Assessing the Usability of M-Health Applications: A Comparison of Usability Testing, Heuristics Evaluation and Cognitive Walkthrough Methods

Obead Alhadreti
Dept. of Computers-Engineering and Computing College, Umm Al-Qura University, Al-Qunfudah, Saudi Arabia

***Abstract*—Mobile health applications have increasingly become an important channel for providing services in the health sector. However, poor usability can be a major barrier for the rapid adoption of mobile services. The purpose of this study is to compare the relative performance of three usability evaluation methods, namely, usability testing, heuristics evaluation, and the cognitive walkthrough methods in determining the usability level of mobile health applications. The study also explores the relationship between the metrics of usability testing and the current level of mobile health applications in Saudi Arabia. An experimental approach has been used in this study, which gathered qualitative and quantitative data. The methods were used to assess two mobile health interfaces and were compared on the number, severity, and types of usability problems identified. Correlation tests were also carried out to examine areas of overlap between usability testing metrics. The heuristic evaluation found significantly greater numbers of usability problems than the other techniques. The usability testing method, however, detects problems of greater severity. There is also a significant correlation between the number of usability issues found and how long it takes to perform tasks in usability tests. Moreover, the level of usability of the Saudi applications tested is below expectation and in need of further improvement. Based on the study results, both usability testing and heuristic evaluation should be employed during the design process of mobile health applications for maximum effectiveness. Additionally, it is recommended that SUS questionnaires should not be the sole method of determining the usability level of mobile health applications.**

Keywords*—Mobile health applications; usability; usability testing; heuristics evaluation; cognitive walkthrough

I. Introduction

Digitalization has come to play a prominent role in delivering health services to individuals and communities. It not only improves patient safety and satisfaction but is instrumental in keeping large-scale health statistics up to date. Hospitals and other healthcare service providers are offering more digitalized services, which has fundamentally altered healthcare systems. Among these services are mobile health (m-health) applications, which are becoming a major avenue of healthcare provision, given that approximately six billion people (around 75% of the world's population) have regular access to mobile phones [1]. As a result, m-health is now a rapidly expanding field of research.

The term, m-health refers in general to the use of mobile devices in the provision of healthcare services [2]. The Global Observatory for e-health defines m-health as "medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants (PDAs), and other wireless devices" [3]. M-health applications also include processes of data collection [4], service delivery [5], communication between doctors and patients [6] and support for monitoring and adherence in real time [7]. The market for m-health applications is expected to grow in coming years at a significant rate: from USD 99 billion globally in 2021 to USD 332.7 billion by 2025 [8]. The scope for the adoption of m-health applications is further increased by their diversification into such health sectors as nutrition, sports, productivity and behavioral therapy [9].

A crucial requirement for m-health applications is usability. An information system that people cannot use easily represents a threat to the safety of patients, as well as being inefficient and a contributor to staff burn-out and dissatisfaction. An easy-to-use system, on the other hand, is more efficient, enhances emergency care safety and is a real benefit to staff [10]. Usability is generally considered as "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" [11]. It is clear from this definition that usability (real and perceived) can differ between contexts, target audiences and products, which is especially true in the m-health sector due to its unique characteristics which are as follows. First, unlike general commercial applications, which can use personalized messages to seduce customers into feeling comfortable with the system, it is difficult to establish user satisfaction with m-health applications, which have to give both positive and negative messages about users' health-related behavior. For example, appropriate advice (such as switching off the television and going for a walk instead) may not be what users wish to hear, which has the potential to affect their satisfaction with the system. Second, m-health communication needs to be tailored to individual users' knowledge levels and awareness regarding health, which can vary significantly from one user to the next [10]. Third, these factors are exacerbated when someone has a chronic illness, as this can increase anxiety and stress, which makes the assimilation of information and self-management skills more difficult [11].

The assessment of mobile applications' usability is challenging due to the small screens on which they are viewed and the resolution of their displays, limited input options, restricted processing speeds and power, and connectivity issues [12]. Several methods exist for assessing the usability of mobile applications. The most frequently employed usability evaluation methods (UEMs) are usability testing (UT), heuristic evaluation (HE), and the cognitive walkthrough (CW) [13]. The UT method is a user-based method that is widely used to measure how easily end–users can use an interface. However, recruiting test participants and performing tests can be an expensive process. HE is a usability inspection method that involves having an expert examines an interface against a set of principles. These principles provide a template to help identify issues a user will likely encounter. One of the limitations of HE is that it tends to detect many low-severity problems. CW is also a reviewer-based method, but the emphasis is on tasks. The idea is to identify users' goals, and how they achieve them in the interface, then experts detect issues users would encounter as they learn to use the system [13]. So far, there has not been any research which compares UT, HE, and CW methods in terms of their performance in measuring m-health application usability. This study, therefore, aims to examine the effectiveness of these UEMs in the context of m-health applications used in Saudi Arabia.

The Saudi healthcare sector is undergoing a digital transformation, including an increasing dependency on m-health applications for expanding access to healthcare, health education, communicating with patients, monitoring their conditions and ensuring conformity to treatments. Various m-health applications have been promoted by the Ministry of Health in Saudi Arabia, such as Sehhaty1. These applications have several purposes, such as booking appointments, remote consultations and delivery of medicines, and they became particularly important in the course of the Covid-19 pandemic, when there was a significant increase in Saudis using m-health applications [14], although this was down to necessity, not their actual interest in using the technology. Indeed, research has shown that it was factors such as stress, fear and depression which led to the rapid adoption of m-health applications, not such reasons as usefulness, ease of use, enjoyability or self-interest [15]. It is therefore crucial to establish whether or not users of m-health applications are satisfied and whether they are continuing to make use of them after the pandemic.

The current study's findings contribute significantly to the research literature on the usability evaluation of m-health apps. This will help usability practitioners to make more informed decisions about which of the examined methods to use and in which context. The findings will also be of value to the governmental and non-government organizations providing healthcare in Saudi Arabia. The rest of this paper is structured as follows. Section II reviews related work. Later sections present the study's methodology, data analysis and its results. The final section sets out the conclusions drawn from the study.

II. Related Work

A number of studies have compared the effectiveness of the UT, HE, and CW evaluation methods across different systems. A study by Tan et al., [16] compared UT and HE and found that HE identified a larger number of problems and more severe issues than UT. However, the study discovered that UT found problems missed by HE. Another study conducted by Hasan et al., [17] indicated that HE identified 72% of problems, UT extracted only 10% and 18% of the problems were found by both techniques. A study by Doubleday et al. [18] revealed that 40% of issues detected were unique to HE, whereas 39% were extracted by the UT method. Jeffries et al, [19] stated that the HE method found approximately three times more issues than UT; but UT found more important problems.

Thankam et al. [20] contrasted the performance of HE with UT to find out which usability issues were revealed by both methods. The comparison was conducted on four dental computer-based patient record interfaces. 50% of the issues were identified by each method. The study recommended that HE can be a useful tool to assess design early in the development process. In a paper by Khajouei et al. [21], the HE and CW evaluation methods were assessed based on the number and severity of the problems extracted and the ISO and Nielsen usability attributes. The number of issues related to the "satisfaction" attribute detected by HE was significantly higher than those identified using CW. However, CW identified a greater number of problems concerning the "learnability" attribute. In addition, Maguire and Isherwood examined the results of UT and HE, and it was found that HE detected approximately five times more problems than UT, thus it could be seen as more effective.

Few studies have explored the user-friendliness level of Saudi m-health applications. AlanziI [8] found that Saudi users were reasonably satisfied. Furthermore, Arafa et al. [22] studied barriers to the use of Saudi m-health applications, along with their personalization and usability, and found that usability scores were low, whereas Shilbayeh and Ismail found an average usability score of 76.8% for the CATA mobile application overall, suggesting general satisfaction [23]. However, most of the research which has been carried out have the significant limitation of having used only subjective data (e.g. from questionnaires), which was not validated against such objective data as expert inspections or task performance in the context of usability testing. It is, therefore, important that this study sheds a light on the usability levels of the m-health applications in use in Saudi Arabia.

The study's research questions are therefore as follows.

- Is there a difference between the performance of UT, HE, and CW methods in evaluating m-health applications?
- Is there a correlation between UT metrics?
- What is the current usability level of m-health applications in the Kingdom of Saudi Arabia?

III. Methodology

A. Study Approach and Variables

This study adopted an experimental approach, using both quantitative and qualitative data collection techniques [24]. The type of evaluation method (UT, HE or CW) formed the independent variable for the study and there were three dependent variables measured: number of usability issues

[1] https://apps.apple.com/sa/app/%D8%B5%D8%AD%D8%AA%D9%8A-sehhaty/id1459266578

detected, problem severity and problem type. Data on participants' task performance and their satisfaction with the usability of the test system was also gathered in the UT sessions to explore how these outcome variables are related to other metrics.

B. Test Objects and Tasks

From a careful assessment of m-health applications used in Saudi Arabia, two were selected as test subjects. It was decided to assess two m-apps instead of one to gain more reliable results pertaining to the UEMs' effectiveness. The two applications were chosen because they had a broad user base, which simplified the participant recruitment process, whilst also making the sample more likely to be representative of actual users. The two applications also had many similarities, which facilitated the formulation of tasks which resembled each other in terms of focus and difficulty. One test subject, App A, was developed by the ministry of health; the other, App B, comes from the private sector. They both have similar features, allowing the booking of GP appointments, the viewing of laboratory reports and health condition monitoring. The names of the applications are anonymized for confidentiality.

An independent usability expert, with extensive health system knowledge, performed a preliminary examination of both applications (and was not involved in later stages of the study). This examination was mainly to ensure that the two test subjects were of sufficient complexity and had sufficient scope for user interaction and the emergence of problems, although the expert did not make any predictions concerning potential problems, nor did she report any.

An analysis of the context of use was then carried out for each application to identify the characteristics of a representative user and select appropriate tasks [25]. Next, each application was examined to reveal typical use cases in order to set the tasks. 15 tasks were then formulated for each application, covering varying degrees of difficulty, as a long list from which the actual tasks set in this study were subsequently selected by the aforementioned independent expert. Equivalence of difficulty between the tasks set for each application was ensured by matching tasks according to the level and depth of their solution within each app. Five tasks were set for each application and their design was done carefully to avoid any bias related to task-related cues or language. These were then piloted by three representative users prior to the main test sessions. Two task examples are given below.

You wish to book an appointment with your general practitioner. What would you do? (App A)

You wish to seek a remote consultation from your doctor. What would you do? (App B)

C. Participants

For the UT evaluation, statistical validity was ensured by recruiting 20 participants [26]. For the HE and CW evaluations, which are done by experts, between three and five evaluators are generally considered sufficient [27], thus four usability experts were recruited for each of the HE and CW tests. 28 participants were therefore included in this study in total. They were recruited by means of convenience and snowball sampling [24].

All of the participants were native Arabic speakers and, for the UT evaluation, averaged 22 years of age (ranging from 18 to 26). All of them had more than five year's daily use of mobile applications and almost 95% had used an m-health application previously, but not the ones under evaluation.

For the HE and CW evaluations, the two evaluator groups were matched with respect to general HCI knowledge and their expertise in relation to user interface design and usability of m-health applications specifically. Of the evaluators, two in each UEM test (four in total) had a PhD related to HCI. The other four had an HCL-related MSc. They all had extensive experience of conducting evaluations by HE and CW, and all had at least seven years' experience of usability evaluation. All of the participants affirmed their informed consent in writing before the study commenced and none were offered, nor received any incentive for their involvement.

D. Experimental Procedure

Due to the risk of a participant being influenced in a second test by their experience in the first [24], a two-week break was inserted between the evaluation sessions. In addition, half of each participant group used App A in the first session and App B in the second, with the other half doing the reverse. The same type of mobile phone was used by all participants (iPhone 15), chosen because of it being the most common type in Saudi Arabia [28].

1) UT evaluation: The setting for the usability testing of both applications in this study was a laboratory. The participants were asked to make themselves familiar with the phone used and then to perform an initial pilot task. After that they were asked to read a sheet of task instructions before setting out to complete the five tasks, which were presented in a different order to each participant to control for any effect on results of task order [26].

The performance measures for the UT condition were 1) the rate of completion of each task, 2) the time each task took to complete and 3) navigational behaviour (such as how many clicks were made and which screens were browsed). After completing all five tasks, the participants were invited to watch a muted recording of their performance and give a retrospective commentary. Participants then completed a System Usability Scale (SUS) [18], which is designed to assess user satisfaction with application usability. Subsequently, all of the test data were reviewed to extract evidence of usability issues.

2) HE evaluation: This study followed the HE procedure as recommended in the work of Nielsen and Molich [29]. 1) a list of ten heuristic principles was distributed to the evaluators as a guideline by which each evaluator then evaluated the user interface, independently. The evaluators all presented a list of the problems they had identified with the system's usability, each with a description, including its frequency, persistence and likely impact on the user, and illustrative screenshots. They were, however, instructed not to share their thought with one another during the session, as one evaluator might miss several problems and each evaluators may identify a broad spectrum of unique problems [24]. The results tend to be more comprehensive, therefore, when the findings of several evaluators are combined. However, once the independent evaluations were over, the evaluators were asked to collaborate on producing one list of usability issues. That done, each evaluator estimated the severity of every problem detected and classified it by type. They all met finally to determine average severity of the problems identified and the type classifications [29].

3) CW evaluation: In the CW condition, the applications were assessed following Blackmon et al.'s methodology [30]. The evaluators used the five tasks from the UT evaluation and answered the following questions as they did so.

- Will users attempt to achieve the correct result?
- Will users see that a necessary action is available to them?
- Will users connect the desired result with the action necessary to achieve it?
- Are users given confirmation that they have made progress towards the desired result once a necessary action has been carried out?

The overall user goals and the requisite subgoals for each task were determined and the necessary actions were identified in the way illustrated by Blackmon et al. [30]. The evaluators then examined each task systematically, noting 1) the goals users are expected to seek, 2) their subgoals and 3) actions, 4) the responses from the application and 5) potential problems with user interaction. Each evaluator produced a list of problems independently and recorded information about the issue in the same way as for the HE evaluation. Following the completion of each task, the list of usability issues identified was reviewed by the evaluator, adding or correcting items if necessary. The five task-specific lists from each evaluator were gathered and compared communally and a consolidated list of issues was established. As with the HE evaluation, the evaluators determined the types and severity of problems independently, before meeting to agree the average severity of each problem and classify all those listed [30].

E. Analysis of Usability Problems

Usability problems found in the UT evaluation were extracted in a structured manner, as employed in [31] to mitigate biases (i.e. evaluator effect) and enhance data validity and reliability. An inter-coder check on reliability was also conducted, by an independent evaluator, on the UT usability problem analysis. This evaluator coded the usability problems experienced by the first participant in the experiment and discussed them with the researcher, before performing an independent analysis on two videos of the testing, selected at random. The agreement between the problems identified by the test subject and those revealed in the videos was a respectable 78% [32].

Problem severity in all evaluation conditions was classified according to the following scale [24]:

1) A catastrophic problem, which prevents users reaching their goal and has to be remedied.

2) A major problem, which leads to user frustration and difficulty in continuing, which should be remedied.

3) A minor problem, which leads to user frustration and difficulty in continuing, which could be remedied.

4) A cosmetic problem, which leads to minor issues for users and which can be remedied easily.

The problems were also classified in four types, navigation, layout, content and functionality (as in Table I), derived from prior research [31].

TABLE I. PROBLEM TYPE CODING SCHEME

	Type	Problem Definition
1	**Navigation**	Users have difficulty moving between pages or finding the right links for specific functions or information.
2	**Layout**	Users have difficulties in respect of the interface, such as display and visibility problems, inconsistent design and awkward design of structures and forms.
3	**Content**	Users either find unnecessary information or expect information which is not there, or they do not understand the information due to its terminology or style.
4	**Functionality**	Users have difficulties because some functions are missing or otherwise problematic.

IV. RESULTS

A. App A

1) Usability problems identified: A total of *66* problems were identified with App A in the test sessions, 56 with the HE method, 46 with CW and 44 with UT. HE therefore identified a wider range of problems than either UT or CW, significantly more so according to a Kruskal Wallis H test ($p < 0.0001$). The HE evaluators each found an average of 25 problems, while the CW evaluators found an average of 18 problems. In the UT evaluation, 11 individual issues arose in each session. However, HE found 14 unique issues which were not found by the CW and UT approaches. UT identified six issues not found in the other evaluations and CW found four. All of the methods were able to detect 38 of the total problems. This is illustrated in Fig. 1.

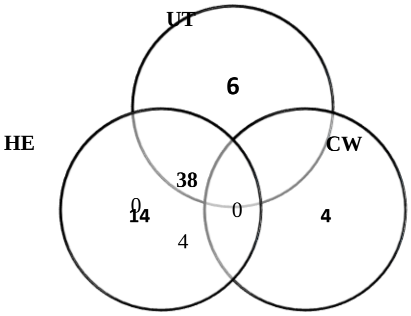

Fig. 1. Venn diagram of the numbers of problems identified by the three evaluation methods (App A).

2) Problem severity: 26 (59%) of the final problems identified in the UT evaluation were of high impact, i.e. either major or catastrophic). The remaining 41% had low impact, i.e. either minor or cosmetic. However, HE and CW both found 20 (36% and 43% respectively) problems of high impact. In fact, all the problems which were only found by UT method were of high impact, whereas those only found by HE and CW were of low impact (Table II).

3) Problem types: The 66 final problems found on App A were classified as 22 navigational, 19 layout issues, 16 content-related and 9 functional. HE found more layout and content problems, as well as identifying more problems of those types uniquely (Table III).

TABLE II. ISSUE SEVERITY

	UT		**HE**		**CW**	
	Unique	**Common**	**Unique**	**Common**	**Unique**	**Common**
Cosmetic	0	4	5	8	2	8
Minor	0	14	9	14	2	14
Major	4	19	0	19	0	19
Catastrophic	2	1	0	1	0	1
Total	6	38	14	42	4	42

TABLE III. PROBLEM TYPES

	UT		**HE**		**CW**	
	Unique	**Common**	**Unique**	**Common**	**Unique**	**Common**
Navigation	4	13	3	13	2	13
Layout	1	9	6	11	1	11
Content	1	7	5	9	1	9
Functionality	0	9	0	9	0	9
Total	6	38	14	42	4	42

4) User task performance and satisfaction: Table IV presents the descriptive statistics for task performance and user satisfaction in the UT evaluation. The rate of successful completion indicates that participants encountered difficulties in executing the tasks, as only half were completed, on average. It is clear that the fourth and fifth tasks were found most difficult, being completed only 10% and 8% of the time respectively. The first and second tasks were easier, being completed by 82.1% and 77.4% respectively. This explains why the UT evaluation found more catastrophic usability problems. However, the UT participants evaluated the application's usability highly, giving it an average score of 85, which exceeds the global average SUS score of 68 by a large margin.

TABLE IV. USER TASK PERFORMANCE AND SATISFACTION STATISTICS

	UT	
	Mean	**SD**
Tasks completed	2.50	1.00
Time to complete tasks (m)	33.04	7.35
Number of clicks	170.00	26.63
Number of screens browsed	15.15	4.34
SUS	85.35	10.10

B. *App B*

1) Usability problems identified: A total of 57 problems were found across the three evaluations for App B. 44 were found by HE, 36 by CW and 32 by UT. Once again, HE found significantly more issues than the other two techniques, which was confirmed by a Kruskal Wallis H test ($p < 0.0001$). The HE evaluators each found 21 problems on average, while those using CW found 16 and each UT session detected nine. The UT and CW methods both failed to spot 15 problems detected by HE, but found five and six, respectively, not found in the other tests. 24 problems were identified by all three evaluation methods (see Fig. 2).

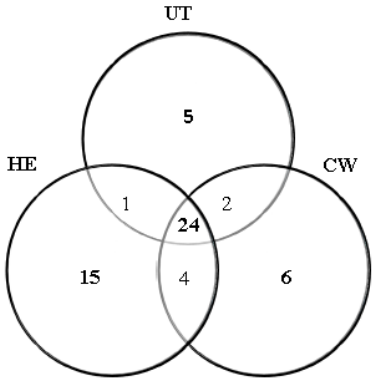

Fig. 2. Venn diagram of the numbers of problems identified by the three evaluation methods (App B).

2) Problem severity: 20 problems identified by the UT method (62%) were of high impact, while 14 (31%) and 15 (41%) of those identified by HE and CW, respectively, were of high impact. The UT method therefore performed better at identifying more severe problems and all of the problems found uniquely by UT were of high impact, whereas all of those found uniquely by HE were of low impact, as shown in Table V.

TABLE V. PROBLEM SEVERITY

	UT		HE		CW	
	Unique	Common	Unique	Common	Unique	Common
Cosmetic	0	3	3	3	3	2
Minor	0	9	12	12	3	13
Major	4	14	0	13	0	14
Catastrophic	1	1	0	1	0	1
Total	5	27	15	29	6	30

TABLE VI. PROBLEM TYPES

	UT		HE		CW	
	Unique	Overlap Problems	Unique	Overlap Problems	Unique	Overlap Problems
Navigation	2	8	3	10	0	10
Layout	3	4	6	6	3	5
Content	0	2	5	2	3	2
Functionality	0	13	1	11	0	13
Total	5	27	15	29	6	30

3) Problem types: App B's 57 final problems were classified as 15 navigational, 18 layout-related, 10 content-related and 14 functional (see Table VI). As with App A, HE found more layout and content problems, as well as identifying more problems of all types uniquely.

4) User task performance and satisfaction: Table VII presents the descriptive statistics for task performance and user satisfaction in the UT evaluation. The rate of successful completion indicates that participants encountered difficulties in executing the tasks, as only 3.2 were completed, on average. As with App A, the fourth and fifth tasks were most difficult, having completion rates of 14% and 11% respectively, while the first and second tasks were easier. Participants rated the application with a SUS of 77, which is also above the global average SUS score.

C. Correlation Analysis

Spearman's correlation coefficient was used across the two application case studies to examine any correlations that exist between the measures employed in the study [33]. As can be seen from Table VIII, one correlation that is statistically significant was found, between the time spent by participants on tasks and the number of problems found.

TABLE VII. USER TASK PERFORMANCE AND SATISFACTION STATISTICS

	UT	
	Mean	SD
Tasks completed	3.20	1.10
Time to complete tasks (m)	35.16	11.10
Number of clicks	173.30	28.45
Number of screens browsed	17.35	3.88
SUS	77.65	11.74

D. The Applications' Usability Levels

As described above, the usability evaluations revealed 123 problems with the two applications under test. This is a significant number of issues, which corresponds to a low level of usability. 46 of these problems (nearly 38%) were severe issues, having a significant impact on task performance. The majority of problems were navigational and related to the layout of the interface (both 30%), which indicates that users are likely to find it difficult to navigate within the applications. This finding is also supported by the rates of successful task completion. The results therefore clearly show that there needs to an improvement of both of these applications' usability.

TABLE VIII. CORRELATIONS BETWEEN MEASURES IN THE UT METHOD

	Tasks completed	Time to complete tasks	No. of clicks	No. of screens browsed	SUS	No. of problems
Tasks completed	1	-0.02	0.21	0.1	0.01	0.12
Time to complete tasks	-0.02	1	-0.18	-0.17	-0.25	0.39*
No. of clicks	0.21	-0.18	1	0.08	0	-0.02
No. of screens browsed	0.1	-0.17	0.08	1	0.1	-0.16
SUS	0.01	-0.25	0	0.1	1	0.05
No. of problems	0.12	0.39*	-0.02	-0.16	0.05	1

* The significance level is 0.05

V. Discussion

No prior research has compared the UEMs considered in this study in relation to m-health applications in Saudi Arabia, so it is no possible to draw any comparisons with comparable studies.

A. Is there a Difference between the Performance of UT, HE, and CW Methods in Evaluating M-Health Applications?

In general, the findings of this study are common to both m-health applications studied. HE found the most problems overall and detected more problems than either UT or CW techniques, but UT was able to detect more severe problems. Similar findings emerged from studies of other systems [19,21,34], where HE was found to detect more minor issues than did UT. This is also in line with Farzandipour et al. [35], who found that HE was better at detecting usability issues than the CW technique.

HE's identification of greater numbers of problems can be attributed to the fact that the HE evaluator were able to explore the applications, whereas the CW and UT participants were given specific tasks to do, so that there was a limit to the kind and number of usability problems that they would encounter. It is therefore recommended that the design of m-health applications should utilise both HE and UT methods in order to derive the maximum benefits.

B. Is there a Correlation between UT Metrics?

A strong correlation was found between the time taken to complete tasks and the number of usability problems encountered on these two m-health applications. This result indicates that time taken is a better indicator of the presence of usability problems than other measures.

It was also evident that the SUS does not serve well as a usability metric for m-health applications on its own. This is because it does not predict the presence of usability problems as well as task completion time. Usability practitioners should therefore at least report task completion times in addition to SUS in usability reports. This is in agreement with the findings of recent research into SUS. For instance, Broekhuis et al. [36] also found SUS to be inadequate by itself for e-health system usability evaluations.

SUS's poor predictive power may be due to a number of factors. 1) it is subjective, which means that usability is but one factor influencing the perception of the assessor, along with, for instance, usefulness and enjoyability. 2) SUS is generalised and does not reflect participants' actual performance; greater difficulty and more problems were encountered with App A, yet it received a higher average SUS. 3) SUS does not recognise such inclusivity factors as accessibility (e.g. for people with learning difficulties or visual disabilities) and information overload, even though they affect usability. Future research should include a comprehensive assessment of such factors and their impacts on usability, which is especially important in the context of health-related applications and systems.

C. What is the Current Usability Level of M-Health Applications in the Kingdom of Saudi Arabia?

This study found that the m-health applications targeted do not meet acceptable usability standards and fail to conform to appropriate design principles. According to the UEMs applied, these two applications are very hard to use, due to the large number of usability problems which arise, which was especially clear from the inspection by an independent expert and the task performance data in the UT sessions, even though the applications received good SUS ratings, which, is not a reliable indicator. These findings contradict previous research into user satisfaction with m-health applications in Saudi Arabia [8, 23], which found general satisfaction, perhaps because those studies relied on questionnaire-based, subjective data.

D. Recommendations

On the basis of the results of this study, a number of recommendations is presented below.

1) The varying effects of different UEMs should be considered seriously when evaluating the usability of m-health apps, as the findings suggest that results may differ depending on the method used. Therefore, practitioners should consider the pros and cons of each approach when deciding on an evaluation method.

2) Consider using the UT method when interested in identifying high severity usability problems.

3) Consider using the HE method when seeking to find higher numbers of low severity usability problems—particularly those relating to content and layout.

4) Both UT and HE should be employed during the design process of m- health applications for maximum effectiveness.

5) Usability practitioners should be aware of the fact that participants' satisfaction with the perceived usability of m-health application does not correlate with the number of usability problems that found on the interface. This implies that SUS questionnaires should not be used as a sole metric for determining the usability of the m-health interfaces.

6) There is a need to an improvement of m-health applications' usability in Saudi Arabia. In particular, the navigation and layout aspects of the interface should be given more attention.

7) Web developers are key to ensuring the usability of m-health apps. If there is to be a positive effect, it is important for developers to enhance their awareness of usability standards.

E. Limitations and Future Work

There are, inevitably, some limitations to the present study. First, the UT sessions recorded various task performance parameters, but did not measure behavioural factors, such as attention levels, which other studies have estimated using eye tracking technology [37]. Second, the study was limited to two Saudi m-health applications, which, whilst in common use, may not reflect the overall usability of m-health applications in Saudi Arabia. Therefore, more research with a broader spectrum of m-health applications across different healthcare domains would be necessary to further assess the generalizability of the results.

VI. Conclusion

This study compared three usability evaluation methods, namely, usability testing, heuristic evaluation, and cognitive walkthrough in terms of their effectiveness in assessing m-health application usability, and to assess the usability of such applications in the Saudi Arabian context. It also explored the relationships between usability metrics. The study finds that heuristic evaluation is able to identify a larger number of usability issues, which is because it takes a broad overview of system design, with predefined principles, rather than focusing on the performance of specific tasks. However, it does appear to identify more minor problems than the usability testing method, which is better at detecting more severe issues. The study also identified a significant relationship between the number of usability problems found and how long participants spend carrying out tasks using m-health applications, which suggests that the existence of usability problems. The findings also show that the two applications tested have low usability levels and need to be improved.

References

[1] Islam MN, Karim MM, Inan TT, Islam AN. Investigating usability of mobile health applications in Bangladesh. BMC medical informatics and decision making. 2020 Dec;20:1-3.

[2] Istepanian R, Laxminarayan S, Pattichis CS, editors. M-health: Emerging mobile health systems. Springer Science & Business Media; 2007 Jan 4.

[3] Kay M, Santos J, Takane M. mHealth: New horizons for health through mobile technologies. World Health Organization. 2011 Jun 7;64(7):66-71.

[4] Tomlinson M, Solomon W, Singh Y, Doherty T, Chopra M, Ijumba P, Tsai AC, Jackson D. The use of mobile phones as a data collection tool: a report from a household survey in South Africa. BMC medical informatics and decision making. 2009 Dec;9:1-8.

[5] Rotheram-Borus MJ, Le Roux IM, Tomlinson M, Mbewu N, Comulada WS, Le Roux K, Stewart J, O'Connor MJ, Hartley M, Desmond K, Greco E. Philani Plus (+): a Mentor Mother community health worker home visiting program to improve maternal and infants' outcomes. Prevention Science. 2011 Dec;12:372-88.

[6] Siedner MJ, Haberer JE, Bwana MB, Ware NC, Bangsberg DR. High acceptability for cell phone text messages to improve communication of laboratory results with HIV-infected patients in rural Uganda: a cross-sectional survey study. BMC medical informatics and decision making. 2012 Dec;12:1-7.

[7] Haberer JE, Robbins GK, Ybarra M, Monk A, Ragland K, Weiser SD, Johnson MO, Bangsberg DR. Real-time electronic adherence monitoring is feasible, comparable to unannounced pill counts, and acceptable. AIDS and Behavior. 2012 Feb;16:375-82.

[8] Alanzi TM. Users' satisfaction levels about mHealth applications in post-Covid-19 times in Saudi Arabia. PloS one. 2022 May 4;17(5):e0267002.

[9] Shati A. Mhealth applications developed by the Ministry of Health for public users in KSA: a persuasive systems design evaluation. Health Informatics Int J. 2020;9(1):1-3.

[10] Broekhuis M, van Velsen L, Hermens H. Assessing usability of eHealth technology: a comparison of usability benchmarking instruments. International journal of medical informatics. 2019 Aug 1;128:24-31.

[11] Georgsson M, Staggers N. Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics. Journal of the American Medical Informatics Association. 2016 Jan 1;23(1):5-11.

[12] Harrison R, Flood D, Duce D. Usability of mobile applications: literature review and rationale for a new usability model. Journal of Interaction Science. 2013 Dec;1:1-6.

[13] Mubeen M, Iqbal MW, Junaid M, Sajjad MH, Naqvi MR, Khan BA, Saeed MM, Tahir MU. Usability evaluation of pandemic health care mobile applications. InIOP conference series: earth and environmental science 2021 Mar 1 (Vol. 704, No. 1, p. 012041). IOP Publishing.

[14] Alharbi A, Alzuwaed J, Qasem H. Evaluation of e-health (Seha) application: a cross-sectional study in Saudi Arabia. BMC medical informatics and decision making. 2021 Dec;21:1-9.

[15] Zhou L, Bao J, Setiawan IM, Saptono A, Parmanto B. The mHealth app usability questionnaire (MAUQ): development and validation study. JMIR mHealth and uHealth. 2019 Apr 11;7(4):e11500. Tan WS, Liu D, Bishu R. Web evaluation: Heuristic evaluation vs. user testing. International Journal of Industrial Ergonomics. 2009 Jul 1;39(4):621-7.

[16] Hasan L, Morris A, Probets S. A comparison of usability evaluation methods for evaluating e-commerce websites. Behaviour & Information Technology. 2012 Jul 1;31(7):707-37.

[17] Doubleday A, Ryan M, Springett M, Sutcliffe A. A comparison of usability techniques for evaluating design. InProceedings of the 2nd conference on Designing interactive systems: processes, practices, methods, and techniques 1997 Aug 1 (pp. 101-110).

[18] Jeffries R, Miller JR, Wharton C, Uyeda K. User interface evaluation in the real world: a comparison of four techniques. InProceedings of the SIGCHI conference on Human factors in computing systems 1991 Mar 1 (pp. 119-124).

[19] Thyvalikakath TP, Monaco V, Thambuganipalle H, Schleyer T. Comparative study of heuristic evaluation and usability testing methods. Studies in health technology and informatics. 2009;143:322.

[20] Khajouei R, ZahiriEsfahani M, Jahani Y. Comparison of heuristic and cognitive walkthrough usability evaluation methods for evaluating health information systems. J Am Med Inform Assoc. 2017;24(e1):e55–60.

[21] Maguire M, Isherwood P. A comparison of user testing and heuristic evaluation methods for identifying website usability problems. In: Marcus A, Wang W, editors. Design, user experience, and usability: theoty and practice: 7th international conference (DUXU 2018), Las Vegas, NV, USA, 15–20 July 2018, Part I. p. 429–438.

[22] Arafa A, Mostafa ZM, Sheerah HA, Alzahrani F, Almuzaini Y, Senosy S, Hassan RI. mHealth app barriers, usability, and personalization: a cross-sectional study from Egypt and Saudi Arabia. Journal of Personalized Medicine. 2022 Dec 9;12(12):2038.

[23] Shilbayeh SA, Ismail SA. Patient experience with an educational mobile health application: A pilot study on usability and feasibility in a Saudi population. Cogent Psychology. 2020 Dec 31;7(1):1843883.

[24] Lazar J, Feng JH, Hochheiser H. Research methods in human-computer interaction. Morgan Kaufmann; 2017 Apr 28.

[25] Maguire M. Context of use within usability activities. International journal of human-computer studies. 2001 Oct 1;55(4):453-83.

[26] Sauro J, Lewis JR. Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann; 2016 Jul 12.

[27] Reynoso JM, Olfman L, Ryan T, Horan T. An information systems design theory for an expert system for training. Journal of Database Management (JDM). 2013 Jul 1;24(3):31-50.

[28] The Communications, Space, and Technology Commission. The Saudi Internet Report. 2024 April. Online available at https://www.cst.gov.sa/en/mediacenter/pressreleases/Pages/2024042402.aspx Accessed [May. 18, 2024]

[29] Nielsen J, Molich R. Heuristic evaluation of user interfaces. InProceedings of the SIGCHI conference on Human factors in computing systems 1990 Mar 1 (pp. 249-256).

[30] Blackmon MH, Polson PG, Kitajima M, Lewis C. Cognitive walkthrough for the web. InProceedings of the SIGCHI conference on human factors in computing systems 2002 Apr 20 (pp. 463-470).

[31] Alhadreti O. Comparing two methods of usability testing in Saudi Arabia: concurrent think-aloud vs. co-discovery. International Journal of Human–Computer Interaction. 2021 Jan 20;37(2):118-30.

[32] Hertzum M, Jacobsen NE. The evaluator effect: A chilling fact about usability evaluation methods. International journal of human-computer interaction. 2001 Dec 1;13(4):421-43.

[33] Sedgwick P. Spearman's rank correlation coefficient. Bmj. 2014 Nov 28;349.

[34] Wang E, Caldwell B. An empirical study of usability testing: heuristic evaluation vs. user testing. InProceedings of the Human Factors and Ergonomics Society Annual Meeting 2002 Sep (Vol. 46, No. 8, pp. 774-778). Sage CA: Los Angeles, CA: SAGE Publications.

[35] Farzandipour M, Nabovati E, Sadeqi Jabali M. Comparison of usability evaluation methods for a health information system: heuristic evaluation versus cognitive walkthrough method. BMC Medical Informatics and Decision Making. 2022 Jun 18;22(1):157.

[36] Broekhuis M, van Velsen L, Hermens H. Assessing usability of eHealth technology: a comparison of usability benchmarking instruments. International journal of medical informatics. 2019 Aug 1;128:24-31.

[37] Ţichindelean M, Ţichindelean MT, Cetină I, Orzan G. A comparative eye tracking study of usability—towards sustainable web design. Sustainability. 2021 Sep 18;13(18):10415.