Measuring Software Usability

Hanan Hayat[1], Russell Lock[2] and Ian Murray[3]

Department of Computer Science, Loughborough University,
H.Hayat@lboro.ac.uk
R.Lock@lboro.ac.uk
I.R.Murray@lboro.ac.uk

Abstract

In recent years the increasingly competitive nature of the software industry has led to greater emphasis on software quality, causing software developing organizations to shift their attention towards usability, which is recognized as one of the key characteristics of software quality. The growing importance attached to software usability has resulted in a plethora of different usability conceptualisations that have led to considerable variation in testing methods throughout industry. These organizations, however, are struggling with usability testing due to the difficulties they face in choosing appropriate usability evaluation methods. This is in part due to the diversity of these testing methods and the increasingly distinctive types of software and software development life cycles. This paper will critically explore the commonly used standardized survey-based usability evaluation methods: SUMI (Software Usability Measurement Inventory), WAMMI (Website Analysis and Measurement Inventory) and TAM (Technology Acceptance Model). Additionally, a contrasting usability evaluation method 'Think Aloud' will be discussed, which is a laboratory based rather than field based usability test. The paper will then outline a possible route to ensuring organisations apply the right evaluation process for their individual development context. Finally, the paper will provide recommendations for future research areas, including the formal definition of usability concepts, existing usability evaluation methods and application to common software development lifecycles.

1. Introduction

A completed set of functionality is not enough to ensure the success of a piece of software, it is quality that determines whether software is considered either a success or a failure [1]. The importance of usability has been widely recognised by both academia and industry due to its potential to make a significant impact on the quality, and therefore success of a piece of software. Therefore, it is no longer a

luxury, as Abran et al. [2] noted in 'Usability Meanings and Interpretations in ISO Standards'. It is rather a basic quality pertaining to the productivity and acceptance of software. Hence, since 1979 when Bennett published an article in which the first recorded use of the term usability was made in the context of software development, there has been a major shift in the design, development and deployment of software applications to focus on usability. End-users of software are now recognised as playing an important role in the software evaluation process [2]. However, capturing better decisions relating to usability can generally be obtained by inviting user involvement in the software development process. In attempts to achieve that, many standardized usability evaluation methods were introduced into the field of Human-Computer Interaction. Commonly used standardized survey-based methods include SUMI, WAMMI and TAM. Owing to their questionnaire based methodology, they are fairly easy to conduct and time efficient. Nevertheless, the extent to which these methods can capture usability issues is questionable, and clearly highly dependent on the context in which they are applied. In contrast, the Think Aloud approach has been merged into the field of usability evaluation methods, offering better insight with respect to usability issues due to its laboratory-based nature. However, the time and resources allocated to conduct the method could be considered excessive for many small software developments.

This paper advocates the development of a structured approach to reason about the applicability of usability testing approaches. In order to develop an approach to provide guidance for developers in reaching an educated decision regarding usability, it is necessary to understand what should be implemented in terms of usability evaluation methods, and secondly, when in the software development lifecycle they should be applied. The paper first critically analyses four commonly used methods, with the goal of categorising them according to suitable selection criteria that will eventually lead to the matching of these methods to a given project context, and, the appropriate stage of its given software development lifecycle.

2. SUMI (Software Usability Measurement Inventory)

SUMI is a survey based method for measuring software usability. It uses a 50-item questionnaire that makes use of five defined subscales for Efficiency, Affect, Helpfulness, Control and Learnability for surveying users that relies on an attitude scale. Work on SUMI began in 1986 by Kirakowski, who was entrusted with a project that had two objectives:

- To examine the competence scale of Computer User Satisfaction Inventory
- Achieving an international standardization database for a new questionnaire.

2.1 Empirical Support and Criticism

Three types of studies have been conducted to establish its validity to test usability: (1) the industrial partners within MUSiC, (2) laboratory-based studies, and (3) theory-based validation. Thus, many early studies have recommended it. Preece et

al. [3] suggested SUMI should be adopted as a standard method for assessing user attitudes. Dzida et al [4] acknowledged it as a way of gaining a measure of user acceptance in the context of the Council Directive on Minimum Safety and Health Requirements for Work with Display Screen Equipment [5]. Davies & Brailsford [6] also recommend the use of SUMI in their series of guidelines for multi-media courseware developers.

Despite the above-mentioned benefits of SUMI, those applying the method face numerous challenges, and a number of drawbacks have been pointed out by researchers who have critically evaluated SUMI. It only allows for objective assessment therefore, the assessment process seeks to address classical end-user issues. In other words, only a small portion of the evaluation process evaluates different aspects and features of the software itself, resulting in insufficient information being made available as a basis for making a consistent decision about the validity of the software and pointing out the potential of its features. Secondly, the results generated by SUMI are often only valid under certain conditions, such as minimum sample size, and their generalizability depends on how well the software plan has been designed, and the extent to which the context of use of the software has been studied [7]. Moreover, SUMI does not offer an opportunity for the respondents to give their personal opinion outside of the structured evaluation questions stated in the questionnaire as it provides for only a fixed three row system of making responses (where agree, undecided and disagree are the only options), end-users with varied degrees of agreement or disagreement with the evaluation questions are not adequately catered for resulting in difficulties in tracking non-direct user responses, such as why and to what extent they found the task easy or difficult.

The use of fixed and inflexible questions in the SUMI method of usability testing limits its potential usage in recent software applications, because some of the aspects that SUMI questions focus on are no longer valid for end-users. For example, the statement “It is relatively easy to move from one part of a task to another” suggests that doing a single task goes through different fixed stages, which would no longer be applicable in the case of certain applications, such as interactive learning software. Moreover, the questions and the wording used in SUMI tests are often confusing. Some of the questions that are written in the questionnaires are of very high level for the sake of creating an adequately accommodative tool to which most software developers can identify their stage of development. Owston [8] stated that this high level of language use creates an efficient evaluation criterion. This can however also result in ambiguity. Questions such as “The software has at some time stopped unexpectedly” are extremely vague. The examinee is not given any details or clues of when it may have stopped; whether it stopped while entering an input, performing a task, or presenting an output.

Finally, SUMI provides an inflexible process of evaluation process as SUMI’s results need to be interpreted by experienced psychometricians who are experts in the science of measuring mental capacities and processes. This means that assessment has to be submitted to the Cork University College’s HFRG for

evaluation and rating. This not only makes the evaluation process tedious but also too procedural for inexperienced programmers. This process also raises important security concerns due to sharing nature of it that might impact adversely on their reputation as developers.

2.1.1 SUMI: Strength

- It is easy to conduct from the perspective of developers, organizations, end-users due to its questionnaire based methodology
- Its objectivity helps set a baseline for further usability tests

2.1.2 SUMI: Weaknesses

- Its dated and restricted question set prevents it being used in many developmental contexts, relegating its role to a small subset of program types
- The complex and relatively opaque process of evaluation prevents application without using the HFRG,
- Requires trust in a 3rd party by the organisation undertaking the test

2.1.3 SUMI Suitability Context

- ✓ Suitable for use in largely procedural systems which do not allow flexibility in completing tasks
- ✓ More suitable for stand-alone software than for web or mobile application based systems
- ✓ Suitable for use after development but not during unless development has taken an incremental approach
- ✓ Cannot be undertaken entirely by the development company
- ✓ Suitable to summarize perceived usability due to its objectivity

3 WAMMI: Website Analysis and Measurement Inventory

WAMMI was developed in 1996 for the purpose of evaluating the benefit of incorporating Human Computer Interaction input into the development of websites. It was originally based on the same factor structure as SUMI, defined by the same Human Factors Research Group. The current version includes 20 items that are assessed using a 5-point Likert scale which ranges from 'strongly agree' to 'strongly disagree' as the response options. The subscales remain the same as in SUMI, namely (1) attractiveness, (2) controllability, (3) efficiency, (4) helpfulness, and (5) learnability.

3.1 Empirical Support and Criticism

WAMMI is available online and the cost of running it starts at 250 Euros per test report providing fairly easy access for developers and usability specialists. Hence, it has proved a relatively popular instrument for assessing website usability, and it is a well-research validated questionnaire [9]. In addition, WAMMI provides a global measure as an overall usability score across all dimensions to make it easy

to compare usability between different sites, the entries of textual responses, cross-tabulations, and other custom analyses [10]. Its standardisation allows results to be easily compared with the results obtained from the usability assessment of other similar websites that have also been assessed using WAMMI. The validity and reliability of the single-item measure of usability given by using WAMMI were both confirmed by different studies such as, [11], who analysed the first version of WAMMI, found the instrument to be valid, sensitive, and reliable and [12], who confirmed them while assessing the usability of an online store.

However, WAMMI shares a number of drawbacks with SUMI i.e. its objectivity, its structure inflexibility, and the ambiguity that lies behind the evaluating process. Additionally, the usefulness of the WAMMI instrument is also questionable as it may not be possible to ascertain the reliability of the response rate if the number of genuine visitors to a site is unknown. A study by [13] concluded that although the results indicated user satisfaction being positively affected by a high level of perceived aesthetics, these did not lead to a high level of perceived usability, which indicated a possible deficiency with the assessment tool. The study involved examining the interaction between user satisfaction and perceived usability for various websites, which was done by assessing scores obtained from applying WAMMI. The selected websites ranged from high to low with respect to aesthetics and usability. The results suggest that a website may be aesthetically pleasing but not necessarily be so usable at the same time, or it may not be aesthetically pleasing yet it may still be highly usable. Their study supports that the objective nature of the instrument that does not provide sufficient information in order to reach a valid decision regarding usability.

3.1.1 WAMMI: Strengths

- WAMMI provides the possibility of measuring user satisfaction of websites cost effectively
- Compares the perceived usability of different sites objectively
- Provides formative evaluation information
- Allows benchmarking of average perceived usability

3.1.2 WAMMI: Weaknesses

- Although it allows the addition of a certain number of custom questions for an extra cost, WAMMI, like SUMI uses a largely inflexible set of questions
- It limits the ability of a developer to understand the user's perspective, and limits their engagement

3.1.3 WAMMI Suitability Context

- ✓ Suitable for use in a standard web usability testing context
- ✓ Suitable for use when launching but not during development unless a pilot version is presented
- ✓ Cannot be undertaken entirely by the development company

4 TAM: Technology Acceptance Model

The TAM (Technology Acceptance Model) provides another standardised usability testing instrument. It was developed by [6] based on Fishbein & Ajzen's Theory of Reasoned Action, which links beliefs, intentions and attitudes with a person's behaviour. It was designed to explain computer usage behaviour; provide a way of predicting the acceptability of an information system as well as to identify any necessary modifications so as to make it more acceptable to its users. It suggests that two main factors determine the acceptability of an information system: (1) perceived usefulness, which refers to the degree to which a person considers the use of a system can improve performance, and (2) perceived ease of use, which refers to the degree to which a person considers the use of a system to be effortless. These factors are captured by 20 items with a standardised 7-point semantic differential rating scale. The questionnaire also includes 5 items to capture the attitude towards using and 2 items for the frequency of using a system.

4.1 Empirical Support and Criticism

TAM has been examined widely by others and accepted as a valid model for predicting acceptance behaviour in the context of various information technologies and types of users [14, 15, and 16]. It has been especially regarded for its high predictive power [17]. However, TAM has been criticised for discounting the role of attitude in explaining behaviour with respect to technology acceptance [18]. [17] Even eliminated the attitude construct altogether in their version with the argument that its role is very limited. On the other hand, research by [18] examined the role of attitude strength in explaining the effect of attitude on the behavioural intention of users, and found that attitude is the most important determinant of this intention to use the system. Similar results were also found by [19] who stated that its effect is actually stronger than that of usefulness. This suggests the decision may have been controversial.

Results to the contrary for perceived usefulness have also been found in studies by others, which cast doubt on the validity of the model. For instance [20] found that usefulness actually affects attitude negatively; [21] found no evidence for any relation between perceived usefulness and attitude, and [22] found no evidence of any relation between perceived usefulness and either behaviour intention or actual use. These studies show that [17] views were not accepted universally, as contradictory results also exist.

4.1.1 TAM: Strengths

- TAM has been elaborately discussed in literature detailing the process of formatting the questions and scaling system, it offers a flexible set of questions,
- It can be tailored accordingly to the system in test,
- It allows a customized evaluation which may result in beneficial usability improvements.
- Its questionnaire based nature is time efficient and reasonably easy to conduct.

- It shares the evaluating psychometric nature with SUMI and WAMMI. Nevertheless, it is defined and discussed elaborately in literature allowing better insight understanding to the process and further criticism.

4.1.2 TAM: Weaknesses

- Some researchers share scepticism regarding the application and theoretical accuracy of the model.
- Its subjectivity raised questions regarding the validity of its results

4.1.3 TAM Suitability Context

- ✓ Suitable for use in a standard usability testing context
- ✓ Suitable for use during and after development
- ✓ Testing can be controlled entirely by the development company

5 Think Aloud

[23] defined Think Aloud as: "In a thinking aloud test, you ask test participants to use the system while continuously thinking out loud – that is, simply verbalising their thoughts as they move through the user interface." In collecting usability data, test participants are often asked to think aloud in order to gain insight into their thought processes. It is used for testing the usability of software as well as websites, interfaces, and instructional documents. One important aspect of making the Think Aloud approach effective is the extent to which the instructions are given. These instructions can range from a simple request to a more explicit one that may include certain content. A study by [24] compared the two extreme approaches, and found that although explicit instructions have a greater mental workload, they help to focus better on the actual problems and yield more utterances related to user experiences, expectations, and behaviour explanations. In this study, 16 participants were involved with an equal number divided between the two different conditions.

The Think Aloud Approach can be divided into the Retrospective Think Aloud Approach (RTA) and the Concurrent Think Aloud (CTA) Approach. The difference between the two is that in the former, the verbalisation takes place continuously whilst completing the set tasks whereas in the later, the verbalisation of the user's performance is done afterwards.

5.1 Empirical Support and Criticism

[23] Regarded this approach as "the single most valuable usability engineering method". Given that the data reflects actual usage and focuses on cognitive processes, the method has a high validity [25]. Nevertheless, it is not a method in terms of having clearly defined rules, which makes comparisons difficult [26]. Additionally, as a direct method, the Think Aloud approaches provides a rich source of data, but this comes at the expense of time consumption [27].

5.1.1 Concurrent Think Aloud

The main drawback of CTA is that by speaking concurrently, the attention and concentration of users is distracted, and users' task performance may therefore be compromised. Also, by fully verbalising how tasks are to be performed, the users may inadvertently change their approach accordingly so it can be an interference [3].

5.1.2 Retrospective Think Aloud

There has not been much work done to confirm the validity and reliability of RTA. Instead, most studies have compared RTA with other methods such as CTA based user testing rather than empirical studies, such as [25] and [28]. On the other hand, it was noted that with Retrospective Think Aloud, some information is omitted during verbalisation. Two case analyses revealed this occurs only when users struggle to complete tasks and that the reporting in these cases tend to be highly abstract and less dense. Such instances are still useful as they are indicators of possible usability weaknesses. Another weakness with it is that the protocol not only results in gaps, but also distortions [29].

5.3 Combining Both Think Aloud Approaches

The complementary nature of both methods has also led to studies that have combined both instead of adopting one or the other. For instance, [30] examined the utility of combining the two for studying dual verbal elicitation. There were overlaps in the type of utterances, and the retrospective method produced more verbalisations relevant to analysing usability. However, the combination of both methods helped to better understand usability issues. Whereas the concurrent phase yielded more usability issues, the retrospective data helped improved overall understanding of them through reinforcement, elaboration and contextual information.

5.1.3 Think Aloud: Strengths

- The novelty of data collected allows detailed results to be reached with respect to usability
- It encourages end-user’s engagement that leads to better usability related decisions which are beyond classical issues.
- The explicit and real time instructions and interactions allow gathering precise data that presumably accelerate enhanced usability related decisions.

5.1.4 Think Aloud: Weaknesses

- Highly complicated and rigid process
- High cost
- Requires trained participants and test coordinators

5.1.5 Think Aloud Suitability Context

- ✓ Suitable for use in planning stage and while developing
- ✓ Suitable for lab based testing
- ✓ Suitable in situations where significant funds are available for testing
- ✓ Suitable for standard usability context

6 Categorizing the Methods Discussed

The question of how best to assess each method is not an easy one to answer and the research itself is still at a relatively early stage. In terms of evaluation, it is important to realize that each method incorporates its own unique set of methodology and assumptions. The significance of choosing an appropriate methodology when carrying out an evaluation cannot be overemphasized. To better understand when to use which method, it is helpful to view them as part of a wider framework considering the following factors, as explored in Table 1:

- Level of priority
- Level of Resources needed
- Context of Use

Table 1: The Categorization of SUMI, WAMMI, TAM and Think Aloud

	Point & Description	**Usability Testing Methods**			
		SUMMI	**WAMMI**	**TAM**	**Think Aloud**
Priority Level	1 Aesthetic problem to be corrected if time permits	No	Yes	Yes	No
	2 Low priority usability problem	Yes	Yes	Yes	No
	3 High priority usability problem	No	No	Yes	Yes
	4 Major usability problem that is imperative to correct	No	No	No	Yes
Resources Constraints	Time	Low	Low	Low	High
	Money	Low	Low	Low	High
	Human	High	High	Low	Low
Context of Use	Standard Usability Test	Yes	Yes	Yes	No
	Perceived Usability Test	No	No	Yes	Yes

Priority of the usability issues can affect which method is chosen as well as time and cost constraints [31], and training and skill requirements [32]. Thus, simple methods may suffice for dealing with low priority issues, or when time and cost are major constraints, whereas high priority issues would ideally require more

sophisticated methods. Table 1 represent a partial, limited, view of the types of data that will be considered in the construction of an envisaged usability method framework.

7 Future Work & Conclusions

Although the outline of the framework has been limited in depth by the constraints of the paper length, once the framework has reached a level of maturity it will be developed further to provide a decision making system. The objective of the system will be to ease the process of choosing the suitable usability evaluation method under a customized individual set of constraints of a software project. Hence encouraging small to medium size organizations to adapt the usability approaches they take to the individual context of their software development processes.

8 References

1. Larman, C. (2002). *Applying UML and patterns: An introduction to object-oriented analysis and design and the unified process*. Second edition. Prentice Hall
2. Abran, A.; A. Khelifi & W. Suryan. (2003). *Usability Meanings and Interpretations in ISO standards.* Netherland: Kluwer Academic Publishers.
3. Preece, Jenny; Yvonne Rogers & Helen Sharp. (2011). *Interaction design: Beyond human-computer interaction*. John Wiley & Sons.
4. Dzida et al. (1993)
5. IEEE Std. 1061. (1992). *IEEE standard for a software quality metrics methodology*. IEEE Computer Society Press.
6. Davis, F.; R. Bagozzi & R. Warshaw. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, vol. 35, pp. 982-1003.
7. Jordan. P; B. Thomas, B. Weerdmeester & I. MsClelland. (1996). *Usability Evaluation in Industry.* London: Taylor & Francis Ltd.
8. Owston, R. D. (1987). *Software Evaluation: A Criterion-Based Approach.* Canada: Prentice-Hall.
9. Satar, Nuhizam Safie Mohammed. (2007). Does e-learning usability attributes correlate with learning motivation? Open University Malaysia Knowledge Repository. Available at http://eprints.oum.edu.my/7/1/Does_e-learning.pdf (accessed June, 2014).
10. Wilson, Chauncey. (2013). *Credible checklists and quality questionnaires: A user-centered design method*. Newness.
11. Kirakowski, J.; N. Claridge & R. Whitehand. (1998). Human Centered measures of success in web site design. Conference Proceedings. Available at http://research.microsoft.com/en-us/um/people/marycz/hfweb98/kirakowski/ (accessed July, 2014).

12. Christophersen, Timo & Konradt, Udo. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human-Computer Studies*, vol. 69, issue 4, pp. 269-280.
13. Lindgaard, G. & C. Dudek. (2003). What is the evasive beast we call user satisfaction? *Interacting with Computers*, vol. 15, no. 3, pp. 429-452.
14. Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behaviour. *Information Systems Research*, vol. 3, no. 3, pp. 173-191.
15. Segars, A. H. & V. Grover. (1993). Re-examining perceived ease of use and usefulness: a confirmatory factor analysis. *MIS Quarterly*, vol. 17, no. 4, pp. 517-525.
16. Doll, W. J.; A. Hendrickson & X. Deng. (1998). Using Davis's perceived usefulness and ease-of-use instruments for decision making: a confirmatory and multigroup invariance analysis. *Decision Sciences*, vol. 29, no. 4, pp. 839-869.
17. Venkatesh, V. (2000). Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, vol. 11, no. 4, pp. 342-365.
18. Kim, Yong Jin; Jae Uk Chun & Jacki Song. (2009). Investigating the role of attitude in technology acceptance from an attitude strength perspective. *International Journal of Information Management: The Journal for Information Professionals*, vol. 29, issue 1, pp. 67-77.
19. Agarwal, R. & J. Prasad. (1999). Are individual differences germane to the acceptance of new information technologies? *Decision Sciences*, vol. 30, no. 2, pp. 361-391.
20. Bajaj, A. & S. R. Nidumolu. (1998). A feedback model to understand information system usage. *Information and Management*, vol. 33, pp. 213-224.
21. Jackson, C. M.; N. Zinatelli, P. Cragg & A. Cavaye. (1997). Toward an understanding of the behavioural intention to use an information system. *Decision Sciences*, vol. 28, no. 2, pp. 357-389.
22. Lucas, H. C. J. & V. K. Spitler. (1999). Technology use and performance: A field study of Broker Workstations. *Decision Sciences*, vol. 30, no. 2, pp. 291-311.
23. Nielsen, J. (1993). *Usability Engineering.* London: Academic Press Limited.
24. Zhou, Xiao-Ying. (2009). Usage-Centered Design for Government Websites - A Practical Analysis to Canada Government Website. *Second International Conference on Information and Computing Science*, ICIC 2009, pp. 305-308.
25. Haak, Maaike J. Van Den; Menno D. T. De Jong & Peter Jan Schellens. (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology*, vol. 22, no. 5, pp. 339-351.
26. Boren, M. Ted. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261-278.

27. Jaspers, Monique W. M. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics*, vol. 78, pp. 340-353.
28. Gero, J. S. & H. H. Tang. (2001). Differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. *Design Studies*, vol. 21, no. 3, pp. 283-295.
29. Hoc, J.M. & Leplat, J. (1983): Evaluation of different modalities of verbalisation in a sorting task. *Intl. J. Man-Machine Studies*
30. McDonald, Sharon; Tingting Zhao & Helen M. Edwards. (2013). Dual verbal elicitation: the complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction*, vol. 29, pp. 647-660.
31. Mueller, C. J. (2009). An economical approach to usability testing. *33rd Annual IEEE International Computer Software and Applications Conference 2009*, pp. 124-129.Nielsen (1993)
32. Bak, Jakob Otkjaer; Nguyen, Kim; Risgaard, Peter & Stage, Jan. (2008). Obstacles to usability evaluation in practice: a survey of software development organizations. *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, held in New York.