IMPROVING THE RELIABILITY AND COST-EFFECTIVENESS OF USABILITY INSPECTION METHOD FOR MOBILE APPLICATION WITH INTEGRATED INSPECTION FRAMEWORK

CHENG LIN CHOU

UNIVERSITI SAINS MALAYSIA

2015

IMPROVING THE RELIABILITY AND COST-EFFECTIVENESS OF USABILITY INSPECTION METHOD FOR MOBILE APPLICATION WITH INTEGRATED INSPECTION FRAMEWORK

by

CHENG LIN CHOU

Thesis submitted in fulfilment of the requirements for the degree of
Master of Arts

March 2015

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my research supervisor, Dr Sam Muhizam Mustafa. His insight and guidance had rendered much value to this research and ensured its timely completion. Dr Sam has taught me well about the goals of academic writings, as all research findings are worthless if no one can comprehend them. This is the one point I kept in mind during the course of this research. I would also like to thank Ms Sumetha Nagalingam, who readily took me in and nurtured me into the field of usability research. I would like to thank both Dr Sam and Ms Sumetha for their time in reviewing this thesis and its earliest drafts with great acuity.

I would like to thank Ms Jenny Ling for her kindness in assisting to improve the readability of my writings from time to time; being a linguist herself, Ms Ling sometimes wonders I ever learnt English before. Also, I would like to thank Dr Sarena Abdullah, Ms Noor Azlina, Prof. Madya Dr Shanti Balraj Baboo, Dr Mohd Asyiek Mat Desa and all the faculty members in the School of The Arts for the helpful suggestions given to this thesis during the postgraduate seminars.

Last but not least, I would like to thank my parents and (especially) my wife, Su Yin Ng, for their encouragement and love throughout all these years.

TABLE OF CONTENTS

CHAPTER 7 - CONCLUSION AND DISCUSSIONS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CW	Cognitive Walkthrough
HE	Heuristics Evaluation
IIF	Integrated Inspection Framework
MMV	Mobile Mesh Viewer
SUS	System Usability Scale
TA	Task Analysis
UEM	Usability Evaluation Method
UT	Usability Testing

LIST OF APPENDICES

MENINGKATKAN KEBOLEHPERCAYAAN DAN KEBERKESANAN KOS UNTUK KAEDAH PEMERIKSAAN KEBOLEHGUNAAN BAGI APLIKASI MUDAH ALIH DENGAN RANGKA PEMERIKSAAN BERSEPADU

ABSTRAK

Permintaan berterusan bagi perkhidmatan dan aplikasi mudah alih yang inovatif telah mewujudkan satu peluang ekonomi baru dalam perniagaan pembangunan aplikasi. Walau bagaimanapun, gangguan teknologi yang berterusan dan isu fragmentasi dalam peranti mudah alih telah mencipta masalah teknologi yang serius, di mana para pemaju perlu bersaing untuk melancarkan aplikasi mereka sebelum kitaran hidup platform sasaran mereka mengalami perubahan. Perlumbaan melampau sedemikian telah memaksa sekumpulan pemaju syarikat teknologi yang kecil melangkau ujian kebolehgunaan demi penjimatan kos. Perbuatan yang sedemikian adalah disebabkan oleh kaedah ujian kebolehgunaan yang sedia ada terlalu memakan masa dan agak mahal untuk dikendalikan. Oleh itu, terdapat keperluan untuk pendekatan jaminan kualiti lebih tangkas dalam menilai kebolehgunaan aplikasi mudah alih, yang boleh menampung model pembangunan masa-ke-pasaran yang lebih pendek. Kajian ini bermula dengan penyelidikan ke dalam mekanik kaedah pemeriksaan kebolehgunaan yang dikenali dengan pendekatan pendiskaunan kos silih gantinya. Dari segi perspektif mengawal kos, kaedah pemeriksaan kebolehgunaan adalah lebih menjimatkan jika dibandingkan dengan kaedah ujian kebolehgunaan tradisional. Walau bagaimanapun, terdapat

kebimbangan sah tentang kaedah pemeriksaan kebolehgunaan untuk sama ada ia masih sesuai untuk menilai aplikasi mudah alih hari ini, kerana kaedah tersebut telah dibangunkan sebelum kemunculan budaya aplikasi dan peranti skrin sentuh. Sebagai tambahan kepada perkara tersebut, kaedah pemeriksaan kebolehgunaan cenderung kepada kekurangan kebolehpercayaan jika dibandingkan dengan kaedah ujian kebolehgunaan yang mahal, di mana penilai berbeza menilai aplikasi yang sama dengan kaedah pemeriksaan yang sama cenderung untuk mendapatkan hasil rumusan yang berbeza. Oleh itu, kajian ini telah menampilkan Rangka Kerja Pemeriksaan Bersepadu (IIF), yang bermatlamat untuk menangani kedua-dua isu-isu kesahan dan kebolehpercayaan yang ditemui dalam kaedah pemeriksaan kebolehgunaan. Bagi menentukan keberkesanan IIF, kajian ini telah menjalankan dua kajian kebolehgunaan, dengan satu kajian menggunakan kaedah IIF yang dicadangkan manakala kajian lain menggunakan kaedah ujian kebolehgunaan untuk menilai aplikasi mudah alih yang sama. Kemudian, kaedah analisis regresi linear akan digunakan bagi kajian kekorelatifan antara kedua-dua set penemuan berasingan. Dari penemuan yang dianalisis, IIF didapati amat kos efektif, dan ia adalah boleh dipercayai sebagaimana kaedah ujian kebolehgunaan sampel besar untuk mengesan masalah kebolehgunaan dalam aplikasi mudah alih. Kesimpulannya, IIF adalah kaedah pemeriksaan yang sah dan kos efektif untuk menentukan kebolehgunaan aplikasi mudah alih.

IMPROVING THE RELIABILITY AND COST-EFFECTIVENESS OF USABILITY INSPECTION METHOD FOR MOBILE APPLICATION WITH INTEGRATED INSPECTION FRAMEWORK

ABSTRACT

The on-going demand for innovative mobile services and applications has created a whole new economic opportunity in the business of app development. Nevertheless, the rapid evolution of technology disruption and the fragmentation of mobile devices have posed a serious technological challenge for many tech startups, where the developers have to race to deploy their apps before the lifecycle of their targeted platform changes. In order to keep up with the pace of change, the indie developers and tech startups usually resort to cut corners and skip usability tests, as the existing methods of usability testing are too time consuming and costly. Thus, there is a need for a more rigorous approach in accessing the quality and usability of a mobile app, while shortening the time-to-market process. This research started out by exploring the mechanics of the various usability inspection methods which are known for their alternate cost discounting approaches. From the perspective of cost containment, these methods of usability inspection are much economical compared to the traditional usability test methods. However, there is a validity concern over the usability inspection methods; are they still suitable to evaluate the present mobile app? This is because the methods were developed way before the emergence of app culture and touch screen devices. In addition to that, the usability inspection methods tend to come short in term of reliability when compared to the costly

usability test-methods, whereby different evaluators are involved in evaluating the same application with the same inspection method, which tend to lead to different concluding results. As such, this research would propose the Integrated Inspection Framework (IIF), which aims to address both the issues of validity and reliability found in the usability inspection methods. To determine the effectiveness of IIF, this research conducted two usability studies, one study using the proposed method of IIF while the other study using the usability test method to evaluate the same mobile app. Linear regression analysis was then applied to study the correlation between these two sets of separate findings. IIF was found to be highly cost-effective, and it is as reliable as the large sample usability test method for detecting usability problems within a mobile app. In conclusion, IIF is a valid and cost effective inspection method for determining the usability of a mobile app.

Chapter 1

Introduction

1.1 Introduction

Tablet computing has become a formidable trend where its portability and functionality has gradually become a "fourth screen" in most household right after TV-screen, personal computer and mobile phone. 'The whitepaper: Mobile Future in Focus' has highlighted that tablet devices have grown tremendously in 2011. It took less than two years to reach nearly 40 million users among the United States mobile population (comScore Inc., 2012). The adoption rate of tablet devices has significantly outpaced the growth of smart phone devices which took 7 years to reach the same level of user-adoptions (comScore Inc., 2012). Evidently, the web-based analytic data by StatCounter (See Appendix A) has suggested there are more people in South East Asia who are connected to the internet via mobile devices than traditional desktop devices (StatCounter, 2014).

This increasing use of mobile devices like tablet computer is likely to double in Asia when cloud computing becomes a gold standard in the developing countries (Morgan Stanley, 2011 a, 2011 b). The rapid adoption of tablet computing and other smart mobile devices has given rise to the "Apps Culture" that has never existed before (Purcell, Entner, & Henderson, 2010). An "app" is a software application that operates within the ecosystem of mobile devices. The emergence of apps culture and tablet computing has signalled that desktop computers could be superseded in a

short span of time. The support of new input methods such as touch-screen interaction and voice control in mobile apps has revolutionized the ways we access and exchange information. The new lightweight ways of tablet computing have made operating system (OS) giants like Microsoft and Apple to adopt a touch-based interface over their Windows 8 and Mac OS X Mountain Lion respectively.

The moves by Microsoft is an obvious sign that consumerism of traditional desktop computing is phasing out as more of its user base is shifting towards mobile platform. The design and development of mobile applications would need to be more rapid and rigorous to meet the ever growing number of mobile users and their demand for apps. For the past 4 years from 2008 to 2012, both Google Play store and Apple App Store has each hits the milestone of 25 billion downloads of mobile apps (Newton, 2012). Based on the Android Market Insights report (Research2guidance, 2011), there are over 70,000 active Android apps publishers at the end of September 2011. Android developers are much more active app producers compare to other app developers in Apple App Store, Microsoft Windows Phone 7 Marketplace and Blackberry App World. On average, Android publishers have published at least 4.38 apps in 2011. For instance, there have been 42,000 new Android apps being published within a single month of September 2011. Interestingly, these record-breaking volumes of Android apps are also notably having the highest removal rate. 37% of the Android apps were deactivated and subsequently removed due to its inferior quality. The Android based mobile apps are widely regarded as having poorer quality of use and lacking in aesthetic appeal (Venturi, 2011). However, the perceptual complains about unusable mobile apps was not only limited to the Android's platform. The recent apps-crashed analytic studies by the analysis firm Crittercism, has revealed that iOS based mobile

apps also crash like Android apps (Geron, 2012). In September 2011, Apple App store too has removed 24% of apps despite Apple having the most stringent submission requirements for all its apps (Research2guidance, 2011).

A usable mobile app is critical as more important business transactions and production tasks are moving onto mobile platforms. The lack of usable qualities in mobile apps is a design issues regardless of platform. It is all back to traditional software engineering challenge where quality of use in an application is not being thoroughly evaluated before deployment (Hooper, 2012). With the convenience of self-publishing feature offered by various app stores, most mobile apps developers would release their apps without formal testing or usability evaluation. They depended on the crowd of end-users as their beta-testers to identify any inherent bugs and problems for them (Yap, 2012). With the gathered feedbacks, the developers would then push the revised version of their app at much later stage as "new update" in the app store. No doubt such direct users' feedback are cost effective but the "launch then fix later" approach would backfire where end users get put off and are unwilling to re-engage with apps that have poor usability. This could explain why most apps have low retention rate and would lose 76% of its users after three (3) months of launch (Flurry Analytics, 2011).

Tim Shepherd, a Senior Analyst at Canalys, has acknowledged that low quality apps in the market are caused by the tough economic condition found in mobile apps development (Yap, 2012). The growing demands and competitions of mobile apps have put most development teams on average to output 6.4 releases of apps per year, while 29 percent of other developers are expected to deliver ten (10) or more new

releases of apps within a year (Dubie, 2012). The extreme working pace in mobile apps development has meant that these developers have to release almost one new app every month. A typical usability evaluation exercise for a mobile app would require an additional 30-40 percent of a total development time, and the use of time could go up to 60 percent if the complexity of an app increases (Yap, 2012). The escalating development cost and narrowing profit margins in making apps have left developers no choice but to skip usability evaluation for meeting deadlines and budget. However, the mounting pressure to deliver high-quality apps within budget constraint is not the only reasons why developers forgo usability evaluation. The survey study led by Coleman Parkes has revealed that most software enterprises are not equipped with the right resources and methods to effectively evaluate the usability of their apps within a tight environment (CA Technologies, 2012). In the survey study, there are more than 56% of developers who reported that the existing application development and testing methods are out-dated and not efficient for shorter development cycles. 70 percent of the three hundred and one (301) respondents in the study have projected that the quality apps could be further increased if there is a more agile method for better quality assurance.

The concept of usability thinking is not new. However, the usable experience of mobile application is improving at a surprisingly slow rate, as the classic triangle development model of “cost, quality and time” is getting difficult to attain. Based on the mobile usability studies in 2009 and 2011, Nielsen (2011) has disappointingly reported that mobile usability has only improved three percent over the years (from 59%-62%). Most users still face severe usability challenge when utilizing their mobile applications. Nielsen (2011) has predicted that mobile usability would only reach the

84% of high usability rate of current desktop computer if the advancement of mobile usability moves within the same pace by year 2026. The practice of usability evaluation for mobile app is still a young research area, as the previous research trend has suggested 61% of mobile research prioritized on mobile system engineering than focused on the actual aspects of mobile usability (Kjeldskov & Graham, 2003). As a result, the issues of mobile application usability would continue to be overlook by most developers.

1.2 Research Background

Usability is an important term in Human-Computer Interaction (HCI) as many efforts have been put in to defining the term in its broadest means. According to International Organization for Standardization (ISO 9241-11, 1998), usability has formally defined 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." The term usability usually refers to the quality of an interactive software application that is easy to be used, learned, understood, and attractive to its user under specific conditions. In software engineering, usability is about "quality in use" which enabled specified users to achieve specific goals with its *effectiveness*, *productivity*, *safety* and *satisfaction* in specified environment (ISO/IEC 9126-1, 2001; Bevan, 1995). Both the usability definitions have the similar descriptive components which are all context dependent: specific user with specific goals of use for specific environment (Newman & Taylor, 1999).

The idea of usability was first conceived under the heading of human ergonomics for designing usable computing terminal (Shackel, 1959). The idea was further developed into the concept of "ease-of-use" (Miller, 1971) and later fully expanded into its own distinctive research area in HCI (Bennett, 1979; Shackel, 1981). The objectives of usability are mainly to improve the usable quality of interactive systems, and focus on how well a user can learn and use that particular system in achieving their goals with higher level of satisfaction. Besides the broad design aim of advocating quality interactive systems, usability is also a benchmark for measuring users' experience when interacting with electronic product or system like website, software application, mobile technology and any other user-operated devices (U.S. Department of Health and Human Services, 2006.).

Usability is a philosophical belief in designing to meet user needs with the utmost pleasing experience of use (Quensenbery, 2003). The concept of usability would need specific methods and processes to translate these "intangible values" into design. The practice to improve the usable quality of an application begins by identifying its existing level of usability. The process requires analysing a set of qualitative and quantitative responses that derived from the end users' behavioural actions and their state of satisfaction while they interact with the application. Such inquisition process is known as usability evaluation based on the five most common usability metrics, which formalized by the United States Department of Health and Human Services (*See Table 1.1*).

Table 1.1 The 5 General Usability Metrics

Dimension	Definition
Ease of Learning	How fast can a user who has never seen the user interface before learn it sufficiently well to accomplish basic tasks?
Efficiency of use	Once an experienced user has learned to use the system, how fast can he or she accomplish tasks?
Memorability	If a user has used the system before, can he or she remember enough to use it effectively the next time or does the user have to start over again learning everything?
Error frequency and severity	How often do users make errors while using the system, how serious are these errors, and how do users recover from these errors?
Subjective satisfaction	How much does the user like using the system?

Source: *What Does Usability Measure?* U.S. Department of Health and Human Services. http://www.usability.gov/basics/index.html

Based on the above metrics, many approaches have been derived to critically evaluate the quality of interactive systems. Within the metrics of quality, two types of data can be obtained during an evaluation study: user performance data (what actually happened during user interaction) and user preference data (what users thought after the interaction). The return of either type of the data during a usability test will then become a set of validated guidelines to improve the usability of application software. The above 5 general usability metrics are based on the ISO definitions, and in practice could be further distilled to suit the contextual needs of an application. For instance, Quesenbery (2004) has proposed the *5Es,* a more

generic set of usability metrics that are more user experience focused and less complicated to use (See Table 1.2).

Table 1.2 The 5Es of Usability Metrics

Dimension	Definition
Effective	How completely and accurately the work or experience is completed or goals reached?
Efficient	How quickly this work can be completed?
Engaging	How well the interface draws the user into the interaction and how pleasant and satisfying it is to use?
Error Tolerant	How often do users make errors while using the system, how serious are these errors, and how do users recover from these errors?
Easy to Learn	How well the product supports both the initial orientation and continued learning throughout the complete lifetime of use?

Usability evaluation or testing is a scope of activities that focuses on observing users working with a product, performing tasks that are real and meaningful to them (Barnum, 2011). The term testing or evaluation is a form of interaction studies that differ in their sample size. In usability evaluation, there is a collective set of techniques that is used to assess the usability of product, by focussing on how well its users can complete specific and standardized tasks during their first time use with the product (Cooper, Reimann & Cronin, 2007). Usability evaluation was started as part of experimental design and only became a formal process during the 1990s. Then, the tests were conducted by "usability experts" who typically are trained as cognitive scientists, experimental psychologists or human factor engineers (Barnum,

2011). By large, usability evaluation is an expensive research activity that is rigorous and time consuming. It is a lab-based research experiment that required 30 to 50 testers and as a result, not many developers could afford usability testing (Barnum, 2011). The process of usability evaluation usually takes place in the later stage of design cycle, and it requires a near complete and coherent design artefact to test against. The techniques of usability evaluation could be classified into these two main types:

- **Formative Usability Evaluation Methods**

  Formative evaluation is a quick "find-and-fix" qualitative diagnostic technique, which focuses on identifying usability issues of a product before it is completed (Reddish et al, 2002). The method is based on repeated small studies during a development. The entire evaluation process does not require the input of end users, but depends on the judgement of usability experts.

- **Summative Usability Evaluation Methods**

  Summative evaluation is a quantitative study that uses statistical significance in summarizing overall usability of a completed product. The study requires broad sample data for establishing statistical validity. In practice, the evaluation process usually is conducted and documented by a group of third-party professional moderators (Cooper et al, 2007). It requires the presence of end users and the evaluators for moderating the end users' interaction with the evaluated user interface.

In his book *Usability Engineering*, Nielsen (1993) provides a distinction between formative and summative evaluation; summative evaluation is the testing of completed product, whereas formative evaluation is an inspection method for diagnosing a product's usability that is still in development. Both streams of technique can be conducted in fixed lab environment or in the field with portable equipment (Koyani, 2006). The aim of usability evaluation is to identify or predict usability problems of a user interface by checking people's inputs against the established usability metrics. Alternatively, the process of evaluation can be done remotely through Internet or distance communication with or without any forms of moderation (Bolt & Tulathimutte, 2010).

In the design and development of a new technological application, usability evaluation is also a yardstick to measure the performance of a prototype application within a given developmental stage, whether it has met the desired level of expectations. Through this research, the aim is to present an improved evaluation method that has been used during the real development of a mobile application for tablet device. The research would include a case study of a real-time 3D graphics viewer prototype where its usability is evaluated with the proposed methods.

1.3 Problem Statements

The heterogeneity of mobile devices and their relative fast evolution have made it very challenging for developers to design and market an app within a short period of time (Biel & Gruhn, 2010). To stay in the global competition and to stay afloat among

the swarm of applications that being marketed around the clock, usability assurance is a must for any mobile apps. However, in practice usability evaluation is a much neglected aspect in reality. To the many developers who have to wrestle with tight deadlines, a formal usability evaluation is just a luxury that they cannot afford (Nielsen, 2008). By and large, usability evaluation is still a time-consuming process where it needs to be repeatedly carried out to reliably measure users' performance and emotional response (Hussain et al, 2012).

Although there are several cost effective and agile evaluation methods which derived from "*discount usability engineering*" since 1989 (Nielsen, 2008), none of these formative approaches contributed to better cost efficiency. A formative evaluation is often cost effective through the use of a relatively small sample study (4-5 evaluators) as compared to a summative evaluation that requires 30-50 testers. To date, there are three most widely used formative evaluation methods, namely *cognitive walkthrough* (CW), *heuristic evaluation* (HE) and *task analyses* (TA). The formative evaluations are documented to have reliability issue such as *evaluator effect* that found in HE, where different evaluators evaluating the same application with the same method would derive at different concluding results (Jacobsen, Hertzum & John, 1998; Hertzum & Jacobsen, 2001). Both Hertzum & Jacobsen (2001) have summarised that this is due to vague evaluation procedures and problem criteria for the evaluator's reference during the process of evaluation. Of the 102 papers that Kjeldskov & Graham (2003) have reviewed, 41% of the usability research is based on empirical approach of "trial and error" and without any formative usability criteria. Intrinsically, most findings in formative evaluations tend to fall short

of being reliable and not significant enough to facilitate design decisions to improve the usability of an application (Hornbaek, 2005). Through his active review of 180 usability studies from core HCI journals and proceedings, Hornbaek (2005) has concluded that most evaluation methods are weakened due to its sole reliance or conflation use of either objective or subjective approach. The objective approach is an analytic-base measurement, where it has been widely used for measuring effectiveness and efficiency through quantitative means, such as task completion time (Cockton et. al., 2003). The subjective approach, on the other hand is an empirical-base study of usability evaluation methods that use qualitative interpretation to gauge users' subjective satisfaction through their interaction with an interface (Cockton et. al., 2003).

From another perspective, the evaluator effect that is found in the formative evaluation could be traced back to the usability guidelines itself. By and large, most usability evaluation methods were based on previously established guidelines that existed way before the emergence of Apps Culture. These usability guidelines were meant for desktop applications and websites, and might not be valid in the context of mobile apps (Zhang & Adipat, 2005). The distinct features of mobile computing devices, such as screen resolution, mobility, and its input model etc. have created a whole new usability challenge, which cannot be addressed with the former standards (Gómez, Caballero, & Sevillano, 2014).

The reviews from mobile HCI literature have evidently suggested that the existing usability evaluation methods lack contextual validity to be applicable for the assessment of mobile applications (Zhang & Adipat, 2005; Gómez et. al., 2014). The

referred contextual validity is about the evidence of external validity about the inspection procedure of a particular study that can be effectively applied and replicated across different kind of apps by the usability analysts confidently. As long the evaluation methods are not relevantly valid, the findings will not be reliable. The evaluator effect that is inherent in the formative evaluation has caused the efficiency of the methods to be heftily discounted, as the methods would take much longer time and higher cost than expected. More resources are necessary to establish larger quantifiable experiment for reliable findings (Kock, Biljon & Pretorius, 2009). Hence, there is a need to improvise and adapt the existing formative evaluation methods.

1.4 Research Questions

The scope of this research will focus on usability evaluation methods for mobile app, which tailored for shorter time-to-market development model. The objective of the research will centre on how to reliably evaluate the usability of a mobile application with greater cost-effectiveness, based on three key research questions (RQ):

RQ1. How to improve the overall cost-effectiveness of usability evaluation for mobile app?

- With the given time and cost constraints in present mobile application development pipeline, cost-effective evaluation method that has low overhead cost and less time consuming is critical to advocate the practice of usability assurance to mobile app developers.

RQ2. How to increase the reliability of a small sample usability study for mobile application?

- A reliable evaluation method would contain no discrepancy in its finding even with small sample study. As result, the cost and time invested in usability evaluation could be reduced, as smaller study is generally less resource intensive.

RQ3. Are the existing usability heuristics still valid for evaluating mobile application?

- A reference to relevant usability guidelines is essential as valid usability criteria are the cornerstone to valid inquisition.

1.5 Chapter Overviews

The emergence of mobile devices and app culture has created a whole new range of innovative products with fresh new usability challenges. The traditional methods of usability evaluation design and its standard practices are still worth to be revisited to pave the future construct of better evaluation methods. In Chapter 2, the thesis would review the established formative usability evaluation methods (UEMs) of *cognitive walkthrough, task analysis* and *heuristics evaluation*. In Chapter 3, the thesis would focus on mobile usability requirements and the proposal of an improvised evaluation method in this research. A case study would be presented in Chapter 4 with the application of the previously proposed method. The findings of the case study will be reported and followed up in Chapter 5. And lastly, the research would summarize and discuss the findings of the proposed method in Chapter 6 and 7

Chapter 2

Formative Usability Evaluation Methods

2.1 Introduction

Usability evaluation is a form of user context analysis, which mainly draws from the direct observations of users' task performance when they are interacting with an application. A usability evaluation can be a quantitative experiment or a formative qualitative study which comprises of large or small sample sizes (Nielsen, 1992; van Greunen & Wesson, 2002). The methods to evaluate usability are usually a complex construct which would need a considerable amount of resources to be administered. With the given constraints of costs, there is usually only one evaluation method used in many occasions (Kock et al, 2009). Thus, it is important for every app developer to select the appropriate usability evaluation methods (UEMs) that is the most cost effective.

The referring costs in usability are the total cost that has been spent on an evaluation cycle, where the definition of costs are based on man-hours (time) and the value of money that has being used during the evaluation activities. The costs of usability evaluation are usually recurring spending, as building usability is an iterative process. To determine the usable quality of an application, the evaluation process would involve several repeated testing sessions. Based on the data collected from 863 usability design projects, Nielsen has found that, on average, the costing of usability would siphon an additional eight to thirteen (8-13) percent of a project's total budget (Nielsen, 2003). Although the cost of usability testing does not increase linearly with

project size, Nielsen advocated that it would be best to devote additional ten (10) percent of a project's budget to usability testing (Nielsen, 2003). The spending on usability largely goes to a series of evaluation activities, which include:

- Planning of evaluation process
- Creating test tasks, recruiting test users and evaluators
- Analysing the evaluated results
- Preparing the recommendation report for revising the application

The similar process would then be repeated to find out whether the revised application is more usable as compared to its previous version. Based on a documented usability evaluation experiment by the Technical University of Denmark, the average time spent for evaluating a website's interface was 39 hours (Nielsen, 1998). The evaluators of the experiment comprised of fifty (50) teams of user interface design students. Prior to the experiment, the students underwent 15 hours of training in user-test methodology. This would equate to a total of 6.75 workdays if both the evaluation time and training hours were to be combined and calculated. These man-hours are an upper estimate of the required time for a first run of usability test, and the investment of time could be reduced to two (2) work days if experienced evaluators were being employed (Nielsen, 1998). This is an inevitable part of the usability testing, and this would drive up the total cost of usability testing as real work costs real money.

In the evaluation of mobile application, the development team can either self-manage the evaluation or outsource it to a usability consulting firm. To assemble an internal usability team for short-term used can be very costly, as it would involve the

one-off expenditure in setting up a usability lab. A two-room usability lab that is furnished with one-way mirror and testing equipment can cost as much as USD 100,000 in the early 1994 (Barnum, 2011). As the technology advances according to Moore's Law (Moore, 1965; Hutcheson, 2005), the cost of computing equipment is getting lower, and the comparable quotes to build an usability lab today is in the USD 25,000 range (Barnum, 2011). This one-off setup cost is not inclusive of the expenditure of hiring usability testing personnel. According to Nielsen (2012), a usability testing staff with five years' experience would cost as much USD 84,000 per annum to be hired. Although the salaries of usability testing practitioners are lower outside the United States, it would not be any lower than other regional standard of an IT professional's earning.

On the other hand, the outsourcing solution for usability evaluation is not any cheaper as well. In fact, the price tags for outsourced evaluation activity are quite steep. For instance, the leading usability consulting firm - the Nielsen Norman Group, which was co-founded in 1998 by the renowned usability gurus: Jakob Nielsen, Don Norman and Bruce Tognazzini, has their services priced between USD 10,000 - USD150,000 per project (See Figure 2.1).

	Type of Usability Evaluations	Price (USD)
Formative	Qualitative Usability Tests	$20,000-$40,000
	Iterative Design Usability Tests	$40,000-$70,000
Summative	Competitive Benchmarking	$50,000
	Quantitative Tests	$70,000
	Remote Usability Testing	$10,000-$70,000
	International Tests	$50,000-$150,000

Source: http://www.nngroup.com/consulting/usability-evaluations/

Figure 2.1 Pricing for Usability Evaluations, Nielsen Norman Group

With reference to the above offered services, there is no best possible evaluation method. Each type of usability evaluation methods has its advantages and disadvantages. In a tight development framework, the best applicable method to evaluate mobile application would probably be the quickest and the most inexpensive method that meets the developer's timeline and budget for their product. Based on the price list above, the five different evaluation services can be broadly classified into two distinct categories: formative and summative methods. These categorical terms of formative and summative are mainly adopted from assessment design in education (Scriven, 1967), where they are used to classify the approach that set to evaluate a student's learning. In education, the formative methods use a test and its result to inform one's learning with immediate feedback for self-improvement, whereas summative methods use a test's grade to summarize how much a student has learnt. The nature of assessment design in education is actually quite similar to usability evaluations; the only difference is that the test subject in usability is an application and not students or the user himself/herself. In the context of usability, the summative UEMs are set to measure how usable an interface is, whereas formative UEMs are used to identify what is not usable within an application (Sauro, 2010a). During the preliminary study of this research, the costing issue in usability evaluation is mainly driven by these two variables:

- Types of usability evaluation methods (UEMs);
  - E.g. formative or summative approach
- Operational cost;
  - E.g. man-hours (evaluators), equipment and lab for usability evaluation activities

The costing for operational equipment in usability evaluation is mostly an one-off expenditure, which can be brought down and substituted with other alternatives. For instance, the expenditure to furbish a usability lab is approximately USD 3,800 as quoted in Table 2.1.

Table 2.1 Equipment costing for desktop-based usability evaluation.

Equipment	Vendor	Cost (US$)	Quantity	Total (US$)
Desktop Computer - S30 Workstation with display unit	Lenovo	1,154.99	1	1,154.99
Webcam - Logitech Webcam Pro 9000	Logitech	49.90	1	49.90
Visualizer - Elmo P10 XGA Visualiser	Elmo	2,184	1	2,184
Tablet (Testing Device) - Nexus 7 16 GB Tablet - Wi-Fi only	Google	339	1	339
Total				$3727.89

The provided quotation is an actualized do-it-yourself desktop setup where this research will be using (See Figure 2.2). These operational set comprises of a set of recording devices, which are used to record the evaluation sessions. A recorded session would allow the evaluator to revisit a particular usability problem and highlighted it in their reports and presentations (Barnum, 2011). In some situations, specialized sound recording equipment would be needed if the evaluation focuses on voice-based interaction system, such as the application for personal voice-assistance.

Hypothetically, the costing of the equipment can be further reduced if the research adopts a portable solution than the current desktop-based setup. The portable solution could be setup with less than USD 2,500. This is done by substituting the digital visualizer with a mobile model, as quoted in Table 2.2.

Figure 2.2 A desktop-based usability evaluation setup.

Table 2.2 Portable equipment solutions for usability evaluation.

Equipment	Vendor	Cost (US$)	Quantity	Total (US$)
Laptop (with built-in webcam) - Dell XPS XPS15-9375sLV	Dell	1,110.46	1	1,110.46
Luggage bag - Samsonite Spinner Boarding Bag	Samsonite	93.99	1	93.99
Mobile Visualiser - Elmo MO - 1 Visualiser	Elmo	624	1	624
Tablet (Testing Device) - Nexus 7 16 GB Tablet - Wi-Fi only	Google	339	1	339
Total				$2167.45

The portable equipment pack enables usability evaluation to be carried in the field without confining to a physical space. Although such bare-minimum setup is comparatively more budget-friendly, it actually costs more as the additional travelling time and expenses in commuting to different locations for data-collection must be factored in as well. Besides the procurement cost for equipment, the other heavy expenses in usability evaluation are mainly the pay-outs for man-hours supporting the evaluation activities. The man-hours expenses are basically a floating

operational cost that scales according to the selected type of evaluation methods. In the field of usability engineering, the summative UEMs are commonly known as the *test method*s. Such classification is mainly derived from its quantitative mode of practices in data collection and analyses. The summative test methods generally tend to be more cost-intensive as compared to the formative methods. For instance, based on the sample price list from Norman Nielsen Group (See Figure 2.1), the median cost for summative and formative UEMs is at the distinct price mark of USD 60,000 and USD 40,000 respectively. This is because all summative test methods required large quantity of samples to conclude its finding through discrete statistical distributions (Hofman, 2011). A typical summative usability test would be a one-to-one session that involves two (2) hired individuals: a moderator and a recruited tester (user). Each of the test sessions will have different testers and same pool of moderators separately moderating the test for at least 30-50 sessions. In its formal procedure, the usability test would continue to be run even the findings are about the same after several rounds of initial testing. As such, the spending on such mode of recurring testing would have set a total cost of $30,000 if each of the outsourced test sessions is worth $1000 and is being rendered for thirty (30) times.

The formative UEMs on the other hand, have much lower overhead cost as compared to the summative UEMs. This is largely because a standard formative evaluation session requires only one (1) evaluator who is experienced in usability without the need to recruit any real users. Furthermore, the entire evaluative process can be concluded within three to five (3-5) rounds of separate inspection by these experienced individuals. Hence, the formative UEMs are also best known as

discounted inspection methods (Nielsen & Mark, 1994). Although formative UEMs seem to be more cost efficient as compared to the summative test methods, the challenge lies with hiring the experienced usability experts as there are no standards to qualify such expertise (Chattractichart & Lindgaard, 2008), and it would not be cheap to engage one (Nielsen, 2007; Nielsen, 2012). Besides, the effectiveness of formative UEMs has long been questioned and criticised by some usability scholars and practitioners (Jacobsen, Hertzum & John, 1998; Hertzum & Jacobsen, 2001; Kjeldskov & Graham, 2003; Cockton et. al., 2003; Hornbaek, 2005; Kock et. al., 2009). Many formatively inspected findings are found to lack objectivity, as different evaluators who inspect the same exact application would end up with different opinions. By and large, usability evaluation is a systematic exercise that is based on the scientific inquisition method, where objectivity and validity are the two cornerstones. Therefore, any inquisition methods that lack reliability or validity would be rendered invalid and not fit to be used for measuring usability. Although there are some biases found in formative UEMs, interestingly, the inspection techniques for usability were deemed more favourable (See Table 2.12).

The objective of usability evaluation is to discover what is usable and not usable in a application. The gathered insights aimed to aid the developers to improve their design with better informed and conclusive decisions, which would benefit their end users. In today's applied practice of usability, there is no need to purposefully bifurcate the evaluation methods for usability as the boundary between the formative inspection and summative testing is blurring (Sauro, 2010a & 2010b); all usability activities can be simply addressed as usability testing (Barnum, 2011).

However, the high cost of gathering data for usability evaluation would render impossible for many independent developers. As such, there is a need to come out with a common framework that is beneficially affordable for all developers of mobile applications.

In the search for UEMs that are cost effective, the research has identified gaps and opportunities in the classic formative evaluation frameworks. In this chapter, the research will relook into the root of several known formative UEMs.

2.2 Taxonomy of Usability Evaluation Methods (UEMs)

As discussed in the previous chapter, a usable mobile app would need to have widely accepted usability traits, or the *5Es*: *effectiveness*, *efficiency*, *user engagement*, *error tolerant* and *ease of learn* (Quesenbery, 2004). In order to assess these five quality traits in an application, a measurement framework known as UEMs is used. UEMs are a collective set of evaluation techniques, which designed to audit the usability of a user interface. According to Holzinger (2005), UEMs can be classified into the inspection and the test approach (See Table 2.1). Both of these approaches have been long used to examine the usability of various types of user interface which range from electronic products to application software that are still under development or are about to be released. (Desurvire, Kondziela & Atwood, 1992; Hornbaek, 2005; Coursaris & Kim, 2011).

Table 2.3 Taxonomy of Usability Evaluation Methods

Usability Evaluation Methods (UEMs)						
	Inspection Approach (Formative)			Test Approach (Summative)		
	Heuristic Evaluation	**Cognitive Walkthrough**	**Task Analyses**	**Thinking Aloud**	**Field Observation**	**Questionnaires**
Applicably in Phase	all	all	design	design	final testing	all
Required Time	low	medium	high	high	medium	low
Needed End Users	None	None	None	3+	20+	30+
Required Evaluators	3+	3+	1-2	1	1+	1
Required Equipment	Low	Low	Low	High	Medium	Low
Required Expertise	Medium	High	High	Medium	High	Low

Adapted from "Usability Engineering Methods for Software Developers", by A. Holzinger, 2005, *Communications of the ACM*, 48, p.72.

The inspection approach is a set of formative techniques, which frequently used to inspect user interface that is still under development. The idea of usability inspection is actually similar to the process of quality control (QC), but instead of inspecting a product's defects, a typical usability inspection is set to identify any potential user interface problems that would hinder its users' task performance when the users are using the interface. The inspection approach comprises of several techniques like *heuristic evaluation* (HE), *cognitive walkthroughs* (CW) and *task analysis* (TA). Each of these techniques can be combined or applied separately to inspect the usability of an interface. In practice, the techniques of usability inspection are usually being carried out by two to three experienced evaluators who have the expertise or knowledge about usability. As such, a typical inspection process would not involve real users, as often the professional judgement by the evaluators is regarded to be sufficient.