and POWER, 2012). However, applying frequent tests with users are expensive for the economical scope of Siena Idea; and applying discount usability evaluations as heuristic evaluation is still expensive because of the dependence on usability expert, who are rare to find in the Brazilian market. For this reason, the company had to apply a heuristic evaluation counting on novice evaluators to conduct it together with only 2 experts. The method applied was the Collaborative Heuristic Evaluation – CHE (BUYKX, 2009; PETRIE and BUYKX, 2010). The goal of this study is to investigate the validity of the heuristic evaluation in this case, in order to support Siena Idea in a decision on how to conduct periodic UEMs.

3 METHODS AND MATERIALS

3.1 EXPERIMENTAL DESIGN

The usability of a prototype of the *Painel Educativo* application was evaluated using two different UEMs. First, test with users were carried out with the voluntary participation of 9 users. Later, 2 usability experts and 3 usability novices carried out a CHE.
Both UEMs resulted in different sets of usability issues. The purpose of this study was to identify whether the development team of *Painel Educativo* could apply CHE using group of both expert and novice evaluators in the cycle of usability evaluation, because of the high cost of conducting periodically test with users.

We calculated the overlap of problems found by CHE and test with users to analyze the pros and cons of applying each one of these UEMs in our context of development. Both evaluations followed the same list of predefined tasks.

3.2 THE PROTOTYPE OF *PAINEL EDUCATIVO*

The prototype of *Painel Educativo* regards only the audience as a user profile for instance. This version does not contemplate the lecturer as a user. Specifically, this version considers the students and professors of IFSP São Carlos as audience and potential users. It was developed using the tool JustinMind[1].

The version of the prototype evaluated implements features as: login and logout; list events according to the period of its occurrence; access to the feedbacks of a specific event; comment, like and dislike a specific feedback; list the participants of a specific event; and download the content of a specific event. Prototype screens are presented in the Figure 1.

Figure 1: Prototype screens of *Painel Educativo.*

3.3 METHOD FOR TEST WITH USERS

9 participants took part in the test with users voluntarily. 3 participants (3 men) were university professors. 6 participants (2 women and 4 men) were university students. All them had previous experience with mobile applications.

The tests were carried out by Siena Idea, as part of the development process of *Painel Educativo*. The Siena Idea needed 3 working days to conduct all the tests, due to the availability of the participants. Because of available time to schedule the tests, it was not possible to ensure that all

1 http://www.justinmind.com/

participants used the same hardware during the evaluations, what implies on a limitation of this study. For this reason, 2 tests were conducted using a computer to access the prototype and the other 7 were conducted using a mobile phone.
The test sessions lasted from 5 to 20 minutes, depending on the participant. A moderator was responsible for explaining the reasons of the tests to all participants before each test. In addition, the moderator was told to remember the participants to follow the Think-Aloud technique and speak out loud his/her thoughts during the interaction. The moderator noted all the feedback received from the verbalizations of participants' thoughts.

3.4 METHOD FOR *CHE*

6 participants took part at the CHE session. 5 participated as evaluators and 1 as the scribe. Among the evaluators: 2 were usability specialists with more than 3 years of research experience in usability related area and several previous participations in usability evaluation; and 3 were novice in usability area that work for Siena Idea. The scribe was also a worker from Siena Idea.
The CHE session took part inside the workplace of Siena Idea. A computer with a wide monitor display accessing the prototype was available for the evaluation. Each evaluator was provided with a severity rating form to rate the severity of each problem. The evaluators used the traditional heuristics of Nielsen, because of its wide adoption in the literature of mobile usability (SALGADO and FREIRE, 2014). The scribe used her own computer to note the usability issues using worksheet software. The CHE session lasted 50 minutes.

3.5 DATA ANALYSIS

The procedure of data analysis was carried out using the criteria of Gray and Salzman (1998) and Hartson et al. (2003) for assessment of different UEMs:

- *False Alarms*: Issues reported by CHE that were not reported by the test with users;
- *Misses*: Issues reported by the test with users that were not reported by the CHE;
- *Hits*: Issues reported by the test with users that were reported by the CHE.

Matching the similarity among different problems reports is a difficult task and the literature does not show a consensus on how to conduct it (HORNBÆk, 2010). To find the number of each one of these terms, we used the two following criteria for matching problems from reports of different UEMs: strict matching criteria and relaxed matching criteria (BUYKX, 2009; PETRIE and BUYKX, 2010; PETRIE and POWER, 2012).
In the strict matching criteria, multiple problems are identified as similar only if they refer to the same design element and to the same problem, at the same level of abstraction. In the other hand, the relaxed matching criteria consider multiple problems as similar if they refer to the same design element, or to the same problem, considering different levels of abstractions and considering cases where the same underlying problem is referred (BUYKX, 2009; PETRIE and BUYKX, 2010).
In addition, we created the term *Positive Alarms*. Our previous experience shows that, using the relaxed matching criteria, it is possible that one issue reported by a specific UEM *hits* more than one issue at the set of issues of the UEM considered as base for the comparison. For this reason, in this study *Positive Alarms* are: issues reported by the CHE that were reported by the test with users.

4 RESULTS AND DISCUSSION

Users identified a 7 distinct usability problems during the tests. Only 1 problem was identified by multiple users. The other 6 problems were not identified by more than one user. 2 users did not find any difficulty using the prototype, and they mentioned no problem. A total of 13 distinct usability problems were reported by the CHE session.

Results of the matching process

Considering the strict matching criteria, 23% of these problems - 3 problems - were *Positive Alarms* and 77% - 10 problems - were *False Alarms*. Analyzing with the strict matching criteria, the percentage of usability problems reported during the test with users that were missed (percentage of *Misses*) by the evaluators during the CHE was 57%. In addition, the percentage of usability problems reported during the test with users that were reported by evaluators during the CHE session (percentage of *Hits*) was 43%. These results show that 43% of the problems that users really care about were identified in this case by the CHE session, based on rigorous comparisons of similarity among problems.
Considering that even experts conducting CHE face difficulties to find all problems users really care about (HUANG, 2012; PETRIE and POWER, 2012), and the results of Othman et al. (2014) were novices covered from 30% to 35% of the problems listed by experts, finding 43% of the problems that users really care about may indicate that the CHE conducted by Siena Idea was a valuable strategy to adapt their process of developing a usable application. This study only contemplates one application and considers a smaller sample of evaluators, for this reason we suggest as future studies to deeper investigate how generalizable are these results.
Using the relaxed matching criteria, 69% of the usability problems - 9 problems - reported by the CHE session were identified as *False Alarms*. In addition, 31% of all usability problems - 4 problems - reported by the CHE session were *Positive Alarms*. 29% of the usability problems - 2 problems - identified during test with users were missed by the evaluators during the CHE session; and 71% of the usability problems - 5 problems - identified during the test with users were identified by the evaluators during the CHE session.
71% of *Hits* is highly satisfactory for the needs of Siena Idea. We understand that in only 50 minutes of CHE, a less expensive group of usability evaluators could find the major part of the problems users would care about. This may indicate that Siena Idea can increase the periodicity of usability evaluation during the development of *Painel Educativo* by having more CHE sessions, according to their economic strategies.

Difference between severity rating by expert and by novice evaluators

We calculated the mean severity of ratings made by expert and ratings made by the novices for each one of the problems listed at the report of the CHE session. In sequence, we generated two lists of mean severity for each problem: the expert list and the novice list. We used the T-Test to compare both lists. No significant difference was found between the expert list and the novice list after a ($p \leq 0.05$; $t = -0.80$). Huang (2012) obtained similar results, no significant difference among the severity ratings of expert and novice evaluators during CHE, but their study conducted a remote CHE (rCHE) and evaluated more applications using a different study design.

These results can indicate that the rating of severities suffered no impact with the presence of novice evaluators in the CHE session. However, our sample is too small to affirm this and future works can investigate this difference with larger samples, and comparing equivalent number of expert and novice. This result is in accordance to the comparisons made by, but their study conducted another variation of the CHE.

5 CONCLUSIONS

We conclude that the strategy assumed by Siena Idea was valid and further CHE with the same structure can be applied in the development process to help the organization to conduct periodic UEMs. As the company cannot apply heuristic evaluation with groups full of expert evaluators, we recommend applying the configurations that supply the presence of many experts by the presence of novice evaluators in order to enhance the periodicity of usability evaluation, the results of this study showed that this configuration provided qualified results.
This study was limited for one mobile application, and for a single group of evaluators. However, the results and conclusions provided good insights to the literature. We suggest to future studies to investigate how generalizable these results are, as if these results and conclusions are applicable for any company of the same characteristics and also if larger companies could apply it without negative implications.

ACKNOWLEDGEMENTS

We thank CNPq, Siena Idea, CAPES their kindly support to this research. We also thank all the participants that voluntarily contributed in this study.

REFERENCES

BUYKX, Lucy. **Improving heuristic evaluation through collaborative working**. 2009. 72 f. Dissertation (MSc in Computer Science) – Department of Computer Science, University of York, York, UK. 2009.

DIX, Alan; FINLAY, Janet Finlay; ABOWD, Gregory D.; BEALE, Russell. **Human Computer Interaction**. Pearson Education Limited. 3. ed. 2003.

ISO/IEC. Software Engineering -- **Software product Quality Requirements and Evaluation (SQuaRE)** -- System and software quality models. 2011.

GRAY, Wayne D.; SALZMAN, Marilyn C. **Damaged merchandise? A review of experiments that compare usability evaluation methods**. Human–Computer Interaction 13, n. 3. p. 203-261. 1998.

HARTSON, H. Rex; ANDRE, Terence S.; WILLIGES, Robert C. **Criteria for evaluating usability evaluation methods.** International Journal of Human-Computer Interaction 15. n. 1. p.145-181. 2003.

HORNÆk, Kasper. **Dogmas in the assessment of usability evaluation methods**. Behaviour & Information Technology 29. n. 1. p. 97-111. 2010.

HUANG, Bo. A **Comparison of Remote Collaborative Heuristic Evaluation by Novices and Experts with User-based Evaluation.** 2012. 128 f. Dissertation (MSc in Computer Science) – Department of Computer Science, University of York, York, UK. 2012.

OTHMAN, Mohd Kamal; MAHUDIN, Fadhullah; AHAGUK, Cassandra Henry; RAHMAN, Abdul; FARHAN Muhd. **Mobile guide technologies (smartphone apps): Collaborative Heuristic Evaluation (CHE) with expert and novice users.** In *User Science and Engineering (i-USEr), 2014 3rd International Conference on*. IEEE. p. 232-236. 2014.

PETRIE, Helen; BUYKX, Lucy. **Collaborative Heuristic Evaluation: improving the effectiveness of heuristic evaluation.** Proceedings of UPA 2010 International Conference. Omnipress. Available at: http://upa. omnibooksonline. com/index. htm. 2010.

PETRIE, Helen; POWER, Christopher. **What do users really care about?** Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12, 2012.

ROGERS, Yvonne; SHARP, Helen; PREECE, Jenny. **Interaction design: beyond human-computer interaction.** John Wiley & Sons. 3. ed. 2011.

SALGADO, André L.; FREIRE, André P. **Heuristic evaluation of mobile usability: A mapping study**. In 16th International Conference, HCI International 2014, Heraklion, Crete, Greece. Springer. p. 178-188. 2014.