USABILITY TESTING OF FITNESS MOBILE APPLICATION: METHODOLOGY AND QUANTITATIVE RESULTS

Ryan Alturki and Valerie Gay

School of Electrical and Data Engineering,
University of Technology Sydney, Sydney City, Australia

ABSTRACT

Obesity is a major health problem around the world. Saudi Arabia is a nation where obesity is increasing at an alarming rate. Mobile apps could help obese individuals but they need to be usable and personalized to be adopted by those users. This paper aims at testing the usability of a fitness mobile app" Twazon", an app in Arabic language. This paper presents an extensive literature review on the attributes that improve the usability of fitness apps. Then, it explains our methodology and our set up of a trial to test the usability of Twazon app that is popular in Saudi Arabia. The usability attributes tested are effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load. The trial is done in collaboration with participants from the Armed Forces Hospitals - Taif Region in Saudi Arabia. The results highlight that the app failed to meet with the usability attributes.

KEYWORDS

Usability, Mobile Application, Obesity, User Experience

1. INTRODUCTION

Obesity is defined as an excessive storage of energy in the form of fat [1]. According to the facts provided by World Health Organization (WHO) Media Centre, 13% of world's adult population is considered obese, and 39% of the adult population is believed to be overweight. The prevalence of obesity around the world has doubled between 1980 and 2014 [2]. Saudi Arabia is one such country where obesity is increasing at an alarming rate. The study by Coronary Artery Disease in Saudis (CADISS) in 2005 found that 35.5% of people in the country are obese which means every third person in the country is affected. A National Nutrition Survey of 2007 mentioned that obesity is a significant concern because the prevalence of obesity in the men (14%) and women (23.6%) in Saudi Arabia [3]. Overweight and obesity are considered as major risk factors for various chronic diseases such as cancer, cardiovascular diseases and diabetes [4-8].

The health problems and diseases that result from obesity have encouraged a lot of researchers to discuss how the condition can be overcome or prevented [9-14]. Most of the research work states that obesity can be overcome by increasing physical activity and changing eating behaviour. However, it is sometimes very difficult to motivate obese individuals to change their lifestyle and become involved in physical activity. Research has shown that the most effective behaviour change related to fitness and health occurs through behaviour interventions [15-18]. Mobile technology such as mobile applications (apps) have been found to be a very useful intervention

Natarajan Meghanathan et al. (Eds) : ICCSEA, WiMoA, SPPR, GridCom, CSIA - 2017
pp. 97– 114, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.71108

tool for increasing physical activity because through their unique features these apps motivate individuals to achieve their fitness goals [19-22].

Fitness apps are becoming increasingly popular both around the world and in Saudi Arabia. Smartphones and their apps have seen an exponential growth in their usage in Saudi Arabia in recent times. Researchers ranked the country third overall in terms of global smartphone usage penetration [23]. In 2016, smartphone users in Saudi Arabia are estimated to be near 15.9 million and this figure is estimated to increase to 19.1 million by 2019 [24]. Because of cultural restrictions many people, but especially women, find it easier to interact publically and socially in a virtual environment through mobile apps on smartphones. The increasing ubiquity of smartphone technology provides an opportunity to develop an Arabic app to help fight obesity.
Sometimes apps development can cost millions of dollars but most of the apps fail miserably [25]. Of all the branded apps, 80% are downloaded less than one thousand times and only 1% has been downloaded one million times or more. After downloading, 25% of mobile apps are never used again [26]. Ample market research suggests that the main reason of failure of mobile apps is the lack of usability [25, 27-28]. The usability of mobile apps enhances user experience (UX) and can play a significant part in the success of the mobile apps [29-32].

The population of obese and overweight individuals in Saudi Arabia is increasing at an alarming rate. Therefore, the increasing use of health and fitness apps in Saudi Arabia is an opportunity to introduce a technological solution which involves developing an app will be popular among obese individuals. There are many Arabic health and fitness apps available but to our knowledge none of these apps have been built with the purpose of enhancing the usability of the app to motivate people to lose weight by considering usability attributes and factors. Moreover, there is no research on fitness and health apps that outlines guidelines for usability. Moreover, to the best of our knowledge, there is no Arabic app that uses any specific features that enhances UX for obese individuals. This leads to our research problem:

- How to improve mobile fitness apps usability to help obese users to reach their health and fitness goals?

In order to solve the above issue, we will start by testing the usability of a popular fitness mobile app that is targeted Saudi users. Then we will develop a new usability guideline that will be specifically designed for fitness apps that help obese users to lose weight. Based on this guideline, we will develop a new fitness app. Our primary focus will be users in Saudi Arabia because the country has a high percentage of obesity rate among its citizens.

This paper presents our method and results of the usability of “Twazon”, an Arabic-language fitness mobile apps. Seven usability attributes: effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load were tested collaboration with the Armed Forces Hospitals in the Taif Region of Saudi Arabia, which provided the participants. All participants are people suffering from obesity and are motivated to lose weight.

2. RELATED WORK

2.1. Fitness Apps in Saudi Arabia

Alnasser et al. examined 65 Arabic fitness apps to determine the level of adherence for each app to the 13 evidence-informed practices [33]. The Centers for Disease Control and Prevention, National Institutes of Health, the Food and Drug Administration and the US Department of Agriculture determined these 13 evidence-informed practices [34]: 1-BMI is determined and explained; 2-fruits and vegetables are recommended and tracked for daily servings; 3-Physical

activities are recommended for daily use; 4-recommendations for drinking water and tracking the daily consumption; 5-recording and tracking the daily consumption of food; 6- a calorie tracker is provided for maintaining calorie balance; 7-advising goal-setting to lose 1 to 2 lb per week; 8-portion control information is provided; 9-Advising users about ways to read and understand nutrition labels; 10- a weight-tracking feature should be provided; 11-physical activities are tracked for daily use; 12-recommending and providing a tool for planning meals; 13-providing a social network among users or allowing users to share via popular social networks, for example Facebook, Twitter, Instagram or Snapchat.

The result of this study stated that there is no app that has more than six evidence-informed practices and only nine apps had between four to six. Therefore, it is clear that there is no Arabic fitness app which successfully adheres to all 13 practices and there is an essential need to address this issue.

2.2. App Selection Process

All of the apps in Google Play or Apple Store are not free. Fitness apps can be divided into three levels:

- At Level 1, apps are free to download but they do not have all the features. The user needs to subscribe and make payments to access extra features;
- At Level 2, the apps are not completely free. The user must pay to download the apps;
- In Level 3 the apps are completely free to be downloaded.

We selected a fitness app to examine its usability and identify how the features in the app affect UX. The app has a high level of adherence to the 13 evidence-informed practices. The app selected is popular in Saudi Arabia and has high ratings on both Google Play and Apple Store so we expect this app to have special features and to be usable. The study of the app will help us determine how the usability in fitness apps can be increased. The app selected is free so that the participants in the usability testing can access them without cost to themselves.

Twazon: We chose Twazon because it is an app developed by academics and it has ten evidence-informed practices out of 13. Another advantage of using the Twazon app is that we can measure the impact of language on the UX. The app is built to make simpler the necessary changes in key diet and exercise behaviour amongst Saudi adults whilst also considering cultural norms. It is also compatible for integration with the Health app on the iPhone. The app is compatible both with Android and iOS operating systems. Twazon has a 4+ rating on both Google Play and Apple Store [35-37].

2.3. Usability

Usability is considered one of the main factors that define the success of a smartphone [38]. Usability can be defined as a multidimensional characteristic of any product. International Standards Organization (ISO) standard 9241-11 gives the meaning of usability as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [39-40]. This definition of usability has been accepted widely [41]. In 2011 it was replaced by ISO/IEC 25010. This form includes a model of software quality that portrays usability as the degree to which a satisfied user can efficiently and effectively attain certain goals under specific conditions. The term UX is used extensively in contrast to usability. The two terms are used interchangeably. However, UX has a much broader meaning than the term usability [42]. It can say that usability is more concerned

with how easy the product and display features are to use. UX includes the user and the product's complete interaction as well as the thoughts, feelings and perceptions that are the results of this interaction [43].

2.4. Usability Models

2.4.1. ISO Usability model

The ISO98 identified three usability attributes, which are effectiveness (demonstrating the level of accuracy and completeness of goal achievement); efficiency (how well resources were used for the sake of effectiveness); and satisfaction (relief, and positive user interaction whilst operating the software). The ISO98 further outlined those usability factors that needed to be considered. These were: user (the person interacting); goal (or main objective); and the background of use (including users, tasks, tool used, and environment). Each one of these factors affects overall how the software will be designed. Specifically, it affects user interaction with the system [39, 44].

Figure 1. IOS usability model

2.4.2. Nielsen Usability Model

Nielsen was one of the first to identify the attributes of usability. Whilst Nielsen's earlier model had only four attributes, which are Effectiveness, Efficiency, Satisfaction and Learnability [45]. However, he later removed Effectiveness and included both Memorability and Errors in his new model. He identified the following attributes [46]:

- Efficiency: Resources used in completing a task accurately to achieve user goals;
- Learnability: The ease with which the system can be learnt so that the user can start to use it to perform tasks in the minimum amount of time;
- Satisfaction: The product should provide comfort and also give the user a positive attitude towards using it;
- Errors: The error rate of the system should be minimal so that the user makes the least number of errors when using the system. If some errors are made, they should be recovered from easily. Finally, catastrophic errors should be avoided;
- Memorability: One should be able to easily memorise the system to the extent that when even a casual user begins using it after a substantial period of time, they do not have to put effort into learning everything from the beginning.

Nielsen's research defines the term utility as how effectively the system can meet user needs. This is not part of the usability but rather it is an entirely separate system attribute. The product

that has no utility for the user lacks the functions and features required. Such a product therefore has a superfluous utility and will not help the user achieve their goals [47].

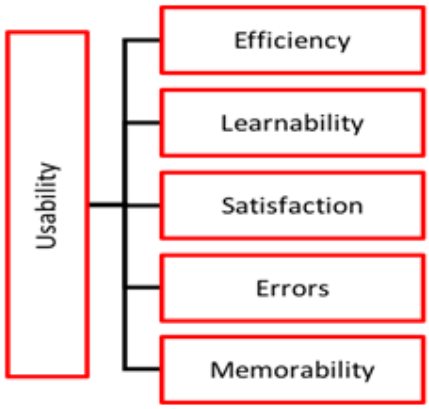


Figure 2. Nielsen usability model

2.4.3. PACMAD Usability Model

This is one of the latest and most frequently used models in recent research on usability. The PACMAD (People At the Centre of Mobile Application Development) model of usability was introduced to overcome issues that have emerged because of the advent of mobile apps. The model aims to include other attributes, which were ignored by other models [47]. Limitations of previous usability models are addressed by PACMAD when they are applied to mobile devices. They included cognitive load in their model because it is likely to have the most significant influence on either the success or failure of an app. The PACMAD model includes attributes for both the ISO and Nielsen models but also incorporates the attributes of both Nielsen's model and ISO standard. The model states that overall usability of a mobile app is affected by 3 factors: task, user, and context of Use. These are recognised by both the ISO and Nielsen. Each of the three factors includes seven attributes. Six are discussed and included in the Nielsen and ISO models. Cognitive load is therefore the new entry and so its inclusion is considered to be PACMAD's main achievement and contribution. Cognitive load is defined as the cognitive processing level that the user requires to use the app. In traditional models, it is assumed that the user is involved in or is performing one task at a time; however, in the context of a mobile device, a user can perform multiple actions whilst using mobile apps. For example, a user might be cooking food whilst he is listening to the stereo.

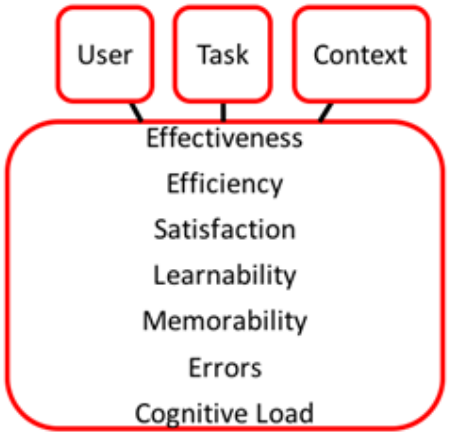


Figure 3. PACMAD usability model

2.5. Discussion

Most of the research discussed enhancing the usability of mobile apps to help people to be motivated and to meet its desired goals. We believe it has missed one very important aspect, which is that of social and cultural norms. For an app to be usable, it should meet with the social and cultural norms for users - be in the users' native language and considering the social customs. Soroa-Koury and Yang recruited 343 participants in their study to find how traditional views and social customs play a part in the prediction of a user's response to mobile apps. The study demonstrated that social norms predicted perceived ease of use (PEOU) and perceived usefulness (PU) [48]. Many researchers in the past have found that social norms affect human behavior [49-51]. Various studies are available which show that social norms have been used to intervene in undesirable behaviors such as drinking, smoking, and sexual proclivity. Researchers investigating the acceptance of technology have incorporated social norms as important predictors of user behavior when adopting a particular technology [52].

Despite the importance of usability, social and cultural norms on the success of a mobile app, there are very few apps in the field of health and fitness that consider their impacts of usability and cultural and social norms when designing the apps. A research by Alnasser et al. was involved the development of 'Twazon'; about which they claim to consider social and cultural norms of Saudi society [35]. However, they have not mentioned any new feature or attribute that makes the app socially more acceptable or more culturally relevant. Mostly, they have used cultural norms as a reason why women are not physically active. The only thing that makes this app culturally and socially aware is the use of Arabic language. For example, the app does not include any special timetable and diet plan for the month of Ramadan. Ramadan is one of the most important religious periods in the Islamic year during which Muslims are not allowed to eat from dawn until dusk. The app uses the Gregorian calendar instead of the Islamic one. Even though the language of the app is Arabic, it still uses the English numbers. Therefore, there are limits to existing research done to make it culturally and socially more relevant and acceptable. Moreover, the app was designed without considering usability attributes and factors. Usability is a very important feature for the success of mobile apps. Therefore, these open issues are identified:

- There is no fitness app available to Saudi users that was designed with the main objective of helping or assisting obese individuals to reach their fitness goals;
- There is no fitness app available to Saudi users that was designed considering usability attributes and factors;
- There is no fitness app available that was designed considering the impacts of cultural and social norms that targeting individuals suffering from obesity in Saudi Arabia.

This leads to the focus of this research:

- How to improve mobile fitness apps usability to help users reach their health and fitness goals and more specifically it discusses how we set a trial to identify;
- What makes mobile fitness apps usable and useful to be easier to use.

3. METHODOLOGY

3.1. Research Strategy

This paper uses experiment as the main research strategy [53]. The purpose of experimental research design is to assist the researcher to establish a cause-effect relationship with a lot of

credibility. Experiments have a particular nature; they are conducted in a systematic way and under controlled conditions. An artificial situation is formed and events, which go together or have something in common are pulled apart [54]. A widely-used definition for experimental research strategy is where scientists actively influence something to observe the consequences [55]. Experimental research strategy can be categorized into [54]:

- Laboratory experiments: These are carried out in settings that are specially created and the experimenter has the ability to control a variety of extraneous variables;

- Natural experiments: These are referred to as quasi –experiments. These studies are conducted when a natural event or social policy creates situations suitable for the experiment. The investigator has no control over independent variables. These subjects are neither matched in groups nor randomly assigned;

- Field experiments: In these experiments, independent variables are manipulated by the researcher in a field environment.

Moreover, laboratory testing has been used in a lot of usability research. Six techniques were outlined and evaluated for usability testing in a laboratory environment. These techniques facilitated systematic collection of data and identified usability problems experienced by mobile users. According to the result form the research, laboratory testing methodology has huge advantages [56]. Additionally, a laboratory testing was used in usability study, which aimed to discover the impacts of the small screen size of mobile devices upon web browsing and navigation [57]. Moreover, laboratory testing allows us to use evaluation techniques such as 'think-aloud' and observation, which cannot be applied through other means. These techniques are not possible in any other research choices such as the field setting [58]. Laboratory testing can reveal a lot about usability; even with the minimum number of participants. According to previous usability guides, 80% of the problems related to usability in any product can be revealed by having four or five participants in the experiment. Similarly, other studies showed that 90% of all usability problems can be detected using ten participants [45, 59-60].

Therefore, in this research, usability testing in a laboratory setting is seen as the most appropriate strategy. A laboratory is a peaceful environment where the users can easily concentrate on the tasks provided to them.

3.2. Usability Metrics Selection

The ISO/IEC 9126-4 makes the recommendation that any usability metric must make reference to effectiveness, efficiency and satisfaction. Attributes such as memorability, errors, cognitive load and learnability are linked to the efficiency and effectiveness of the app. Whilst they each measure the effectiveness and efficiency of these apps, they do so from a specific perspective. If an app has less errors, it means that it is effective because the user can perform more tasks in less time without repeating the tasks with errors. Similarly, if an app has better learnability, it helps the user undertake more tasks accurately so it is more effective. The user therefore becomes more efficient in their completion of these tasks. Any of the above features, improve usability and user satisfaction. This is the reason that usability metrics usually include effectiveness, efficiency and satisfaction as the important features for improving usability.

3.2.1. Usability Metric for Effectiveness

Effectiveness can be measured using the completion rate of tasks. However, another measurement that can be used is the number of mistakes that users make when trying to finish a task.

Effectiveness can therefore be defined as a percentage by utilising the simple equation represented below [61].

$$Effectiveness = \frac{Number\ of\ tasks\ completed\ successfully}{Total\ number\ of\ tasks\ undertaken} \times 100\%$$

3.2.2. Usability Metric for Efficiency

Efficiency is used as a tool to measure the time taken to finish a task. It is usually the time taken by participants to complete a task. Efficiency can be calculated using two methods: Overall Relative Efficiency and Time-Based Efficiency [61].

$$Overall\ Relative\ Efficiency = \frac{\sum_{i=1}^{R}\sum_{i=1}^{n} n_{ij}t_{ij}}{\sum_{i=1}^{R}\sum_{i=1}^{n} t_{ij}} \times 100\% \qquad Time\ Based\ Efficiency = \frac{\sum_{i=1}^{R}\sum_{i=1}^{n} \frac{nij}{tij}}{NR}$$

Where:

- R: number of users
- N: number of tasks.
- nij: result for task (i) by user (j). if the task is completed successfully, then nij = 1, otherwise nij = 0.
- tij = time spent by user "j" to complete task "i". If the user does not complete the task successfully, then the time will be measured until the moment the user gave up from the task.

3.2.3. Usability Metric for Satisfaction

Users' satisfaction can be determined through standardized questionnaires that measure s satisfaction. These can be dispensed after each task or following the usability testing session. Once the user attempts a task, they are given a questionnaire to measure the difficulty of task and the task level satisfaction. Post-task questions can take various forms: ASQ, Subjective Mental Effort Questionnaire (SMEQ), Single Ease Question (SEQ), Usability Magnitude Estimation (UME) etc. From the above list, we will use SEQ as recommended by Sauro [62]. SEQ has the advantage in that it is brief and simple to answer as well as being easy for the experimenter to conduct and then tally the results. The SEQ in this case is "Overall, how easy or difficult did you find this task?". This SEQ has a rating scale of 7 points where 1 is very easy and 7 is very difficult. The level of satisfaction is found via a formalized questionnaire for users to gain an overall idea of how easy the app is to use. There are different types of questionnaires available however the choice depends on the budget as well as the degree of significance placed upon the user's perceived level of satisfaction as a factor of the overall project [63].

3.2.4. Usability Metric for Cognitive Load

Cognitive load has been identified as the measure of mental activity on working memory at any particular instance [64]. To determine the app's cognitive load, we will use the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) test. NASA-TLX allows the user to evaluate the situation of the workload after the testing is done. It measures the overall task demands by identifying 3 broad scales, which are task, behaviour and subject-related. Each of the scales has factors. The task-related scale includes mental, physical and temporal demands. The behaviour-related scale includes performance and effort. Subject related includes frustration. A user will need to have description for each of the factors as demonstrated below [65]:

- Mental demand: To what extent did you need to perform mental and perceptual activities (such as thinking and calculating)?

- Physical demand: To what extent did you need to perform physical activities (such as pushing and pulling)
- Temporal demand: To what extent did you feel a time pressure while performing tasks?
- Effort: How hard did you have to work hard (mentally and physically) to perform tasks?
- Performance: How satisfied are you with your performance?
- Frustration level: How stressed or annoyed did you feel while performing these tasks?

The NASA-TLX test contains two stages which are weights and ratings. In the weighting procedure, a user will be required to evaluate the influence of each factor regarding a task. There are 15 potential pairs of factors about which a comparison is made. A user will be giving 15 cards and each card contains a pair of the factors and asked to select the most relevant factor regarding the task. Each time the user selects from a pair, the examiner counts it. The scale for a factor for each user can range from 0 to 15. The total comparisons for all factors should equal 15. In the second stage, a user needs to rate each of the factors above in a scale that is divided into 20 equal intervals and each interval equals 5 points with a total of a 100 on the scale. As it is a post-event test, it is an effective one as it captures the thoughts and interaction of the user.

3.2.5. Usability Metric for Learnability

Learnability is the ability of the interface to help the user accomplish tasks on the first attempt [66]. Learnability can therefore be measured through establishing the task performance of users who have not been exposed to that app before. Another way of looking at usability is through perceiving how usability or task performance has improved after repeated trials.

3.2.6. Usability Metric for Errors

Another usability measurement is measuring the amount of errors made by the user when completing a task. Errors are defined as mistakes that are made by the participant when attempting a task. Counting the errors provides excellent diagnostic information and it should be mapped into usability problems [67].

3.2.7. Usability Metric for Memorability

Memorability measures how easy it is to remember how to perform a task on the app after the casual user returns to the app after a certain period of not using it [47]. Memorability has the same tests of efficiency and effectiveness but these are repeated after some period of time in order to determine whether the user has remembered how to perform the same task; and hence whether this has improved the usability.

3.3. Usability Testing Environment

The tests were conducted in a typical usability test environment. Laboratory settings were controlled in order to ensure that there were no external interruptions such as varying lighting conditions or disturbing noises. Test sessions were completed via Apple's wireless AirPlay technology. A MacBook was used for recording. The first step was to install Reflector, which is a wireless streaming and mirroring receiver that converted a laptop into an AirPlay receiver. This allowed the user to mirror their smartphone's screen onto their laptop. It also eliminated the need to have an external camera to record events. Moreover, it also helped to minimize the distraction for the user. The purpose of using this software and technology was to create the friendly and quiet environment that is essential for usability testing [68-69].

Nine participants tested the Twazon app. While they tested it, their mobile screens were recorded through Reflector software. All participants were asked to use the app 3 times. Each time, all participants were asked to perform 14 tasks, which were the same for all users. The time difference between the first and second sessions was one hour. Between the second and the third sessions, there was a one-week interval.

The Armed Force Hospital in the Taif Region of Saudi Arabia provided candidates who suffered from obesity and were motivated to lose weight in order to have a healthier life style. The usability test was divided into five phases:

- Introduction: In the first phase, both participants and the examiner introduced themselves. The purpose of the introduction phase was to establish a comfortable interaction between the examiner and participants.

- Warm-up: In this phase, participants were asked to download the app "Twazon" and to fill out a brief questionnaire that aimed to collect participants' information such as gender and age.

- Deep focus: During this phase, the examiner gave the users a list of the 14 tasks. The participants used the app with the focus being on what it was doing; how it worked and how the app could be used. The examiner encouraged the participants to think aloud while they were performing the tasks. Moreover, when participants finished a task, they were asked to rate it in an SEQ questionnaire.

- Retrospective: In the penultimate phase, the examiner explained the NASA-TLX questionnaire and asked participants to fill it out.

- Wrap up: In the final phase, the examiner thanked the participants and answered any enquiries.

4. RESULTS

Researchers examined all the videos that were recorded on mobile screens while the participants were performing in the trial. All users who successfully completed a task scored 1 and at the same time we measured how long it took to complete a task. In contrast, users who completed a task in the wrong way or gave up on a task received 0 and the time taken was measured as well. Then the equations for effectiveness, overall relative efficiency and time-based efficiency were applied. Then all errors that participants had made while performing tasks were calculated. Regarding the learnability attribute, we compared participants' performances in the first session with those of the second. Memorability was then measured by comparing participants' performances in the second session with those of the third. Both satisfaction and cognitive loads were applied only in the first session as they measured the performances of participants who had not previously been exposed to an app. If these loads had been applied in the second and third sessions, this condition could not have been met. Next, we examined the data from the SEQ questionnaire that was used to measure satisfaction. The rating for each user was calculated and then divided by 14 to determine the average satisfaction value for each user. We then examined the data from the NASA-TLX questionnaire and applied the roles to determine the total user score for the cognitive load [65].

Table 1. Participants' information.

Users	Gender	Age group	Occupation	Type of phone
1	Male	35 t0 44	Self employed	iPhone 7
2	Male	25 to 34	Teacher at high school	iPhone 7
3	Female	25 to 34	Unemployed	OnePlus 3
4	Female	45 to 54	Government employee	iPhone 6 S
5	Female	25 to 34	Government employee	HTC 10
6	Female	Prefers not to say	Prefers not to say	iPhone 7
7	Female	25 to 34	Accountant in a company	iPhone 7 Plus
8	Female	25 to 34	Receptionist at a hospital	iPhone 6 S
9	Female	Prefers not to say	Prefers not to say	iPhone 7

Nine participants, seven females and two males, were part of the usability of Twazon app. Their information is presented in Table 1.

4.1. Effectiveness

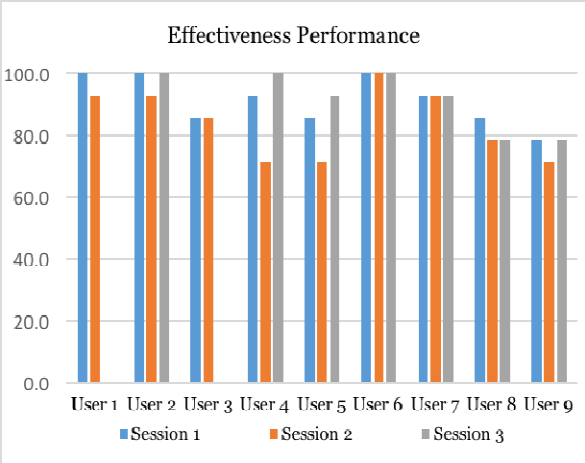


Figure 4. Effectiveness performance

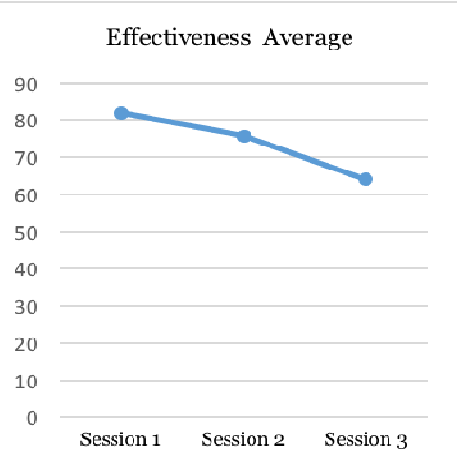


Figure 5. Effectiveness average

Figure 4 describes each user's effectiveness performance over the course of the three sessions. User 6 had the highest percentage of value each time. In session 1, it is 100% and remains constant in session 2 as well as session 3. User 7 showed the same pattern though the value is lower at 92.85%. User 2, user 4, user 5 and user 9 showed positive progress across sessions. However, Users 1 and user 3 had a negative performance because in session 3 they both scored 0%. In addition, User 8's effectiveness performance also slightly decreased. Figure 5 shows the effectiveness performance average, which decreased over each session. In session 1 it was 82%, then it fell to 75.71% and finally in session 3, it reached to 64%.

4.2. Efficiency

4.2.1. Overall Relative Efficiency

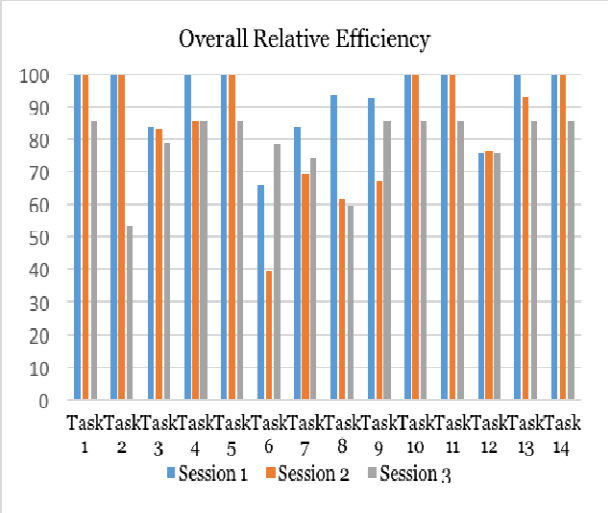


Figure 6. Overall relative efficiency

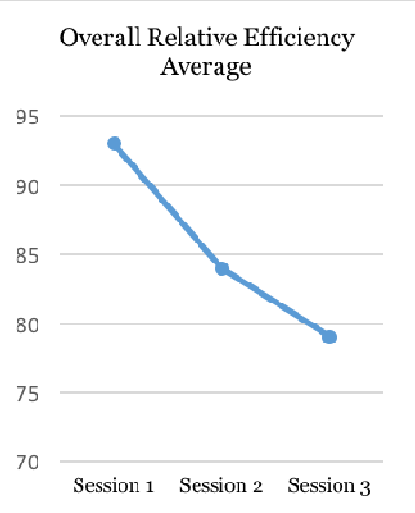


Figure 7. Overall relative efficiency average

Figure 6 demonstrates overall relative efficiency for tasks for the three sessions. In session 1, among the 14 tasks, 8 tasks scored 100% whereas in session 2 and session 3 it was only 6 and 0 respectively. Only in tasks 6 and 12 did the overall relative efficiency percentage improve over each session. However, all the other tasks dramatically decrease between the first and final sessions. Figure 7 shows the overall relative efficiency average, which decreased over each of the three sessions. In session 1 it was 93%, then it fell to 84% and in the final session it reached 79.03%.

4.2.2. Time-Based Efficiency

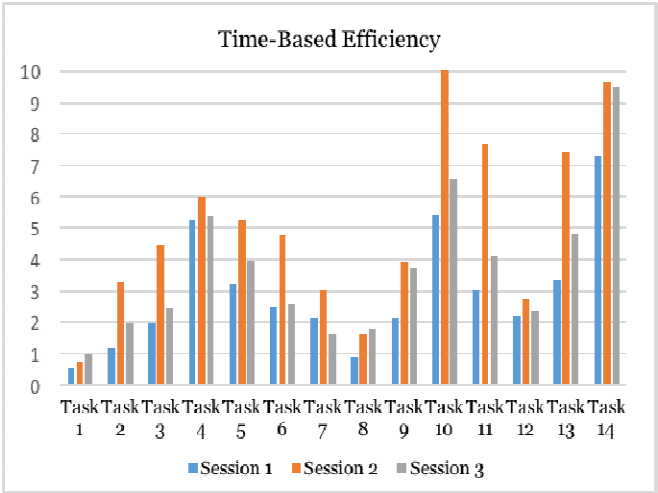


Figure 8. Time-based efficiency

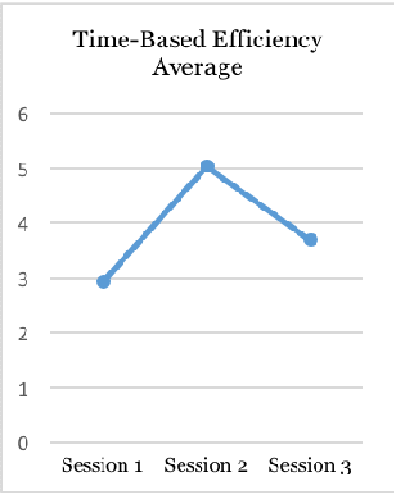


Figure 9. Time-based efficiency average

Figure 8 states time-based efficiency for tasks among the sessions. Task 14 had the highest time-based efficiency score among tasks. In sessions 1, 2 and 3 it was 7.28 goals/sec, 9.66 goals/sec

and 9.48 goals/sec respectively. Task 10 had the second highest time-based efficiency score followed by task 4 and task 13. Interestingly, task 10 reached 10.02 goals/sec in session 2, which was the highest value in all sessions. On the other hand, task 1 got the lowest time-based efficiency followed by task 8 and task 2. Figure 9 shows the time-based efficiency average, which fluctuated across sessions. In session 1 it was 2.93 goals/sec, then it increased to 5.03 goals/sec and finally in the session 3 it decreased to 3.69 goals/sec.

4.3. Satisfaction

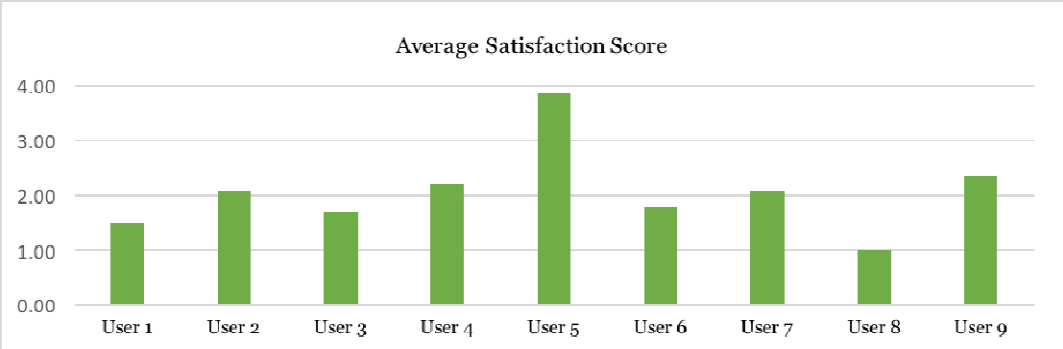


Figure 10. Average satisfaction score

Figure 10 shows each user's average satisfaction score for all tasks. User 5 had highest score at 3.86. User 9 and user 4 scored 2.36 and 2.21 respectively. However, User 8 and user 1 had the lowest score at 1.00 and 1.50 respectively.

4.4. Cognitive Load

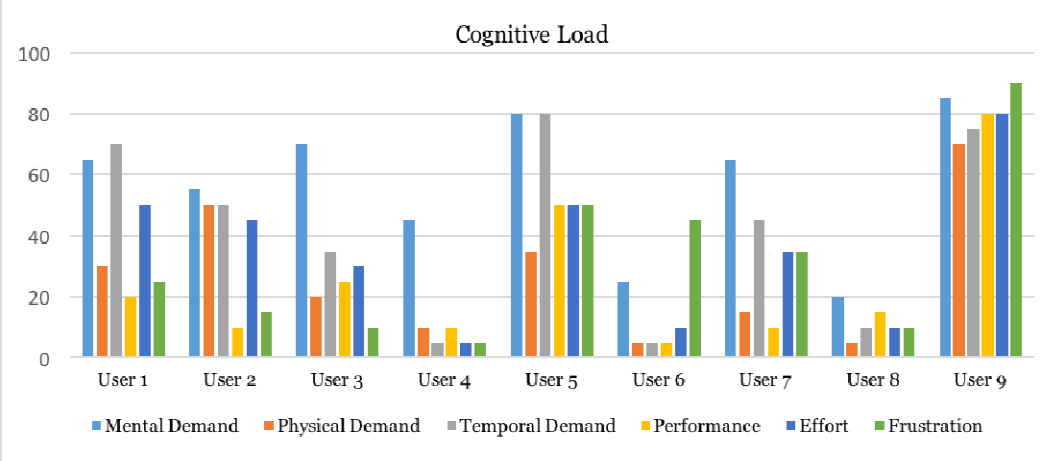


Figure 11. Users' rating for each subscale in cognitive load

Figure 11 shows each user's rating for each subscale in the cognitive load. User 9's cognitive loading is the most consistent. Scores lie between physical demand at (70%) to frustration (90%). However, between user 4 and user 6, the score gap is too high. Mental demand and temporal demand scored the highest value amongst all the subscales. On the other hand, performance and physical demand scored the lowest values.

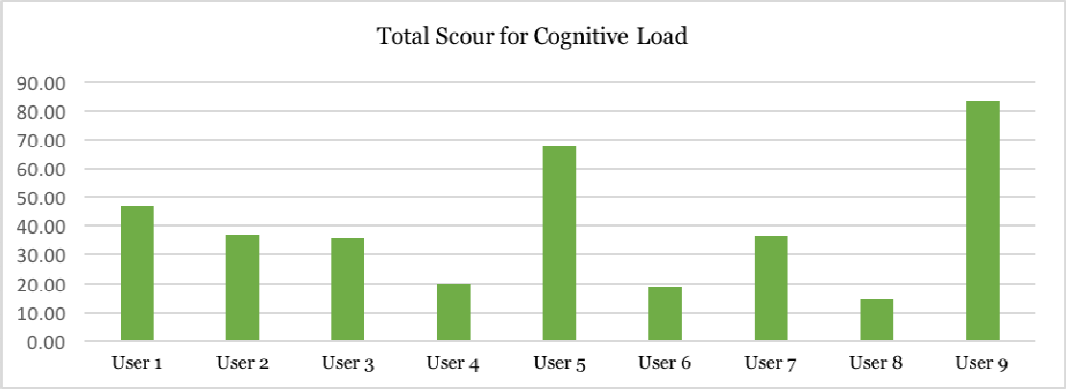


Figure 12. Total score for cognitive load

Figure 12 refers to the total score for cognitive load amongst users. User 9 had the highest value at 83.33%. User 5 and user 1 scored 68% and 47% respectively. However, user 8 had the lowest score at 14.33%.

4.5. Errors

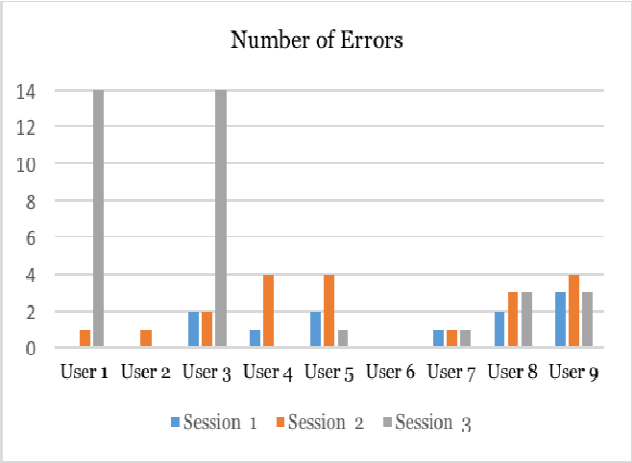


Figure 13. Number of errors

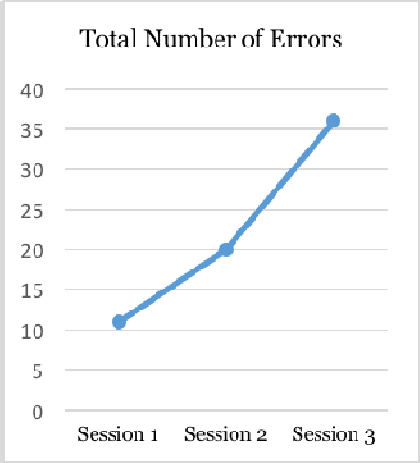


Figure 14. Total number of errors

Figure 13 shows the number of errors made by each user. User 6 is the only user who did not make any errors in all three sessions. User 2 has the second lowest number of errors with just one in session 2. However, User 3 and user 1 had the highest number of errors at 18 and 15 respectively. Figure 14 demonstrates the total number of errors made by all users, which increased over each session. In session 1 it was 11, then it sharply increased to 20 and finally in the third session, it increased to 36.

4.6. Discussion on the Results

One limitation of this study was that users 1 and 3 were not able to participate in the third session as they faced a technical issue with the app. The app did not respond to them when they started performing the first task and after several attempts, they gave up. However, the overall trial for testing the usability of the app succeeded as the level of usability was determined.

Despite the positive increase in the overall score for time-based efficiency between session 1 and session 3, the percentage score for user's effectiveness and overall relative efficiency decreased

over time. Moreover, the number of errors increased from the first session to the second session and did so again from the second to third sessions. As a result of this, the app had a negative association with both learnability and memorability attributes. Furthermore, several participants scored a high percentage in the satisfaction questionnaire, which is negative as a high score means it was more difficult and only one participant rated the whole task as very easy and scored 1 as an average. Besides this, overall cognitive load score was high as the lowest percentage scored by a participant was 14.33%, which means that several participants were not able to perform tasks correctly while doing other activities; for example, speaking to examiners.

The five usability attributes (effectiveness, efficiency, learnability, memorability and errors) did not improve over time. Moreover, both satisfaction and cognitive load scored high percentages because the majority of participants found the app difficult to use. Therefore, the results state that Twazon app has a low level of usability, which is expected due to the fact that it was designed and developed without considering usability attributes and factors.

It is recommended that conducting a qualitative study to determine the reasons and factors that negatively affect the level of usability of the Twazon app. The qualitative study will also consider the importance of social and cultural norms and how they can be applied to improve the usability of the app. A specific usability guideline for fitness mobile apps will then be created, which will help to develop a fitness app that is specially designed for obese individuals in Saudi Arabia.

5. CONCLUSION

The primary purpose of this paper has been to examine the usability for an Arabic fitness mobile app “Twazon”. This paper has highlighted the attributes that are considered to be a crucial for improving the usability of fitness mobile apps through presenting an extensive literature review. The paper has presented the methodology and the procedures for testing the Twazon app. Seven usability attributes, (effectiveness, efficiency, satisfaction, memorability, errors, learnability and cognitive load) were tested. The trial for the test was done in collaboration with the Armed Forces Hospitals - Taif Region in Saudi Arabia, which provides the candidates. The result from this trail stat was that Twazon app failed to meet with the usability attributes and consequently participants found it difficult to use. Future work will include performing a qualitative study for the app to determine how to improve the level of usability and then create usability guidelines for fitness mobile apps. Based on these guidelines, an app that is specifically designed for obese individuals in Saudi Arabia will be developed. Obesity is a major issue for health departments all over the world. Saudi Arabia is a country where the obesity has reached an alarming rate of 35.5% of the population. Better app usability would help keep these individuals motivated to make necessary lifestyle changes.

REFERENCES

[1] A. P. SIMOPOULOS and T. B. VAN ITALLIE, "Body weight, health, and longevity," Annals of internal medicine, vol. 100, pp. 285-295, 1984.
[2] W. H. Organization. (2016, 2 October). Obesity and overweight. Available: http://www.who.int/mediacentre/factsheets/fs311/en/
[3] O. R. Center. (2016, 10 October). Obesity in Saudi Arabia. Available: https://www.obesitycenter.edu.sa/pages/patients.aspx?id=258
[4] A. Afshin, M. H. Forouzanfar, M. B. Reitsma, P. Sur, K. Estep, A. Lee, et al., "Health Effects of Overweight and Obesity in 195 Countries over 25 Years," The New England journal of medicine, vol. 377, pp. 13-27, 2017.
[5] K. Singer and C. N. Lumeng, "The initiation of metabolic inflammation in childhood obesity," The Journal of clinical investigation, vol. 127, pp. 65-73, 2017.

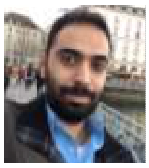
[6] K. R. Fontaine, D. T. Redden, C. Wang, A. O. Westfall, and D. B. Allison, "Years of life lost due to obesity," Jama, vol. 289, pp. 187-193, 2003.
[7] J. Stevens, J. Cai, E. R. Pamuk, D. F. Williamson, M. J. Thun, and J. L. Wood, "The effect of age on the association between body-mass index and mortality," New England Journal of Medicine, vol. 338, pp. 1-7, 1998.
[8] E. E. Calle, M. J. Thun, J. M. Petrelli, C. Rodriguez, and C. W. Heath Jr, "Body-mass index and mortality in a prospective cohort of US adults," New England Journal of Medicine, vol. 341, pp. 1097-1105, 1999.
[9] C. Summerbell, E. Waters, L. Edmunds, S. Kelly, T. Brown, and K. Campbell, "Interventions for preventing obesity in children (Review)," Cochrane library, vol. 3, pp. 1-71, 2005.
[10] W. Saris, S. Blair, M. Van Baak, S. Eaton, P. Davies, L. Di Pietro, et al., "How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement," Obesity reviews, vol. 4, pp. 101-114, 2003.
[11] O. Bar-Or, "Juvenile obesity, physical activity, and lifestyle changes: Cornerstones for prevention and management," The physician and sportsmedicine, vol. 28, pp. 51-58, 2000.
[12] J. L. Anderson, E. M. Antman, S. R. Bailey, E. R. Bates, J. C. Blankenship, D. E. Casey Jr, et al., "AHA Scientific Statement," Circulation, vol. 120, pp. 2271-2306, 2009.
[13] J. O. Hill and H. R. Wyatt, "Role of physical activity in preventing and treating obesity," Journal of Applied Physiology, vol. 99, pp. 765-770, 2005.
[14] J. O. Hill and J. C. Peters, "Environmental contributions to the obesity epidemic," Science, vol. 280, pp. 1371-1374, 1998.
[15] I. Contento, G. I. Balch, Y. L. Bronner, L. Lytle, S. Maloney, C. Olson, et al., "The effectiveness of nutrition education and implications for nutrition education policy, programs, and research: a review of research," Journal of nutrition education (USA), 1995.
[16] G. D. Foster, A. P. Makris, and B. A. Bailer, "Behavioral treatment of obesity," The American journal of clinical nutrition, vol. 82, pp. 230S-235S, 2005.
[17] T. A. Wadden and A. J. Stunkard, Handbook of obesity treatment: Guilford Press, 2002.
[18] K. D. Brownell, LEARN program for weight management 2000: American Health, 2000.
[19] J. Yang, "Toward physical activity diary: motion recognition using simple acceleration features with mobile phones," in Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics, 2009, pp. 1-10.
[20] T. Denning, A. Andrew, R. Chaudhri, C. Hartung, J. Lester, G. Borriello, et al., "BALANCE: towards a usable pervasive wellness application with accurate activity inference," in Proceedings of the 10th workshop on Mobile Computing Systems and Applications, 2009, p. 5.
[21] S. M. Arteaga, M. Kudeki, A. Woodworth, and S. Kurniawan, "Mobile system to motivate teenagers' physical activity," in Proceedings of the 9th International Conference on Interaction Design and Children, 2010, pp. 1-10.
[22] D. E. Conroy, C.-H. Yang, and J. P. Maher, "Behavior change techniques in top-ranked mobile apps for physical activity," American journal of preventive medicine, vol. 46, pp. 649-652, 2014.
[23] F. Raben and E. Snip, "The MENAP region is developing, but can it keep its promise?," Research World, vol. 2014, pp. 6-11, 2014.
[24] Statista. (2017, 8 March). Number of smartphone users in Saudi Arabia from 2014 to 2021 (in millions)*. Available: https://www.statista.com/statistics/494616/smartphone-users-in-saudi-arabia/
[25] D. LLP. (2012, 2 February). So Many Apps -- So Little To Download. Available: http://www.mondaq.com/x/192692/IT+internet/So+Many+Apps+So+Little+To+Dow%20nload
[26] S. Dredge, "Most branded apps are a flop says Deloitte. But why," ed, 2011.
[27] M. Bhuiyan, A. Zaman, and M. H. Miraz, "Usability Evaluation of a Mobile Application in Extraordinary Environment for Extraordinary People," arXiv preprint arXiv:1708.04653, 2017.
[28] R. Youens. (2011, 2 February). 7 Habits of Highly Effective Apps. Available: https://gigaom.com/2011/07/16/7-habits-of-highly-effective-apps/
[29] I. Nascimento, W. Silva, A. Lopes, L. Rivero, B. Gadelha, E. Oliveira, et al., "An Empirical Study to Evaluate the Feasibility of a UX and Usability Inspection Technique for Mobile Applications," in 28th International Conference on Software Engineering & Knowledge Engineering, California, USA, 2016.
[30] H. Hoehle, R. Aljafari, and V. Venkatesh, "Leveraging Microsoft′ s mobile usability guidelines: Conceptualizing and developing scales for mobile application usability," International Journal of Human-Computer Studies, vol. 89, pp. 35-53, 2016.

[31] S. Pagoto, K. Schneider, M. Jojic, M. DeBiasse, and D. Mann, "Evidence-based strategies in weight-loss mobile apps," American journal of preventive medicine, vol. 45, pp. 576-582, 2013.
[32] A. C. King, E. B. Hekler, L. A. Grieco, S. J. Winter, J. L. Sheats, M. P. Buman, et al., "Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults," PloS one, vol. 8, p. e62613, 2013.
[33] A. A. Alnasser, R. E. Amalraj, A. Sathiaseelan, A. S. Al-Khalifa, and D. Marais, "Do Arabic weight-loss apps adhere to evidence-informed practices?," Translational behavioral medicine, vol. 6, pp. 396-402, 2016.
[34] E. R. Breton, B. F. Fuemmeler, and L. C. Abroms, "Weight loss—there is an app for that! But does it adhere to evidence-informed practices?," Translational behavioral medicine, vol. 1, pp. 523-529, 2011.
[35] A. Alnasser, A. Sathiaseelan, A. Al-Khalifa, and D. Marais, "Development of 'Twazon': An Arabic App for Weight Loss," JMIR research protocols, vol. 5, p. e76, 2016.
[36] F. Al-Maarik. (2016, 5 January). Twazon. Available: https://itunes.apple.com/us/app/twazon/id946772876?mt=8
[37] G. Play. (2016, 5 January). Twazon Available: https://play.google.com/store/apps/details?id=com.twazon.twazon
[38] R. Baharuddin, D. Singh, and R. Razali, "Usability dimensions for mobile applications—A review," Res. J. Appl. Sci. Eng. Technol, vol. 5, pp. 2225-2231, 2013.
[39] W. ISO, "9241-11. Ergonomic requirements for office work with visual display terminals (VDTs)," The international organization for standardization, vol. 45, 1998.
[40] S. Ben and C. Plaisant, "Designing the user interface 4 th edition," ed: Pearson Addison Wesley, USA, 2005.
[41] E. Folmer and J. Bosch, "Architecting for usability: a survey," Journal of systems and software, vol. 70, pp. 61-78, 2004.
[42] D. Saffer, "Designing for Interaction: Creating Smart Applications and Clever Devices," New Riders Press,< http://www. designingforinteraction. com, vol. 2, p. 2.1, 2007.
[43] W. Albert and T. Tullis, Measuring the user experience: collecting, analyzing, and presenting usability metrics: Newnes, 2013.
[44] C. Stary and C. Stephanidis, User-Centered Interaction Paradigms for Universal Access in the Information Society: 8th ERCIM Workshop on User Interfaces for All, Vienna, Austria, June 28-29, 2004. Revised Selected Papers vol. 3196: Springer Science & Business Media, 2004.
[45] J. Nielsen, Usability engineering: Elsevier, 1994.
[46] J. NIELSEN. (2012, 18 November). Usability 101: Introduction to Usability. Available: https://www.nngroup.com/articles/usability-101-introduction-to-usability/
[47] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model," Journal of Interaction Science, vol. 1, pp. 1-16, 2013.
[48] S. Soroa-Koury and K. C. Yang, "The Effects of Social Norms on Consumers' Responses to Mobile Advertising," in Proceedings of the 2009 Academy of Marketing Science (AMS) Annual Conference, 2015, pp. 162-166.
[49] M. K. Lapinski and R. N. Rimal, "An explication of social norms," Communication Theory, vol. 15, pp. 127-147, 2005.
[50] G. L. Cohen and D. K. Sherman, "The psychology of change: Self-affirmation and social psychological intervention," Annual Review of Psychology, vol. 65, pp. 333-371, 2014.
[51] R. N. Rimal, M. K. Lapinski, R. J. Cook, and K. Real, "Moving toward a theory of normative influences: How perceived benefits and similarity moderate the impact of descriptive norms on behaviors," Journal of health communication, vol. 10, pp. 433-450, 2005.
[52] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," Management science, vol. 46, pp. 186-204, 2000.
[53] M. Saunders, P. Lewis, and A. Thornhill, "Research methods for business students," Harlow: Prentice Hall, 2009.
[54] M. Rao and S. Shah. (2002, 12 December). EXPERIMENTATION A RESEARCH METHODOLOGY. Available: http://www.public.asu.edu/~kroel/www500/EXPERIMENTATION%20Fri.pdf
[55] O. Blakstad. (2008, 3 January). Experimental Research. Available: https://explorable.com/experimental-research
[56] E. Beck, M. Christiansen, J. Kjeldskov, N. Kolbe, and J. Stage, "Experimental evaluation of techniques for usability testing of mobile systems in a laboratory setting," 2003.

[57] A. Parush and N. Yuviler-Gavish, "Web navigation structures in cellular phones: the depth/breadth trade-off issue," International Journal of Human-Computer Studies, vol. 60, pp. 753-770, 2004.
[58] N. Sawhney and C. Schmandt, "Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments," ACM transactions on Computer-Human interaction (TOCHI), vol. 7, pp. 353-383, 2000.
[59] J. Rubin, "Handbook of usability testing: how to plan, design, and conduct effective tests," Wiley technical communication library Show all parts in this series, 1994.
[60] J. S. Dumas and J. Redish, A practical guide to usability testing: Intellect Books, 1999.
[61] J. Mifsud. (2015, 3 November). Usability Metrics – A Guide To Quantify The Usability Of Any System. Available: http://usabilitygeek.com/usability-metrics-a-guide-to-quantify-system-usability/
[62] J. Sauro. (2010, 9 December). IF YOU COULD ONLY ASK ONE QUESTION, USE THIS ONE. Available: https://measuringu.com/single-question/
[63] A. Garcia. (2013, 18 October). UX Research | Standardized Usability Questionnaire. Available: https://chaione.com/blog/ux-research-standardizing-usability-questionnaires/
[64] J. P. Tracy and M. J. Albers, "Measuring cognitive load to test the usability of web sites," in Annual Conference-society for technical communication, 2006, p. 256.
[65] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," Advances in psychology, vol. 52, pp. 139-183, 1988.
[66] J. Sauro. (2013, 1 December). HOW TO MEASURE LEARNABILITY. Available: https://measuringu.com/measure-learnability/
[67] J. Sauro. (2011, 4 December). 10 ESSENTIAL USABILITY METRICS. Available: https://measuringu.com/essential-metrics/
[68] C. Walsh. (2015, 6 November). A Guide To Simple And Painless Mobile User Testing. Available: https://www.smashingmagazine.com/2015/12/simple-and-painless-mobile-user-testing/
[69] J. Mifsud. (2016, 3 November). Usability Testing Of Mobile Applications: A Step-By-Step Guide. Available: http://usabilitygeek.com/usability-testing-mobile-applications/

AUTHORS

Ryan Alturki is a Teaching Assistant in the department of Information Sciences at Umm Al Qura University, Saudi Arabia. Currently, he is doing a PhD in mobile application usability and its role to motivate people to lose more weight.



Valerie Gay has more than 25 years of research experience in leading research labs in Australia and Europe; Her research focuses on the use of mobile technology to offer more personalised advice and care.