World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:5, 2017

The Correlation between Users' Star Rating and Usability on Mobile Applications

Abdulmohsen A. AlBesher, Richard T. Stone

***Abstract*—**Star rating for mobile applications is a very useful way to differentiate between the best and worst rated applications. However, the question is whether the rating reflects the level of usability or not. The aim of this paper is to find out if the user' star ratings on mobile apps correlate with the usability of those apps. Thus, we tested three mobile apps, which have different star ratings: low, medium, and high. Participating in the study, 15 mobile phone users were asked to do one single task for each of the three tested apps. After each task, the participant evaluated the app by answering a survey based on the System Usability Scale (SUS). The results found that there is no major correlation between the star rating and the usability. However, it was found that the task completion time and the numbers of errors that may happen while completing the task were significantly correlated to the usability.

***Keywords*—** Mobile applications, SUS, Star rating, Usability

I. Introduction

OVER the recent years, mobile devices have been used for various services such as mobile banking, mobile government, and mobile learning. Using mobile services has become more popular than e-services for many reasons.

One reason is that the number of mobile phones is greater than the number of personal computers. The need for mobile services has grown each year [8], and as a result, developing mobile applications were encouraged.

Mobile applications allow users to provide star ratings and reviews. Normally, new users look at the number of people who have tested an application and the star rating that they gave before installing it. The star rating builds a first impression about an application for the new user.

The question is whether this type of rating correlates with the usability of the app or not. The aim of this paper was to find out if there is any correlation between the star rating and the usability of mobile applications.

There was no clear study to show the correlation between the users' star rating and usability on mobile apps. Thus, we believe that our study will add useful knowledge in the field of mobile Human Computer Interaction (HCI).

This paper is organized as follows: First the methodologies that had been used followed by the detailed description of the experimental design. Consequently, details about the participants, the used materials and the experimental setting are presented. Then, the data obtained were discussed and analyzed, and finally, the limitations and future works are discussed.

Abdulmohsen Albesher from King Faisal University, Saudi Arabia (e-mail: albesherkfu@gmail.com)

II. Methodology

The focus of this study was to determine if a users' star rating for a mobile application correlated with the usability of the reviewed application. Thus, we searched for three mobile applications that have different star ratings: low, medium, and high. Table I presents the details of those apps.

Following this, the researchers defined a random task for each application. The selected tasks could be completed, but also involved a fair amount of users' engagement.

TABLE I
The Three Tested Mobile Applications.

App #	App Name	# of Users	Star Rating
A	MOFA*	236	4.8
B	Saudi Post	211	3.5
C	MOHE**	97	2.5

* MOFA Ministry of Foreign Affairs ** MOHE Ministry of Higher Education

Prior to running the user trials, we determined the errors that may occur during the completion of the three tasks. We defined these errors as follows:

1. Typing information in the wrong field.
2. Failing to log in.
3. Going back to the main menu.
4. Writing the email or the cell phone number in the Arabic language.
5. Clicking on the wrong icon.
6. Programming errors.

We then developed an experiment that would require participants to perform a specific single task for each of the three selected applications. During each task, the human application interactions were filmed and both the time and error rate were recorded. After each task, the participants were asked to answer the SUS questionnaire. SUS was selected because it is simple, short (10 items) and has been found to be remarkably robust across a number of studies [1]-[4], and [7].

Later, the score of each questionnaire was collected based on pre-defined criteria. Questions 1, 5, 7, and 9 were scaled from 1 to 5, while the remaining questions were scaled from 5 to1. The score scale was not shown to the users in order to focus on the meaning of the question and not on the score.

III. Experiment

A. Hardware and Software Selection

An Android operating system was used instead of IOS (iPhone OS), because in Android, those applications have more users than IOS. For example, the number of users for

application A in Android was 229 users, while in IOS the number was 14 users.

Nexus-5 phone was used for the experiment since one of the author's has this phone and it is a very new phone. The participants were required to use this phone only to ensure that all participants used a device that had the same processing speed and memory size.

Applications that are designed to one group of users were used in order to make the reliance on the rating given to the three applications more accurate. Furthermore, government applications were selected since the researchers believe that no one would be interested to change the rating of those applications.

Fig. 1 Shows the tested phone on the computer's screen.

B. Experimental Procedure

The experiment was done in a private room in the ISU library. Each participant did the experiment individually. Prior to starting the experiment, the participant had been informed that his purpose was to accomplish three tasks and after each task, he would fill-in a questionnaire to evaluate the task.

A researcher explained a specific script for each participant. It was about how to use the phone in terms of how to go back and forward, type and find special characters, change the language, and locate the tested applications. Furthermore, a researcher notified every participant that each task would be monitored and recorded.

The authors did not ask the participants to work on app A, B then C. However, we followed one strategy that is based on the Probability Law, and the three apps were organized into six sets (ABC, ACB, BAC, BCA, CAB, or CBA). As a result, the first participant did ABC and then the second one did ACB and so on. When the cycle of the six options ended, the cycle was repeated and so on. The aim of following this strategy was to ensure that no bias would enter into the experimental outcomes as a result of order effect.

C. Participants

The number of participants was 15 and they were all males. All of participants were undergraduate and graduate students, attending Iowa State University in the US. They were from various departments in the school. All of them volunteered for the experiment and they were interested to participate. Most of them had used some mobile government applications before.

D. Experimental Setup

We connected the Nexus-5 phone to a computer in order to monitor and record the phone's screen, as shown in Fig. 1. For the purpose of analysis it was necessary to catch the errors and comeback to the recording any time in the future. The connection occurs when the two devices are connected on the Internet via an IP address. For the recording, QuickTime Player was used.

IV. Results

The collected data were organized and analyzed using Excel 2013 and SPSS v22.0.

One of the surprising results is that the usability score of application C was the highest although its star rating was the lowest, as shown in Fig. 2. The participants faced zero or one error while performing the given task for application C. As a result, the error average for all the participants was 0.2, as shown in Fig. 3. Furthermore, we noticed that all participants were able to complete the task for application C. Also, the time taken to complete the task for each one of them was very close to the time that it took the expert to complete the task.

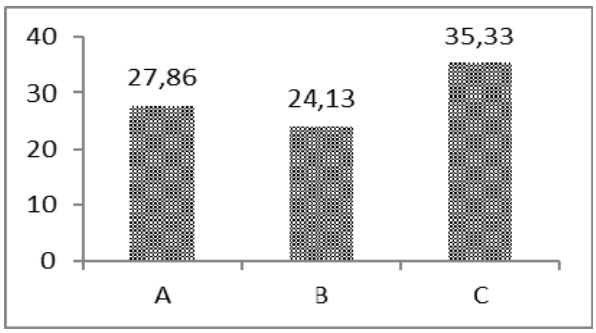

Fig. 2 Comparison of the usability score average.

On the other hand, applications A and B scored the lowest for usability even though they were given the highest star rating. Participants faced a considerable number of errors while performing the given tasks for applications A and B. The error averages were 5.6 and 6.2, respectively, which are significantly greater than the error average of the assigned task for application C. For most of the participants, the time taken to complete each task in A and B was notably much longer than the time that took the expert to complete the task.

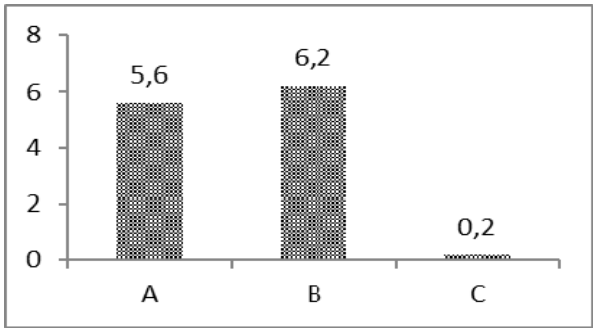

Fig. 3 Error average

Based on a random task for each application, it was found that the star rating of an application does not correlate to the overall usability of that application. However, the time taken to complete a task and the number of errors that may occur while performing the task, were significantly correlated to the usability.

In fact, all of the factors (usability rating (SUS), Time, Number of Errors and Percentage completion) correlated with an R of 0.4358 or higher. This strongly indicates that user experience by task is a confounded function of all of these factors. Add to this that no correlation existed in regards to the star ratings of the application and it would appear that initial task experience does not correlate to the star ratings of motivated and long term application users.

V. Discussion

During the time of the experiment, it was noticed that the more errors the user faced or the more time he spent in a task, the more frustrated he became. We believe that this frustration had a major effect on the overall score of usability for each application. In fact, both Raptis et al. [7] and Mendoza et al. [5] found that task completion times were considerably correlated with the participants' SUS ratings in the results of their experiment.

The researchers propose two reasons that can explain why one application got a high star rating, while it did not get a good usability score. The first reason could be that the users liked the interface but they did not try to accomplish specific tasks. The second reason is that the users had experienced different tasks that were done well. Our view that star ratings are often influenced by users who are required or motivated to use a piece of software beyond the initial experience is supported by the idea that the majority of software users frame their impression of software from their earliest experiences with it [6]. For an application to succeed in a more universal appeal it must create a good initial impression. Hence, a star rating may not be a good indicator of software performance across potential users, but rather in informs on the experience of users who are otherwise motivated to use a given piece of software.

VI. Limitation

The study was limited by the very small number of Saudis residing in Ames, Iowa, and thus, we could not have more than 15 participants. Only Saudis were selected to participate in the study because the examined applications were designed for use by Saudi citizens and residents.

The researchers also struggled to find three different mobile government applications to fit the study criteria of having a low, medium, or high user rating. Also, after identifying the three applications to be examined, it was found that the number of tasks in application A and C were limited. The researchers were looking for tasks that could be completed and have interaction with the users; in other words, tasks that ask the user for inputs. In fact, for application C, there was only one task that could meet the requirements of the study. As a result, we could not examine a wide array of tasks for each application.

VII. Conclusion

It has been shown in this paper that a star rating is not significantly correlated with usability on mobile apps. On the other hand, the task completion time, the number of errors that may occur while performing the task and the SUS ratings were significantly correlated to each other. However, this finding is

based on a single random task for each application. Another study is highly recommended to find out whether the real rating of an application depends on the task that the user needs to complete.

ACKNOWLEDGMENT

The authors like to thank all participants in this experiment. The authors also appreciate the funding provided by King Faisal University.

REFERENCES

[1] Bangor, Aaron, Philip T. Kortum, and James T. Miller. "An empirical evaluation of the system usability scale." *Intl. Journal of Human–Computer Interaction* 24.6 (2008): 574-594.

[2] Bangor, Aaron, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale." *Journal of usability studies* 4.3 (2009): 114-123.

[3] Borsci, Simone, Stefano Federici, and Marco Lauriola. "On the dimensionality of the System Usability Scale: a test of alternative measurement models." *Cognitive processing* 10.3 (2009): 193-197.

[4] Lewis, J.R., and Sauro, J. The Factor Structure of the System Usability Scale. In Proc. HCII 2009, Springer-Verlag (2009), 94-103.

[5] Mendoza, V., & Novick, D. G. (2005, September). Usability over time. In Proceedings of the 23rd annual international conference on Design of communication: documenting & designing for pervasive information (pp. 151-158). ACM.

[6] Morris, M. G., & Dillon, A. (1997). The influence of user perceptions on software utilization: application and evaluation of a theoretical model of technology acceptance.

[7] Raptis, Dimitrios, et al. "Does size matter? Investigating the impact of mobile phone screen size on users' perceived usability, effectiveness and efficiency." *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services.* ACM, 2013.

[8] Turel, O., & Serenko, A. (2006). Satisfaction with mobile services in Canada: An empirical investigation. Telecommunications Policy, 30(5), 314-331.