Developing Quality through Measuring Usability – The UTUM Test Package

Jeff Winter, Kari Rönkkö
School of Engineering
Blekinge Institute of Technology
Soft Center
SE 37225 Ronneby, Sweden
(jeff.winter, kari.ronkko)@bth.se

Mårten Ahlberg, Mark Hinely,
Mats Hellman
UIQ Technology
Soft Center VIII
SE 37225 Ronneby, Sweden
(marten.ahlberg, mark.hinely,
mats.hellman)@uiq.com

Abstract

This paper presents a tool for building quality of use into the software and software process, in the form of a test package for mass market devices. It is developed for measuring user experience, which is seen as a more encompassing term than usability.

The test package, which is under constant development, is the result of a long term cooperation between industry and academia, and is in use in industrial development projects.

It shows usability through the use of metrics, for efficiency, effectiveness and satisfaction, supported by qualitative judgements made by a test leader/usability expert. It gives a clear demonstration of quality, from the customer and end-user point of view.

The case presented here, where the test has been performed concurrently in two countries, has been a quality assurance of the test package, which has been found to work efficiently and flexibly in a complex industrial environment, with complicated relationships between customers, partners and end-users.

1. Introduction

It is a great challenge to achieve quality assurance through a usability test package for mass market products, where economic benefits are gained through approaching the broadest possible category with one product. The test package presented in this paper, the UIQ Technology Usability Metrics (UTUM) is a usability test package for mass market mobile devices. UTUM enables quality assurance by measuring usability empirically on the basis of metrics for satisfaction, efficiency and effectiveness,

complemented by the test leader's observations.

The test package supports building quality into a mass market device, based on end-user testing and feedback, and provides important information to customers, partners, and other stakeholders in the design and development process. It is grounded in usability theory and guidelines, and in industrial Software Engineering practice and developed in a long-term cooperation between academia and an industrial partner with long experience within telecommunications field and usability. UTUM is in use within the industrial setting, and is under a process of continual development.

Although we see UTUM as a method for quality assurance, we see the case presented here as a quality assurance of the UTUM test package itself, showing that it can work efficiently in a complex environment, with many different relationships between different organisations, stakeholders, customers and end-users.

Quality is not simply an attribute of lack of defects and there are clear connections between usability and quality. Bevan [1] states that quality of use is the extent to which a product satisfies stated and implicit needs when used under stated conditions. Quality of use is expressed in the ISO standard as the extent to which specified goals can be achieved with effectiveness, efficiency and satisfaction by specified users carrying out specified tasks in the specified environment [2]. This places the focus of quality on measurable aspects of the product, the tasks and the context in which it is used, rather than on the product in isolation. This is a worthwhile approach but is still too narrow as it does not cover enough of the quality aspects related to the users' experience.

UTUM is an attempt remedy the above shortcoming, whilst still retaining the broader and measurable aspects of usability. It measures what the

industry refers to as User Experience, which is seen as a more encompassing term than usability. User Experience is directed towards the overall experience of having, owning and using a product, where usability is one part of the experience [3]. A system with good usability should give the user a good user experience. UTUM measures the user experience through the use of metrics, supported by qualitative judgements made by a test leader/usability expert. It gives a clear demonstration of quality, from the customer and end-user point of view.

For general purpose products, the fact that products can only have quality in relation to their intended purpose is a problem, and is a key reason for being concerned with user-perceived quality [1]. But telecommunications as a branch focuses on providing the market with new and improved technology rather than fulfilling end-user needs [4]. It is a challenge to develop a usability test for a mass-market product, where specific end-user groups must be targeted [5], whilst being unwilling to exclude other potential end-users [6]. UTUM is a way to meet that challenge.

The paper begins with the background to the development of the UTUM, and continues with a brief look at the relationship between quality and usability. Next follows a description of the UTUM and the methods and measures it contains, and how they are used and presented. Following that is a summary, based on some of the findings from the test series used to exemplify the UTUM.

2. The historical background.

The industrial partner is UIQ Technology, situated in the Soft Center Science Research Park, Ronneby, Sweden. The company was established in 1999, and has as one of its goals to pave the way for the creation of user-friendly, diverse and cost-efficient mobile phones. They focus on “usability design, software engineering, product realization, technical consulting, sales and marketing” [7]. Their product, UIQ, is an open, media-rich, flexible and customizable software platform, pre-integrated and tested with Symbian OS, providing core technologies and services.

The research group U-ODD – Use-Oriented Design and Development [8], is a part of the Department of Interaction and System Design, at the School of Engineering at Blekinge Institute of Technology (BTH), and is part of the research environment BESQ [9]. The research performed by U-ODD is directed towards use-orientation and is influenced by the application of a social science qualitative research methodology and the end-user's perspective.

The process of developing the Usability test package began with an attempt to introduce Personas in the development process, in order to bridge a gap between designers and developers, and other stakeholders in the company. This attempt was unsuccessful for reasons that can be found in [6].

At the same time, an early version of the user test was implemented with the same purpose, as a two part tool consisting of a number of use cases to be performed on a mock up [4], and the use of the System Usability Scale (SUS) [10]. The test was performed at several places in the development process, in order to show improvements in usability, but the results were seen as somewhat predictable and did not contribute greatly to the development process in practice.

The first UTUM version was developed in 2005, consisting of a questionnaire used to prioritize use cases, a performance metric based on completion of specified use cases, and an attitudinal metric based on the SUS. Usability was shown as a value between 1 and 100, but there was still no way of taking the test leader's observations into account. In 2005 and 2006 new versions of the test were developed, which include more metrics, and a new way of presenting the results in a 4-field diagram. The current version of the test is described more fully later in the paper.

3. Quality, usability, and metrics

Bevan [1] states that the objective of software development is quality of use, and software product quality is the way to achieve it. Many aspects of software quality contribute to quality of use, but ease of use is often critical for interactive software. Therefore we will now look at some of the connections between quality and usability, and metrics for illustrating them.

Quality is a complex concept, meaning different things to different people, and in order for others to understand your view of quality, it must be defined in a measurable way [11]. There are however many reasons why it is unclear what a good measure of software quality is. Wong and Jeffery, for example, found that developers and users have different cognitive models of what software quality is, with differences in their choice of characteristics and consequences and in their desired values [12].

Therefore, there is a variety of interpretations of the meaning of software quality, the meaning of the terms to describe its aspects, of which aspects should be included in a model of software, and which software development procedures should be included in the definition. Recent research has adopted the user-based view of quality, recognising that each person is

different and has a different perception of quality [12].

Kitchenham and Pfleeger state that the user view is concrete, and is grounded in product characteristics that meet the user's needs. Users are concerned with aspects of usability, not only with reliability. Studying usability is also related to the user view, as researchers observe how users interact with products [11].

It is apparent that there are many meeting points between the user view of quality and usability. However, confusion also exists as to the meaning of the term usability. This is one of the reasons for problems incorporating usability engineering into the field of software engineering (see for example [1], [13]). Bevan distinguishes two separate approaches to usability, reflecting the problems that are apparent in divergent ideas regarding quality. One is a product-oriented view, which sees usability as ease of use, whilst the other is a "top-down" approach, where usability is seen as the ability to use a product for its intended purpose. In the product-oriented view, which fits well in conventional software engineering, usability is seen as an independent contribution to software quality, whilst from a human factors viewpoint, usability can only be achieved through a process of user-centred design [1].

Metrics are an important factor in finding acceptance for studies of quality or usability. When incorporating usability in a quality system, measurable usability requirements need to be specified, and the fulfilment of goals must be monitored during design [1]. Metrics are important in development processes, and measuring usability requirements leads to greater acceptance for usability work [14]. In a usability test, it is important to measure all three aspects of usability; satisfaction, efficiency and effectiveness rather than simply assuming that they are correlated [15].

4. UTUM

This section presents a brief overview of the test package. UTUM measures quality of use and usability empirically, on the basis of metrics for satisfaction, efficiency and effectiveness, complemented by the test leader's observations and comments. The following key terms are used in accordance with ISO 9241-11:1998 [2]: Usability, Effectiveness, Efficiency, Satisfaction, Context of use, Goals and Tasks.

As an indication of the importance given to UTUM, it was presented to the telecom industry at the Symbian Smart Phone Show in London, October 2006 [16], and a more detailed explanation of the test is available on UIQ Technology's website [17].

4.1 The case

In the test series used to exemplify the test package, the testing was performed by two test leaders, in Sweden and England. It was the first time that the test had been performed by anyone from outside UIQ Technology, and by someone who has not been involved in developing the test package. The English test leader represented a large multinational company who is involved in the development process in the roles of both partner and customer.

The English test leader came to Sweden to learn the method, in order to perform testing independently in England. She was introduced to the method via a visual presentation and discussion, where the composition of the test group was decided. The test group consisted of 48 testers, with 24 testers in each country. They also decided the devices to test and chose the 6 use cases to be performed.

The Swedish test leader performed 4 tests in Swedish with the English test leader as an observer. She then performed a test in Sweden with an English speaker, which the Swedish test leader observed.

The English test leader then performed 24 tests in England, whilst the Swedish test leader performed his remaining 20 tests. The Swedish test leader went to England and observed 4 tests, to see if they had performed the tests in the same manner, and to be able to judge the similarity of the testing and the findings.

The test leaders discussed their findings, and were given access to one another's test data and comments. Data from England was taken to Sweden, and the quantitative measurements were summarised. Data from Sweden were sent to England, and the qualitative data were summarised in a different spreadsheet.

Whilst in England, they discussed and summarised the findings, and planned a telephone conference with the system architect in England and the interaction architect in Sweden, to present and prioritize most important findings. This took place 14 days after the testing period had begun, and provided systems and usability architects with direct input for design decisions. After that meeting, they began to write a report from the findings and notes from the tests.

4.2 The test process

The tester completes a questionnaire collecting data that can have impact on the results, such as information about the tester and which applications are most important. This data can also be used to choose use cases for testing, based on the tester's use patterns. The tester has time to get acquainted with the devices, and then fills in a Hardware Evaluation (HWE), which is a

questionnaire regarding attitudes to the look and feel of a specific device. This evaluation is a new feature of the latest version of the UTUM.

The tester then performs all of the use cases on one device, whilst the test leader times the performance, observes the user's experience, and makes notes about setbacks or problems, or other personal observations about the use case. The time taken is used to calculate the **relative efficiency** for each use case, and the judgement is used to calculate the **specific efficiency**.

The tester fills in a **Task Effectiveness Evaluation** for each use case, based on the user's opinion of how well the phone lets the user perform the use case.

When all of the test cases are performed, the tester fills in the SUS questionnaire [10] which gives an attitudinal metric representing the user's subjective impressions of how easy the phone is to use. The SUS is a quick and dirty usability scale based on ISO 9241:11 [2], resulting in a number that expresses a measure of the overall usability of the system as a whole (see [18] for a detailed description). This is used to calculate the **User Satisfaction Metric**.

4.3. The metrics, their use and presentation

Calculating an average of the **relative efficiency** and the **specific efficiency** gives the **performance efficiency metric** which is a response to the statement "This telephone is efficient for accomplishing the given task".

The **task effectiveness metric** is based on the answers to the **task effectiveness evaluation**. It is determined by looking at each use case and determining how well the telephone supports the user in carrying out each task, and is in the form of a composite response to the statement "This telephone provides an effective way to complete the given task".

The **User Satisfaction Metric**, calculated on the basis of the SUS is a response to the statement "This telephone is easy to use".

The **Total Usability Metric** is a combination of user satisfaction, performance efficiency and task effectiveness. These parameters are expressed on the same scale, but are actually different in nature, and a sum of these results would not express meaningful information. Instead, the method chosen to present the results is to contrast two of the parameters at a time, to illustrate a comparative **effectiveness-efficiency metric** for a use case and a comparative **satisfaction-efficiency metric** for each phone.

To illustrate the effectiveness-efficiency metric, **design effectiveness**, the average of the task effectiveness metric for all of the test cases for a single use case, is contrasted with **task efficiency**, the

average of the performance efficiency metric for all of the test cases for the same use case (see figure 1). Increased usability shown as a move towards the upper right of the quadrant.

Figure 1. The effectiveness-efficiency metric

To illustrate the satisfaction-efficiency metric, the average of the user satisfaction metric for all test cases is contrasted with the average performance efficiency metric for all of the use cases and all test cases (see figure 2). Increased usability is shown as movement towards the upper right of the quadrant. The satisfaction-efficiency metric is referred to as **Total UTUM**, and is seen a useful illustration of total usability.

Figure 2. The satisfaction-efficiency metric

The results of the testing, in the form of data concerning the individual testers, the results measured in the tests, and the test leader's observations, and testers' comments, are stored in spreadsheets. A

selection of this data can be used to show different aspects of quality in use to different stakeholders.

They can e.g. show Total UTUM for a complete test series, Total UTUM by gender, Total UTUM by age groups, the correlation between Hardware Evaluation and satisfaction, or between effectiveness and efficiency. This is a selection of all of the possible ways of using the data to illustrate different facets of quality in use, and the presentations can be adapted to stakeholders' needs. See figure 3 for an example.

Figure 3. Total UTUM

5. The complexity of the testing process

The test situation in this case is complicated. To begin with, there are complex relationships between customers, clients and end users, in different degrees of dependencies, and complexities regarding how and where the results are to be used (see figure 4).

One company, UIQ Technology, is central to the process. In this case, the testing was performed together with an English branch of a multi-national company, who can be regarded as both a new partner in a different country, and as a customer.

Each organisation contributed one test leader, who performed UTUM tests with testers in their respective country. To reach equivalent results this requires the test leaders to work in a compatible way. The test leaders must also cooperate, in the summarising and interpretation of results.

The testers have the important role of both end-users and customers for both organisations, irrespective of which country they live in. The English users are customers of both the English and the Swedish organisation, and vice versa.

The results of the user testing must be compiled and summarised and translated into tasks that are then

divided between the two different organisations to be incorporated into the development processes in the separate organisations.

The requirements that arise as a result of the testing must also be allocated to different phases of the development process, with some being implemented directly in the current development cycle, whilst others are placed in the next development cycle. They must also be placed on two levels, where some of the results have to be incorporated into UIQ as a platform, and some must be on the product level.

Figure 4. A complex flow of tests and results

As a whole, this shows the complexity of the testing situation, and the demands that are placed on the UTUM if it is to work as a quality assurance tool.

6. Summary and conclusions

To summarize, based on some of the results that we have arrived at after the testing period:

Quick and efficient. The testing phase was short. The most important findings had been reported in to the design process and the work of including them in the development process had begun only 14 days after the testing began. This is on the basis of the conclusions reached by the test leaders. The benefit of the continued recording and analysis process is that it

gives a statistical confirmation of the early findings that are made. It is also useful for stakeholders to be able to see figures of the findings that have been reached.

Transferable. The fact that the new test leader could quickly learn the test method and return to England and perform the testing with apparently equivalent results seems to show that the method can be used in different circumstances by people who have not been involved in developing the test package.

Handles complexity. We have seen that the test works in a complex situation, where there are complicated relations between customers, partners, end-users and technologies.

Customer driven. The features to be tested can be decided on the basis of customer requirements, or on end-user patterns. The test can give an end-user perspective to customers' questions regarding the usability and thereby quality of their product, both from a software and hardware point of view.

More quantitative data. The basis in metrics should mean that the test results can more easily find acceptance in the software development process. The amount of measurement is increasing in the test package. In this test series, there are close to 10000 points of measurement. This, combined with the qualitative aspects of the test, gives a sound basis for verifying test results.

Quality assurance. The series of tests that are used to exemplify the UTUM in this paper can be seen as a quality assurance of the UTUM method itself. This result, combined with the construction of the test, with its basis in qualitative research, usability principles, and metrics, lead us to believe that the UTUM can be used as a quality assurance tool in a wide perspective.

To sum up. The UTUM test package measures the user experience quickly and flexibly, and in so doing measures aspects of quality in use. It has been tested in a complex industrial development project and has been found to be a valuable and flexible tool that is easy for a usability expert to learn and use, to measure and help build quality on the customer's terms.

8. Acknowledgements

Thanks to the people in England and Sweden who have helped in this work. This work was partly funded by The Knowledge Foundation in Sweden under a research grant for the software development project "Blekinge – Engineering Software Qualities", www.bth.se/besq.

9. References

- [1] N. Bevan, "Measuring usability as quality of use," *Software Quality Journal*, vol. 4, pp. 115-130, 1995.
- [2] International Org. for Standardisation, "ISO 9241-11 (1998): Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability," 1998.
- [3] UIQ Technology, "UTUM Report: Inside Information," vol. 2007, 2006.
- [4] M. Rettig, "Prototyping for tiny fingers," *Communications of the ACM*, vol. 37, pp. 21 - 27, 1994.
- [5] J. Grudin and J. Pruitt, "Personas, Participatory Design and Product Development: An Infrastructure for Engagement," presented at Participatory Design Conference, 2002, Toronto, Canada, 2002.
- [6] K. Rönkkö, B. Kilander, M. Hellman, and Y. Dittrich, "Personas is Not Applicable: Local Remedies Interpreted in a Wider Context," presented at Participatory Design Conference PDC '04, Toronto, Canada, 2004.
- [7] UIQ Technology, "UIQ Website." www.uiq.com/aboutus, 2006.
- [8] U-ODD, "Use-Oriented Design and Development." <http://www.bth.se/tek/uodd.nsf/>.
- [9] BESQ, "Blekinge - Engineering Software Qualities." www.bth.se/besq, 2006.
- [10] J. Brooke, "System Usability Scale (SUS): A Quick-and-Dirty Method of System Evaluation User Information," Digital Equipment Co Ltd, Reading, UK, 1986.
- [11] B. Kitchenham and S. L. Pfleeger, "Software quality: the elusive target [special issues section]," *Software, IEEE*, vol. 13, pp. 12, 1996.
- [12] B. Wong and R. Jeffery, "Cognitive Structures of Software Evaluation: A Means-End Chain Analysis of Quality," in *Product Focused Software Process Improvement: Third International Conference, PROFES 2001*, vol. 2188/2001, Lecture Notes in Computer Science. Kaiserslautern, Germany: Springer Berlin / Heidelberg, 2001, pp. 6 - 26.
- [13] A. Seffah and E. Metzker, "The Obstacles and Myths of Usability and Software Engineering," *Communications of the ACM*, vol. 41, pp. 71-76, 2004.
- [14] T. Jokela, N. Iivari, J. Matero, and M. Karukka, "The standard of user-centered design and the standard definition of usability: analyzing ISO 13407 against ISO 9241-11," presented at Latin American conference on Human-computer interaction, Rio de Janeiro, Brazil, 2003.
- [15] E. Frøkjær, M. Hertzum, and K. Hornbæk, "Measuring Usability: Are Effectiveness, Efficiency, and Satisfaction Really Correlated?" presented at Conference on Human Factors in Computing Systems, The Hague, Netherlands, 2000.
- [16] Symbian, "Symbian Smart Phone Show." London: <http://www.symbiansmartphoneshow.com/2006/>, 2007.
- [17] UIQ Technology "UTUM website," <http://uiq.com/utum.html>: UIQ Technology, 2006 - 01 - 17.
- [18] J. Brooke, "SUS - A quick and dirty usability scale," www.usabilitynet.org/trump/documents/Suschart.doc.