2020

An Experiment: The Optimal Number of Participants for the Usability Testing of Mobile Apps

Xiaofan Zhao
Auckland University of Technology, zhaoxiaof231@163.com

Ramesh Lal
Auckland University of Technology, ramesh.lal@aut.ac.nz

An Experiment: The Optimal Number of Participants for the Usability Testing of Mobile Apps

Completed research paper

Xiaofan Zhao
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
City Campus, Auckland, New Zealand
Email: zhaoxiaof231@163.com

Ramesh Lal
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Manukau Campus, Auckland, New Zealand
Email: ramesh.lal@aut.ac.nz

Abstract

Mobile applications (apps) are being released rapidly with the development of smart devices. Usability is a critical success factor and it is essential to conduct usability tests before launching mobile apps. The aim of this research was to find the optimal number of participants for the usability testing of mobile apps. This research involved conducting 4 rounds of usability tests using representatives of the users of a mobile app "JB-Career-Connect", which provides a platform for establishing direct connection between graduating students and employers. Each of the testing rounds had different number of participants. Our test results show that 2 testers detected 16% of the usability issues; 5 testers detected 36% of the usability issues; while 9 and 12 testers each detected 64% of the usability issues. 9 testers appear to be more cost-effective since they performed well in other usability metrics. Our research provides evidence that 9 testers is the optimal number of participants required for testing the usability of mobile apps.

Keywords Mobile Applications, Usability, Usability Testing, User Interface

1 Introduction

Mobile applications (apps) are impacted by quality issues, fierce competition, and lack of customer loyalty (Alshamari & Mayhew, 2009). Hence, adopting usability testing as part of the development process for mobile apps is critical for success at the marketplace (Hwang& Salvendy, 2010). Usability testing ensures that applications are easy to use (Bevan, 1995). Usability testing requires representatives of each user group to accomplish a series of pre-identified tasks while observing them interact with the system (Alshamari & Mayhew, 2009). The number of usability testers is an important factor requiring balancing costs against benefits of the usability testing (AlRoobaea and Mayhew, 2014).

This research seeks to answer the following question "what is the optimal number of usability testers required for mobile apps". We conducted an experiment involving four different testing groups with different number of testers to evaluate the usability of a mobile app "JB Career Connect". It provides a platform to establish a direct connection between graduating students and potential employers. Its features include: user registration, CV uploading, job searching, employment information posting, QR code scanning, adding friends, and profiles printing. We tested the app with 28 final (3rd) year students from the Bachelor of Computer and Information Sciences (BCIS) degree programme at Auckland University of Technology (AUT).

Next, information is provided on the related studies, followed by information on the research methodology. We then present the usability test results and discussion. Finally, the research limitation and suggestion for future work are provided.

2 Related Work

2.1 Usability Testing

Usability is defined by Nigel Bevan (1995) as the *ease of use* and *acceptance of a system*- one impacts the performance and user satisfaction, and the other impacts the actual system use. Usability determines the quality of a system (Alshamari & Mayhew, 2009). Usability testing ensures that issues are detected before it is released (Cazañas et al., 2017). Usability testing has five major parameters: effectiveness, efficiency, learnability, satisfaction and errors (Resnik, 2011). Effectiveness is the accuracy and completeness with which users perform tasks (Simorangkir et al., 2018). Efficiency is the use of resources in relation to accuracy and completeness of tasks (Adhy et al., 2018). Learnability is the difficulty levels to perform tasks (Zhang & Adipat, 2005). Satisfaction is the user perception, feeling, and opinion (Adhy et al., 2018). Error relates to the mistakes made and recovery from it (Zhang & Adipat, 2005).

2.2 Methods for Usability Testing

There are several usability testing methods such as Heuristic Evaluation, Cognitive Walkthrough, Think-aloud protocols and Survey/Questionnaire method. Holzinger (2005) classifies Heuristic Evaluation, Cognitive Walkthrough and Action Analysis as Usability Inspection Methods while classifying Thinking-aloud, Field Observation and Questionnaires as Usability Testing Methods. A combination of these methods enables a reliable usability testing of applications (Goh et al., 2013).

Heuristic Evaluation (HE) is a widely used method for design evaluation. It involves a series of heuristics (guidelines) as part of the procedure to test a system (Khajouei et al., 2017). Cognitive Walkthrough (CW) is based on a cognitive model (Mahatody, Sagar & Kolski, 2010). CW uses a detailed procedure to simulate the user problem solving process including the user goals to guide the next step to accomplish a task. Unlike HE, CW tests the difficulty level, requiring deciding the actions needed to accomplish a task (Khajouei et al. 2017).

Think-aloud (TA) requires testers to do a set of tasks, and asks them to express their thoughts and task performance during the usability testing. TA provides insight into the thought processes and user experience interacting with the targeted system (Alhadreti, 2016). There are two types of Think-aloud method: Concurrent Think-aloud and Retrospective Think-aloud. Concurrent Think-aloud requires the participants to articulate their thoughts during task execution while the Retrospective Think-aloud requires the testers to express their experiences after task completion (Willis & Mcdonald, 2016).

2.3 Mobile Applications Usability Testing

Adhy et al. (2018) report usability test result of a mobile app "WeMo", used to monitor weather. This app allows Diponegoro University to share the weather information. Ten participants were invited for usability testing and two usability testing methods were used: the performance-based evaluation method was used to test efficiency and effectiveness while the questionnaire-based evaluation method was used to test satisfaction and learnability. The completeness and execution time of tasks were used to calculate the efficiency and effectiveness of the app. The satisfaction and learnability were tested via a questionnaire. Their score on effectiveness, efficiency, satisfaction, and learnability (93.33%, 91.57%, 83.6% and 83.2%) suggest an easy to use app.

2.4 Tester Size for Usability Testing

To determine the optimal number of testers, economic and scientific implications need to be considered (AlRoobaea and Mayhew, 2014). Nielsen and Molich (1990) suggest 5 testers are likely to find 2/3 of the usability issues. Virzi (1992) proposed 4 or 5 users would detect 80% of the usability problems. Hence, it has become normal to use 4 or 5 testers for usability testing. However, Spool & Schroeder's (2001) study shows only 35% of the usability problems were discovered by the first five participants. Hwang and Salvendy's (2010) recommend 10±2 testers. AlRoobaea and Mayhew (2014) have proposed using 16±4 testers while Cazañas et al. (2017) recommend using 20 testers to discover 90% of the usability problems. There are many factors which may determine the size of the usability testing team such as the budget, time allocated for testing, testing purpose and methods, user groups, and the complexity of system. (AlRoobaea and Mayhew, 2014).

3 Methodology

It was decided to involve 4 rounds of testing of the mobile app with further 2 rounds of testing to follow at a later stage. The app has three different user groups: 1. final year undergraduate students- looking for IT related jobs; 2. employers- want to hire IT graduates; 3. administrator - organiser of the career fair events and the venue. We invited the student user group to test the usability of the app.

3.1 Participants

In total, there were 28 participants used for the usability testing the app, grouped into four rounds: Round 1- 2, Round 2- 5, Round 3- 9, and Round 4- 12, participants. The number of participants for each round and the number of rounds of usability testing for this experiment were based on our findings from the literature. They were undertaking a final year R&D project paper, part of the BCIS degree qualification at AUT. Each round of the experiment was independent and offered easier observation and better comparison of results. Participants carried-out the tasks on the mobile app "JB Career Connect". Each testing round was expected to take between 20-90 minutes, carried-out in the post-graduate computer lab. Participants were explained about the test and were given a description of the app. They were video-recorded when doing the test and filled out the questionnaire after completing the test. The further two rounds of testing at later stage will have the following number of participants- Round 5- 15 and Round 6- 20 participants.

3.2 Usability Testing

We adopted the five parameters (effectiveness, efficiency, learnability, satisfaction, and errors) to test the functionality (Resnik, 2011) and adopted the 10 User Interface Design Guidelines to test the interface (Nielsen & Molich, 1989).

3.2.1 Functionality

The following 11 tasks were to be performed using the JB Career Connect app: 1. New user registration, 2. View the information on events; 3. View the venue map; 4. View the information of presenters; 5. View the information of jobs; 6. Give your feedbacks to the app; 7. Log-out and re-log-in your account; 8. View and edit user profiles; 9. Scan a QR code; 10. Share your identity to others; 11. Add and view friends.

The five parameters enabled to measure the performance and usability for effectiveness, efficiency, learnability, satisfaction, and errors. To measure the effectiveness and efficiency, performance- based evaluation methods were used (Adhy et al., 2018). The ISO 9241-11 standards were used to calculate the scores of efficiency and effectiveness (Simorangkir et al., 2018). For learnability and satisfaction, questionnaire-based evaluation method was adopted, using the data to calculate the ease of use (Adhy et al., 2018). Any errors during the tests were captured.

Effectiveness is the accuracy and completeness with which users perform specified tasks (Adhy et al., 2018). Effectiveness = $\frac{\text{the number of complete tasks}}{\text{the number of all tasks}} * 100\%$.

Efficiency refers to all the resources used in relation to the accuracy and completeness with which users perform tasks. In this research, resource refers to execution time (Adhy et al., 2018). Efficiency = $\frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{ij}t_{ij}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{ij}} * 100\%$. Where R = the total number of participants; N = the total number of tasks; n_{ij} = the result of task completeness. If a participant successfully completes a specified task, then $n_{ij} = 1$; if a participant fails to perform a specified task, then $n_{ij} = 0$. t_{ij} = the execution time spent by a participant to perform a specified task. Time must be recorded for all the participants regardless of a successful completion or a failure to complete a task (quits a task).

Learnabiliity refers to the difficulty levels for users to perform tasks (Zhang & Adipat, 2005). Satisfaction refers to users' perceptions, feelings, and opinions about the product, and it is usually captured through a questionnaire or survey (Zhang & Adipat, 2005).

It is critical to capture errors made during testing to determine the seriousness of each error, and ease to recover from them (Zhang & Adipat, 2005). Errors include slips and mistakes. Slips occur when users aim to do one thing but end up doing another similar but a different task. Mistakes occur when a user aims to perform a task, takes the correct steps but it ends in a failure.

The following data was captured: task completion, execution time and errors made. The data from questionnaire on learnability and satisfaction was compiled. Task completion data was used to calculate effectiveness. Execution time for each task was used to calculate efficiency. All errors occurred in tests was recorded (slips and mistakes). A questionnaire was designed to collect the information on learnability and satisfaction from all the participants. Table 1 shows the learnability and satisfaction questionnaire which includes the Five-score Likert Scale- "1. strongly disagree", "2. disagree", "3. neutral", "4. agree", and "5. strongly agree".

No.	Statements and Questions	Score
	Learnability	
1.	In my opinion, the interface of JB Connect application is easy to learn.	
2.	In my opinion, the menu provided in the application is easy to be used.	
3.	In my opinion, the buttons provided in the application are easy to be used.	
4.	In my opinion, the help documentation provided in the application is helpful.	
5.	I can use the application without help from the developer or technical person.	
	Satisfaction	
6.	In my opinion, the JB Connect application is user-friendly	
7.	In my opinion, my first reaction to the application is good.	
8.	In my opinion, the interface display of the application is good.	
9.	In my opinion, this application makes it easier for me to get and share information of jobs.	
10.	Overall, I am satisfied with this application and I will use it later.	

Table 1. Learnability and satisfaction questionnaire

3.2.2 User Interface Design

The 10 User Interface Design Guidelines (rules) adopted to test the interface usability of the app based on Nielsen & Molich (1989) guidelines were: 1. Visibility of system status; 2. Match between system and the real world; 3. User control and freedom, 4. Consistency and standards; 5. Error prevention; 6. Recognition rather than recall; 7: Flexibility and efficiency of use; 8. Aesthetic and minimalist design; 9. Help users recognize, diagnose, and recover from errors; and 10. Help and documentation. Each participant filled out the questionnaire including giving marks for these 10 rules (from 1 to 10) and listing any other interface issues they encountered or any suggestions for improvement with the UI.

3.3 Data Collection

The following methods were used to collect data with all the four rounds: 1. timekeeping-recorded the execution time of each task (for calculating the efficiency score); 2. video recording- during usability

testing participants' actions and reflections were recorded (to be used for retrospectives, and improve the usability of the features by minimizing time taken to execute them); and 3. written records- participants filled questionnaires after finishing the usability tests (for evaluating learnability, satisfaction, and the interface).

4 Experiment Results

In this section the results of the four rounds of usability testing experiment are presented. Results are presented on effectiveness, efficiency, learnability, satisfaction, errors, and interface satisfaction *(summarized results of Round 2, 3 and 4 usability tests are provided due to page limitation)*.

4.1 Round 1 Testing- 2 Usability Testers

4.1.1 Functionality

- **Effectiveness**

Participant 1 did not complete task 1 while both did not complete Task 11. The rest of the tasks were successfully completed by both participants.

Overall Effectiveness = $\frac{\text{the number of complete tasks}}{\text{the number of all tasks}} * 100\% = \frac{19}{22} * 100\% = 86.36\%$.

- **Efficiency**

Table 2 provides data on execution time relating to 11 tasks.

Overall Efficiency = $\frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{ij}t_{ij}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{ij}} * 100\% = \frac{472}{750} * 100\% = 62.93\%$.

	Participant 01	Participant 02	Average
Task 01	266 (Fail)	155	210.50
Task 02	11	7	9.00
Task 03	14	7	10.50
Task 04	7	4	5.50
Task 05	11	12	11.50
Task 06	18	11	14.50
Task 07	21	22	21.50
Task 08	49	58	53.50
Task 09	17	20	18.50
Task 10	19	9	14.00
Task 11	4 (Fail)	8 (Fail)	6.00

Table 2. Execution time to complete tasks (in seconds)

- **Learnability**

Table 3 provides the scores on learnability.

The learnability percentage (the 3rd column) = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall Learnability = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{70\%+80\%+90\%+60\%+60\%}{5} = 72\%$.

	Participant 01	Participant 02	Percentage
Question 01	4	3	70.00%
Question 02	4	4	80.00%
Question 03	5	4	90.00%
Question 04	3	3	60.00%
Question 05	3	3	60.00%

Table 3. Learnability score

- **Satisfaction**

Table 4 shows the scores relating to satisfaction.

The satisfaction percentage (the 3rd column) = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall Satisfaction = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{80\%+90\%+70\%+70\%+70\%}{5} = 76\%$.

	Participant 01	Participant 02	Percentage
Question 06	4	4	80.00%
Question 07	4	5	90.00%
Question 08	3	4	70.00%
Question 09	3	4	70.00%
Question 10	4	3	70.00%

Table 4. Satisfaction score

- **Errors**

  a) The registration is unclear and confused. "Register" function is used for registering for an event instead of registering a new account. One participant made this mistake. Users should go to "Login" function and use "Sign up" button to create new accounts.

  b) Both two participants in this experiment were not able to add friends after scanning QR codes.

4.1.2 Interface

- **Interface testing based on 10 heuristic guidelines**

Table 5 shows the scores relating to interface testing.

The interface percentage (the 3rd column) = $\frac{\text{Sum of score for each rule}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

User control and freedom reported the lowest score (60%).

Overall UI Design Score = $\frac{\text{Sum of percentage}}{\text{The number of guidelines}} = \frac{75\%+75\%+60\%+70\%+70\%+70\%+70\%+80\%+75\%+70\%}{10} = 71.5\%$.

10 Heuristic Guidelines	Participant01	Participant02	Percentage
1. Visibility of system status	7	8	75.00%
2. Match between system and the real world	8	7	75.00%
3. User control and freedom	5	7	60.00%
4. Consistency and standards	7	7	70.00%
5. Error prevention	7	7	70.00%
6. Recognition rather than recall	7	7	70.00%
7. Flexibility and efficiency of use	6	8	70.00%
8. Aesthetic and minimalist design	8	8	80.00%
9. Help users recognize, diagnose and recover from errors	7	8	75.00%
10. Help and documentation	7	7	70.00%

Table 5. Interface score

- **Issues identified by the participants**

  a) The registration is unclear and confused.

  b) Buttons and input boxes could be bigger.

  c) The system has no "go back" button.

  d) Users may be able to have freedom of some customizations, such as move icons locations.

4.2 Round 2 Testing- 5 Usability Testers

4.2.1 Functionality

- **Effectiveness**

Participants 2, 3 and 5 failed to complete task 1. Participant 1 failed to complete task 8, while participant 2 failed to complete task 9. All 5 participants failed to complete task 11. Overall Effectiveness = $\frac{\text{the number of complete tasks}}{\text{the number of all tasks}} * 100\% = \frac{45}{55} * 100\% = 81.82\%$.

- **Efficiency**

Overall Efficiency = $\frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{ij}t_{ij}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{ij}} * 100\% = \frac{819}{1517} * 100\% = 53.98\%$.

- **Learnability**

The learnability percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall Learnability = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{84\%+80\%+80\%+72\%+84\%}{5} = 80.00\%$.

- **Satisfaction**

The satisfaction percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall Satisfaction = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{84\%+88\%+84\%+80\%+80\%}{5} = 83.20\%$.

- **Errors**

a) The registration functions are confused. "Register" function is used for registering for an event instead of registering a new account. Some of participants made this mistake. Users should go to "Login" function and use "Sign up" button to create new accounts.

b) One participant was not able to save the changes when he selfied and uploaded picture.

c) All participants in this experiment were not able to add friends after scanning QR codes.

4.2.2 Interface

- **Interface testing based on 10 heuristic guidelines**

The interface percentage = $\frac{\text{Sum of score for each rule}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall UI Design Score = $\frac{\text{Sum of percentage}}{\text{The number of guidelines}} = \frac{76\%+78\%+70\%+72\%+80\%+80\%+70\%+80\%+76\%+78\%}{10} = 76.00\%$.

- **Issues identified by the participants**

a) The registration is confused and unclear.

b) Buttons could be bigger.

c) The system has no "go back" button.

d) There is no specified "log out" button.

e) There are some bugs when taking selfies and uploading profile pictures, pictures were not able to be saved, or pictures were squashed.

f) A prompt tone is suggested to be added for selfies, because some users were not aware that they had finished their selfies, and then took again and again.

4.3 Round 3 Testing- 9 Usability Testers

4.3.1 Functionality

- **Effectiveness**

Task 1 was not completed by Participant 3, 4, 6, 7, 8, 9. Task 6 was not completed by participant 6. Task 9 was not completed by participant 5. Task 11 was not completed by all 11 participants. The other tasks were completed successfully by the participants.

Overall Effectiveness = $\frac{\text{the number of complete tasks}}{\text{the number of all tasks}} * 100\% = \frac{81}{99} * 100\% = 81.82\%$.

- **Efficiency**

Overall Efficiency = $\frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{ij}t_{ij}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{ij}} * 100\% = \frac{1556}{2859} * 100\% = 54.42\%$.

- **Learnability**

The learnability percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall Learnability = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{77.78\%+73.33\%+84.44\%+51.11\%+71.11\%}{5} = 71.55\%$.

- **Satisfaction**

The satisfaction percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}}$ * 100%.

Overall Satisfaction = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{75.56\%+75.56\%+77.78\%+66.67\%+62.22\%}{5}$ = 71.56%.

- **Errors**

a) The registration functions are confused. "Register" function is used for registering for an event instead of registering a new account. Some of participants made this mistake. Users should go to "Login" function and use "Sign up" button to create new accounts.

b) Two participants were not able to use camera when they scanned the QR code.

c) One participant was not able to change the number of stars when she gave feedbacks.

d) All participants in this experiment were not able to add friends after scanning QR codes.

4.3.2 Interface

- **Interface testing based on 10 heuristic guidelines**

The interface percentage = $\frac{\text{Sum of score for each rule}}{\text{Sum of full mark} * \text{Number of testers}}$ * 100%.

Overall UI Design Score = $\frac{\text{Sum of percentage}}{\text{The number of guidelines}} = \frac{63.33\%+68.89\%+54.44\%+68.89\%+54.44\%+60\%+56.67\%+80\%+61.11\%+57.78\%}{10}$ = 62.56%.

- **Issues identified by the participants**

a) The registration is unclear and confused.

b) The system does not have a "go back" button.

c) The system does not have a specified "log out" button.

d) A prompt tone is suggested to be added for selfies, because some users were not aware that they had finished their selfies, and then took again and again.

e) After uploading or taking profile picture, it becomes squashed.

f) Some instructions on the app are ambiguous. For example, in profiles, the formulation of "year of study" was easy to be misinterpreted. Some participants filled in "3" because they thought that "year of study" meant the duration of study; while others filled in "2018" because they thought that "year of study" meant the beginning year of study.

g) Possibly the system could have drop-down boxes for degrees and courses in the registration.

h) The system does not provide adequate and good help documentations.

i) Users should be able to sign off their accounts if they do not use the app any more.

4.4 Round 4 Testing- 12 Usability Testers

4.4.1 Functionality

- **Effectiveness**

Task 1 was not completed by Participant 1,2, 3, 4 & 5. Task 3 was not completed by participant 9, 10, 11 & 12. Task 4 was not completed by participant 9 & 11. Task 5 was not completed by participant 9. Task 7 was not completed by participant 5, 6, 7 & 8. Task 9 was not completed by participant 4, 5, 6, 7, & 8. all 11 participants did not complete task 11. . Otherwise, the other tasks were completed successfully by the participants.

Overall Effectiveness = $\frac{\text{the number of complete tasks}}{\text{the number of all tasks}}$ * 100% = $\frac{99}{132}$ * 100% = 75.00%.

- **Efficiency**

Overall Efficiency = $\frac{\sum_{j=1}^{R}\sum_{i=1}^{N} n_{ij}t_{ij}}{\sum_{j=1}^{R}\sum_{i=1}^{N} t_{ij}}$ * 100% = $\frac{2491}{4198}$ * 100% = 59.33%.

- **Learnability**

Learnability percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}}$ * 100%.

Overall Learnability = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{75\%+76.67\%+75\%+66.67\%+60\%}{5}$ = 70.67%.

- **Satisfaction**

Satisfaction percentage = $\frac{\text{Sum of score for each question}}{\text{Sum of full mark} * \text{Number of testers}}$ * 100%.

Overall Satisfaction = $\frac{\text{Sum of percentage}}{\text{The number of questions}} = \frac{68.33\%+78.33\%+70\%+76.67\%+66.67\%}{5}$ = 72.00%.

- **Errors**

a) The registration functions are confused. "Register" function is used for registering for an event instead of registering a new account. Some of participants made this mistake. Users should go to "Login" function and use "Sign up" button to create new accounts.

b) Two participants typed the correct email address but the system showed email address was invalid. They used their AUT email here. This problem could be solved if they replaced "xxx@autuni.ac.nz" with "xxx@aut.ac.nz".

c) Some participants failed to scan the QR code, because it was default front-facing camera and not able to switch to the rear camera.

d) Some participants were not able to see some functions on the homepage, when they completed the login. For example, the icons of "Map", "Presenters" and "Jobs" were missing on the homepage. After refreshing application data, this problem could be solved.

e) All participants in this experiment were not able to add friends after scanning QR codes.

4.4.2 Interface

- **Interface testing based on 10 heuristic guidelines**

Interface percentage = $\frac{\text{Sum of score for each rule}}{\text{Sum of full mark} * \text{Number of testers}} * 100\%$.

Overall UI Design Score = $\frac{\text{Sum of percentage}}{\text{The number of guidelines}} = \frac{66.67\%+70\%+50.83\%+74.17\%+49.17\%+66.67\%+61.67\%+82.5\%+54.17\%+48.33\%}{10} =$ 62.42%.

- **Issues identified by the participants**

a) The registration is unclear and confused.

b) The system does not have a "go back" button.

c) The system does not have a specified "log out" button.

d) The sidebar menu (three white bars) is hard to find.

e) Pressing the black "AUT" button can go to the home page, there is no hint of this.

f) Some prompts and tips are too small to be obvious, such as the prompt at password length.

g) A prompt tone is suggested to be added for selfies, because some users were not aware that they had finished their selfies, and then took again and again.

h) After uploading or taking profile picture, it always becomes squashed.

i) Some instructions on the app are ambiguous. For example, in profiles, the formulation of "year of study" was easy to be misinterpreted. Some participants filled in "3" because they thought that "year of study" meant the duration of study; while others filled in "2018" because they thought that "year of study" meant the beginning year of study.

j) In profiles, the information of "year of study" is not able to be displayed after users have filled it out.

k) The system does not provide adequate and good help documentations.

4.5 The Number of Identified Usability Issues

In the four rounds of usability testing, the 28 participants identified 102 usability issues in total. However, our analysis showed that there were overlapping issues i.e. most of these 102 issues had been identified multiple times by different testers. Once all the duplications were removed there are only 25 unique usability issues. All the usability issues have been reported to the developer team of JB Career Connect for the app upgrade and improvement.

5 Discussion

Table 6 lists the calculated scores on effectiveness, efficiency, learnability, satisfaction, and user interface design using the data collected through the four rounds of usability testing. All scores have been converted into percentages (shown in the table below). Our discussion is based on the overall percentage for each round of testing (the last row in the table).

	Round 1	Round 2	Round 3	Round 4
	2 Testers	5 Testers	9 Testers	12 Testers
Effectiveness	86.36%	81.82%	81.82%	75.00%
Efficiency	62.93%	53.98%	54.42%	59.33%
Learnability	72.00%	80.00%	71.55%	70.67%
Satisfaction	76.00%	83.20%	71.56%	72.00%
UID Scores	71.50%	76.00%	62.56%	62.42%
Overall Percentages	73.76%	75.00%	68.38%	67.88%

Table 6. The overall scores of 4 rounds of testing

Low percentage indicates higher levels/numbers of usability issues serving the goal of identifying and fixing usability issues before deployment preventing the negative impact on users and customers. Round 2 (with 5 testers) has the highest percentage (75.00%) while Round 4 (with 12 testers) has the lowest percentage (67.88%). This result suggests using more testers will likely to identify a higher number of usability issues if cost and time are not factors.

However, when comparing the test result of Round 1 (with 2 testers) with the result of Round 2 (with 5 testers) throws an interesting findings which is different from Nielsen and Molich's (1990) suggestion of using 5 testers and usually taken as the best practice for usability testing. Our findings suggest that 2 testers performed better in usability testing compared to 5 testers.

Our results also show that ***more than 5 testers*** performed better in usability testing and enable more usability issues discovered (9 testers- 68.38% and 12 testers- 67.88%). When comparing with the results of 5 testers, the difference is 6.62% (with 9 testers) and 7.12% (with 12 testers).

In addition, as mentioned earlier in section 4.5, there were 25 unique usability issues of JB Career Connect app identified in the tests after eliminating the issues repeatedly identified. Hence, Round 1 (with 2 testers) detected 4 usability issues; Round 2 (with 5 testers) detected 9 usability issues; Round 3 (with 9 testers) and Round 4 (with 12 testers) each detected 16 usability issues, after eliminating the repeats. For comparison we converted the above into percentages. 2 testers found 16% of the usability issues; 5 testers found 36% of the usability issues; 9 testers and 12 testers each found 64% of the usability issues, distinctly 9 is more cost-effective.

Our result shows a minor difference of 0.5% between Round 3 (9 testers) and 4 (12 testers) in usability testing metrics. These 9 and 12 testers each identified 64% of usability issues. Hence if cost, time or both are a major project constraint, 9 testers appear to be an appropriate group size for usability testing of mobile apps.

While our study suggests a large size (more than 5 individuals) testing group is likely to identify more usability issues, we are not yet certain the size of the testing group that is likely to be cost-effective while achieving the goal to identify most of the usability issues with mobile apps. To gain further understanding we have already begun a separate study to further test the same mobile app with three other different group sizes which are as follows: Round 5- 15 individuals, Round 6- 20 individuals, and Round 7- 25 individuals. The testing individuals would be selected from the same user group (final year BCIS students at university) as for this investigation.

6 Conclusion, Limitation and Future Work

Our research findings suggest that the usability testing of any mobile app will require more than 5 individual testers representing the end-user groups or customers. While our research suggests more usability testers would likely to produce a better result (more usability issues identified), we would recommend 9 usability testers based on the cost and time constraints rather than 12 as a result of our findings showing only 0.5% difference in the overall usability test results between the two group sizes

One limitation of this research was that we did not conduct (due to time issues) more rounds of testing involving larger group size of more than 12 testers to gain a better understanding on the number of participants required for usability testing of mobile apps. Another limitation was that we could not involve two other user groups – the employers and administration. The available time for this research was a major reason for not including them as the participation for usability testing. They were regarded as minor users of the mobile app. We are in the process of addressing our first limitation through conducting another separate investigation to do further rounds of usability testing with three other groups involving more than 12 testers. At a later stage, we are planning to undertake a further investigation targeting the two minor user groups.

Any future research may consider focusing on creating a framework providing procedures for usability testing for an optimal result based on a smallest possible group size considering the cost and time factors. Moreover, it is necessary to have independent but similar research to validate our research findings since we must always strive to improve the usability of mobile apps to promote better user experience.

7 References

Adhy, S., Prasetio, A., Noranita, B., & Saputra, R. 2018. "Usability Testing of Weather Monitoring on Android Application," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS). doi:10.1109/icicos.2018.8621752

Alhadreti, Obead. 2016. "Thinking about Thinking Aloud : An Investigation of Think-Aloud Methods in Usability Testing," January. https://search-ebscohost-com.ezproxy.aut.ac.nz/login.aspx?direct=true&db=edsble&AN=edsble.699593&site=eds-live.

Alroobaea, R., and P.J. Mayhew. 2014. "How Many Participants Are Really Enough for Usability Studies?" 2014 Science and Information Conference, Science and Information Conference (SAI), 2014, August, 48–56. doi:10.1109/SAI.2014.6918171.

Alshamari, M., & Mayhew, P. 2009. "Technical Review: Current Issues of Usability Testing," IETE Technical Review (26:6), 402. doi:10.4103/0256-4602.57825

Bevan, N. 1995. "Measuring Usability as Quality of Use," Software Quality Journal (4:2), 115–130. doi: 10.1007/bf00402715

Cazañas, A., de San Miguel, A. and Parra, E. 2017. Estimating Sample Size for Usability Testing. Enfoque UTE. 8, 1 (Feb. 2017), pp. 172-185. doi:https://doi.org/10.29019/enfoqueute.v8n1.126.

Goh, K., Chen, Y., Lai, F., Daud, S., Sivaji, A., & Soo, S. 2013. "A Comparison of Usability Testing Methods for an E-Commerce Website: A Case Study on a Malaysia Online Gift Shop," 2013 10th International Conference on Information Technology: New Generations. doi:10.1109/itng.2013.129

Holzinger, A. 2005. "Usability engineering methods for software developers," Communications of the ACM (48:1), 71-74.

Hwang, W., & Salvendy, G. 2010. "Number of people required for usability evaluation," Communications of the ACM (53:5), pp. 130. doi:10.1145/1735223.1735255

Khajouei, R., Esfahani, M., & Jahani, Y. 2017. "Comparison of heuristic and cognitive walkthrough usability evaluation methods for evaluating health information systems," Journal of the American Medical Informatics Association (24:1), pp. e55–e60, https://doi.org/10.1093/jamia/ocw100

Mahatody, T., Sagar, M., & Kolski, C. 2010. "State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions," International Journal of Human-Computer Interaction (26:8), pp. 741-785. doi:10.1080/10447311003781409

Nielsen, J., & Molich, R. 1990. "HE of user interface. Conference Proceedings," ACM, pp. 249-256.

Nielsen, J., & Molich, R. 1989. "Teaching user interface design based on usability engineering," ACM SIGCHI Bulletin (21:1), pp. 45–48. doi: 10.1145/67880.67885

Resnik, L. 2011. "Development and testing of new upper-limb prosthetic devices: Research designs for usability testing," The Journal of Rehabilitation Research and Development (48:6), pp. 697. doi:10.1682/jrrd.2010.03.0050

Simorangkir, G. D., Sarwoko, E. A., Sasongko, P. S., Sutikno, & Endah, S. N. 2018. "Usability Testing of Corn Diseases and Pests Detection on a Mobile Application," 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS). doi:10.1109/icicos.2018.8621842

Spool, J., & Schroeder, W. 2001. "Testing Web sites: five users is nowhere near enough," Proceedings of the Conference Extended Abstracts on Human Factors in Computing Systems, CHI'2001, ACM Press, New York.

Virzi, R.A. 1992. "Refining the test phase of usability evaluation: How many subjects is enough? " Human Factors, pp. 457-468.

Willis, L. M., & Mcdonald, S. 2016. "Retrospective protocols in usability testing: A comparison of Post-session RTA versus Post-task RTA reports," Behaviour & Information Technology (35:8), pp. 628-643. doi:10.1080/0144929x.2016.1175506

Zhang, D., & Adipat, B. 2005. "Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications," International Journal of Human-Computer Interaction (18:3), pp. 293-308. doi:10.1207/s15327590ijhc1803_3

Copyright