A Metric-BasedEvaluation Model for Applicationson Mobile Phone

Azham Hussain[1], Maria Kutar[2] and FazillahMohmad Kamal[3]

[1]*Universiti Utara Malaysia, Malaysia, azham.h@uum.edu.my*
[2]*University of Salford, United Kingdom, m.kutar@salford.ac.uk*
[3]*Universiti Utara Malaysia, Malaysia, fazillah@uum.edu.my*

ABSTRACT

Research on HCI, specifically on mobile application has started more than a decade. Usability metrics have been used as guidelines to evaluate the quality of the system as well as mobile applications. However, the metrics used for evaluation method keep changing due to the new inventions on mobile phones. Thus, there is a need to create a dynamic model for evaluation that can grow together with new inventions and technology. In this paper, we created a dynamic usability metrics model and test the model to ensure the model is reliable and effective. The model comprises usability goals, questions and metrics for evaluation of applicationson mobile phone. This paper also reports the usability test results for applicationsinstalled on different mobile phone.

Keywords: Usability, Goal Question Metric, Mobile Application, iPhone, O2 Orbit.

IINTRODUCTION

Usability evaluation has grown into a well-established research area. The first guidelines to evaluate the application system as well as today's mainstream is ISO 9241 – 11 standards (ISO, 1998). For several years, this focus on generic usability metrics has been countered by others who argue in favour of using specific usability metric. The discussion of this difference between generic and specific has mostly been a matter of opinions, and it has not been prominent in the literature on the comparison of usability metric, e.g. (Bertoa et al., 2006) and (Ahmed et al., 2006).

New invention and current features on mobile phone will reflect existing evaluation metric and guideline. Application used built-in digital compass on mobile phone can find the location quickly and accurately, get a direction and see which way user is facing. On the other hand, GPS receiver inside the mobile phone will benefit many users by using the application associate with GPS receiver, for example, SatNav application and find-me application. All this application required new metrics for evaluation method. To the best of our knowledge, there is no published work on how to evaluate the application with these new features on mobile phone.

There are a number of models for usability measurement; for instance, Quality in Use Integrated Measurement (QUIM) developed by (Ahmed et al., 2006). QUIM is a consolidated model for usability measurement and metric; and also appropriate for user who has no or little knowledge of usability. The model consists of 10 factors, which are subdivided into 26 criteria. For the measurement of the criteria, the model provides 127 metrics. The model is used to measure the actual use of working software and identifying the problem; however, the model is not optimal yet and needs to be validated.Many current models and methods which aim to evaluate usability still have some limitations, for instance, they are not intended for developers who are not familiar with the field of HCI, and they are difficult to apply.

In the following section, we examine previous studies on usability metric base on model development for desktop software and mobile application. Section III provides detail about the method used to conduct the study. The results are presented in section IV and sections V discussed a conclusion arising from the study and provide recommendations for further work.

IIUSABILITY MODEL AND METRIC

In this section, we review existing usability model and metric while highlighting some of their contributions and limitation. Metrics for Usability Standards in Computing (MUSiC) develop by Bevan & MacLeod (1994) is another project concerned on defining measures of software usability and was integrated into the original ISO 9241 standard. Examples of specific usability metrics in the MUSiC framework include user performance measures, such as task effectiveness, temporal efficiency, and length or proportion of productive period. However, a strictly performance-based view of usability cannot reflect other aspects of usability, such as user satisfaction or learnability. Software

Usability Measurement Inventory (SUMI) develops by Kirakowski& Corbett (1993) is a part of MUSiC project. SUMI was developed to provide measures of global satisfaction of five more specific usability areas, including effectiveness, efficiency, helpfulness, control, and learnability. Another MUSiC project related to software tool development is Diagnostic Recorder for Usability Measurement (DRUM) developed by Macleod &Rengger(1993). This project concerns with the analysis of user-based evaluations and delivery of these data to the appropriate party, such as a usability engineer. The Log Processor component of DRUM is the tool concerned with metrics. It calculates several different performance-based usability metrics including, task time, snag, help, and search times, effectiveness, efficiency, relative efficiency and productive period.

In addition, the Automated Interface Designer and Evaluator (AIDE) that were developed by Sears (1995) concerns with evaluating static HTML pages according to a set of predetermined guidelines about Web page design. AIDE is a software tool that able to generate alternative interface layouts and evaluate some aspects of a design. Among things that are concerned in these guidelines include the placement and alignment of screen elements, for example, text, buttons, or links. There is two metrics to be evaluated in the design which is task-sensitive metric and task-independent metric. Task-sensitive metrics incorporates task information into the development process which may ensure that user tasks guide the semantics of interface design. Task-independent metrics tends to be based on principles of graphic design and help to ensure that the interface is aesthetically pleasing. AIDE tool can measure a total of five different usability metrics, including efficiency, alignment, horizontal balance, vertical balance, and designer-specified constraints. Subsequently, other models that deals with the analysis of the quality of use for interactive devices was introduced, which is The Skill Acquisition Network (SANe) by Macleod &Rengger(1993). His approach assumes a user interaction model that defines user tasks, the dynamics of the device, and procedures for executing user tasks. Specifically, a task model and a device model are simultaneously developed and subsequently linked. After that, user procedures are simulated within the linked task-device model. A total of 60 different metrics is described in this framework, of which 24 concerns with the quality measures. Scores from the latter are then combined to form a total of five composite quality measures including: efficiency, learning, adaptiveness, cognitive workload, complexity and effort for error correction.

A. Usability for Mobile Application

Study on the challenges and issues of mobile application by Zhang &Adipat(2005) lists nine usability attributes and measuring variables as a part of their studies. All the generic attributes were collected and compiled from existing usability studies, but they were not validated. On the other hand, Gafni(2008)introduced the usability quality characteristic for a mobile wireless information systems. The study focus on the development of a new metric and all metrics was validated theoretically and empirically at least by one of four different experiments performed in diverse devices. However, the device used in the experiment is quite old and this model needs to be updated to provide new metric for new mobile phone. Study byTerrenghiet al., (2005) shows the detail usability metric for mobile devices by refining the usability characteristic from ISO /IEC 9216-1. This study focuses on new issues to encounter usability requirements for mobile computing scenarios. However, they are not validated.

B. International Organization for Standardization (ISO)

The International Organization for Standardization (ISO) is an international standard setting body composed of representatives from various national standards organizations. ISO has developed over 17000 International Standards on a variety of subjects and 1100 new ISO standards are published every year (source from ISO website). Most of the literature in HCI employed ISO9241-11 for usability measurement (Hornbæk & Law, 2007). Table 1 lists the ISO standard related to HCI and ISO9241-11 specifically addresses the definition of usability measurement. Study by Constantinos& Dan(2007) found the highest characteristics in usability evaluation are effectiveness (62%), Efficiency (33%) and satisfaction (20%). These three characteristics reflect the ISO 9241 standard, and thus it is chosen as foundation of the model in this study, similar to (Hornbæk, 2006).

Table 1: ISO Standard Related to Measurement

Usability in ISO Standard	Description
The ISO 9241-11	Identify efficiency, effectiveness, and

(1998)	satisfaction as major attributes of usability.
ISO/IEC 9126-1 (2001)	Define the standard as a software quality attributes that can be decomposed into five different factors, including understandability, learnability, operability, attractiveness, and usability compliance.
ISO/IEC 9126-4 (2001)	Define the related concept of *quality in use* as a kind of higher-order software quality attribute.
The ISO/IEC 14598-1 (1999)	A model for measuring quality in use from the perspective of internal software quality attributes

III METHOD

We develop the model by analyzing the journal related to HCI. The total of 409 journals was reviewed based on keywords "usability", "evaluation" and "metric". Only 26 out of 409 journals selected for further review in obtaining the guidelines for mobile application development. Table 2 below describes the journal papers that were reviewed.

Table 2: Journal Papers

Journal	Year	Candidate	Selected
TOCHI	2006-2010	54	8
HCI	2006-2010	36	2
IJHCI	2006-2010	97	5
IJHCS	2006-2010	222	11
Total		409	26

The review is based on a conception of usability, similar to (Hornbæk, 2006). This conception merely discusses studies related to usability evaluation instead of the broad concept of usability. We analyse the quality characteristic of each measures to ensure there is no duplication. We also refine the measures to simplify the guidelines and to ensure the model not too complex. Interestingly, we found most of the studies employed effectiveness, efficiency and satisfaction as quality characteristics. Thus, we decide to make these three characteristics as a base of guidelines and others become sub guidelines. From seventeen popular guidelines, we summarized it into six usability characteristics to become a goal for the model as shown in table 3 below.

Table 3: Usability Characteristic

Quality Characteristic	Goal	Guidelines
Effectiveness	Simplicity	-Ease to input the data -Ease to use output -Ease to install -Ease to learn
	Accuracy	-Accurate -Should be no error -Successful
Efficiency	Time taken	-To response -To complete a task
	Features	-Support/help -Touch screen facilities -Voice guidance -System resources info. -Automatic update
Satisfaction	Safety	-While using the application -While driving
	Attractiveness	-User interface

We employed Goal Question Metric (GQM) approach by Basili et al., (1994) to develop the model because the approach allows the model to be enhanced at any time. First steps, we input goals from table 3 into the model. In second step, we create questions to assess each goal described in the first step. We carefully created the questions by refining the goal into several questions and ensure the questions we created are measurable. Finally, we develop a set of metrics that provide the information to answer those questions. In this case, we refined all the questions into metrics. The model are shown in Appendix A, consist of 17 objective measures and 19 subjective measures.

To ensure the model is reliable, effective and optimal we carry out experiments to test the usability of applications in two different mobile phones. We conduct the experiment to test whether the metrics can be used to collect the usability data. Besides that we will analysed the data to compare the usability problem of application being installed in current mobile phone and traditional mobile phone.The experiment is divided into two parts; first we collect the objective data through observation, and then we collect subjective data via an interview to assess the perception from participants on mobile application, as recommended by Nielsen (1994).

Picture 1: Participant using CoPilot inside carPicture 2: Participant using Mobile Facebook

We used the Mobile Facebook and CoPilot Live SatNav system installed in an iPhone and O2 Orbit mobile phone device. We conduct the experiment in the comfortable and quiet room for Mobile Facebook and the experiments were conducted inside a car for CoPilot system in order to mirror the way such applications are used in practice.However, participants did not drive the car during the study for safety reasons.We mixed the participants (novice, expert, men, women and age from 20 to 35) and all participants were asked to complete five tasks, and they were given time to explore and learn the application before continue to complete all the tasks.Picture 1 and 2 above are some pictures taken during usability test.

A.**Objective Data**

Most of the data can be collected but some of them were unable to collect, for instance; the metric 'Time taken to install', 'The number of interaction while installing the application' and 'Percentage of Battery used while installation'. Facebook's application on O_2 Orbit is a wireless application and no installation process for Facebook. Moreover, we are also unable to obtain data related to automatic update and influence the metric 'Number of request to update the application'. Sometime we receive automatic update alert from the owner of the application, unfortunately it won't come out while conducting the experiment.

B.**Subjective Data**

We create a semi-structured instrument for the interview session, to test whether we can collect subjective data using GQM model. The questions were designed to be not too technical, and the session was conducted in an informal manner. The overall aim being to obtain participants' opinionswhile using mobile application. Examples of questions include the feeling after completed the task, the comment on the menu arrangement, voice assistance, interface, screen, satisfaction on system speed and safety. We also ask participants to comment on the devices for both iPhone and O2 Orbit in terms of screen size, speed and text size.

Participants were interviewed after they had used each of the applications on different mobile phone. Only a metric is unable to obtain, which is 'Virtual joystick'. We did not ask the participants, whether they satisfy with 'Virtual joystick' because they did not use the joystick for both applications. In the result section, we compare the number of positive and negative comments for applications inside iPhone and O_2 Orbit.

IV Results and Discussion

In this section, we will compare the objective and subjective results for iPhone and O2 Orbit mobile phone. We use SPSS software version 17 to run t-test for each metric to show the different between current mobile phone and traditional mobile phone. We also compare the different on subjective metric using Nvivo 8. We create 19 tree nodes similar to number of subjective metrics in GQM model to check whether the data can be collected using GQM model.

A.**Objective Measure**

We run t-test for all metrics and below are an example t-test for one of the metric.

Metric 1: Time taken to input the data

H_0: $\mu_{iPhone} = \mu_{O2}$	H_1: $\mu_{iPhone} < \mu_{O2}$

Where

H_0= the null hypothesis

μ_{iPhone}= time taken to input the data using iPhone

μ_{O2} = time taken to input the data using O_2

($t_8 = 0.018$, $p < .05$)

A result shows that at 95% confident level, time taken to input the data for an iPhone is shorter than O_2.

We summarized all t-test results and found 13 metrics were tested out of 17 for CoPilot application and 11 metrics were tested for Facebook. Some metrics are unable to run the t-test due to the standard deviation for both groups are 0. We are also unable to run t-test for some metrics(time taken to install) on Facebook due to no data to test. For CoPilot application, we found a significant different between iPhone and O2 orbit for 7 usability metricsand the other 6 are not different. For Facebook's application, we found a significant different between iPhone and O2 orbit for 6 usability metrics and the other 5 are not different. Overall results show that application installed on an iPhone is better than O2 Orbit.

B. **Subjective Measure**

We analyzed interview transcript and categorized the comment base on the nodes we created in Nvivo. We check the comment and identify whether it is a positive or negative comment. Table 4 shows the overall result for both applications in two different mobile phones. iPhone obtains 27 positive feedbacks and 13 negative feedbacks, whereas O2 Orbit gets only 12 positive feedbacks and 39 negative feedbacks.

From the interview transcript, most participants were very happy to use CoPilot inside iPhone except one participant who expressed dissatisfaction with the virtual keypad. All participants were unhappy with CoPilot on O2 Orbit and mentioning screen size, touch screen, tiny virtual keypad and most participants stated that overall they didn't enjoy using CoPilot on the Orbit. For the Facebook on an iPhone, interestingly we found a more equal balance of positive and negative feedback. Participants were unhappy using the virtual keypad on the iPhone, and they noted that the keypad is too sensitive. Most participants gave positive feedback about the content. For Facebook on O2 Orbit, all participants mentioned that the virtual keypad as is too small and they do not like to use the stylus. Some participants still made mistakes while using the stylus and suggested a physical keypad for data entry would be preferable. Participants were also unhappy with the overall navigation and interface design, and they suggested having one main menu for all sub menus on Facebook.

Table 4: Result for Subjective Measure

Application / Device	Positive Feedback	Negative Feedback
CoPilot / iPhone	16	8
Facebook / iPhone	11	5
Total for iPhone	**27**	**13**
CoPilot / O2 Orbit	6	15
Facebook / O2 Orbit	6	24
Total for O2Orbit	**12**	**39**

The result for objective and subject measure shows that the application on an iPhone is better than O2 Orbit in terms of interaction. However, comparison is not the main objectives of this study apart from validation purpose only. We also found that the model could generate too many metrics and become too complex. Thus, we recommend having an optimal number of metrics by reduce or combine the metrics; for example, the metric 'virtual keyboard' and 'virtual joystick' can be combined into 'satisfaction with touch screen'.

V**Conclusions**

We develop GQM model as guidelines to evaluate usability of mobile application and proof that the model can be used to evaluate the application on mobile phone. The model can be edited whether to add or drop the goals, the questions and the metrics. This capability allows a new measure to be inserted into the model by creating a new goal or new questions. The model will benefit usability evaluator as well as a mobile application developer as guidelines while design mobile application. However, the model is only a list of usability metrics, an evaluator still need to set up and plan for experiment method. Moreover, this model focuses merely on interaction between human-computer and could be enhanced to the other area for instance; how the device handle memory load and load the content into the screen. For the future study, we look forward to develop an automated tool to evaluate mobile application using GQM model and the tool will have features to add or drop

themetrics. For further test on the model we suggest to test the model using field method to ensure the model can be used in any conditions.

REFERENCES

Ahmed, S., Mohammad, D., Rex, B. K., & Harkirat, K. P. (2006). Usability measurement and metrics: A consolidated model. *Software Quality Control, 14*(2), 159-178.

Basili, V., Caldeira, G., & Rombach, H. D. (1994). The Goal Question Metric Approach. *Encyclopedia of Software Engineering*.

Bertoa, M. F., Troya, J. M., & Vallecillo, A. (2006). Measuring the usability of software components. *Journal of Systems and Software, 79*(3), 427-439.

Bevan, N., & MacLeod, M. (1994). Usability measurement in context. *Behaviour & Information Technology, 13*(1), 132 - 145.

Constantinos, K. C., & Dan, K. (2007). *A research agenda for mobile usability*. Paper presented at the CHI '07 extended abstracts on Human factors in computing systems.

Drummond, J. S., & Themessl-Huber, M. (2007). The cyclical process of action research: The contribution of Gilles Deleuze. *Action Research, 5*(4), 430-448.

Gafni, R. (2008). Framework for Quality Metrics in Mobile-Wireless Information Systems. *Interdisciplinary Journal of Information, Knowledge, and Management, 3*, 23 - 38.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79-102.

Hornbæk, K., & Law, E. L.-C. (2007). *Meta-analysis of correlations among usability measures*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

ISO. (1998). *INTERNATIONAL STANDARD: ISO 9241-11(Guidance on usability)*. Geneva.

Kirakowski, J., & Corbett, M. (1993). SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology, 24*(3), 210-212.

Macleod, M., & Rengger, R. (1993). The development of DRUM: A software tool for video-assisted usability evaluation. *Journal*. Retrieved from http://www.nigelbevan.com/papers/drum93.pdf

Nielsen, J. (1994). *Usability Engineering*. London: Morgan Kaufmann.

Sears, A. (1995). *AIDE: a step toward metric-based interface development tools*. Paper presented at the Proceedings of the 8th annual ACM symposium on User interface and software technology.

Terrenghi, L., Kronen, M., & Valle, C. (2005). *Usability Requirements for Mobile Service Scenarios*. Paper presented at the Human Computer Interaction 2005 (IHCI'05), Las Vegas, USA.

Zhang, D., & Adipat, B. (2005). Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications. *International Journal of Human-Computer Interaction, 18*(3), 293 - 308.