AI & Soc (2007) 22:227–235
DOI 10.1007/s00146-007-0127-0

OPEN FORUM

How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability

Humberto Cavallin · W. Mike Martin · Ann Heylighen

Received: 31 March 2005 / Accepted: 29 August 2006 / Published online: 29 June 2007

Abstract This paper addresses the possibility of measuring perceived usability in an absolute way. It studies the impact of the nature of the tasks performed in perceived software usability evaluation, using for this purpose the subjective evaluation of an application's performance via the Software Usability Measurement Inventory (SUMI). The paper reports on the post-hoc analysis of data from a productivity study for testing the effect of changes in the graphical user interface (GUI) of a market leading drafting application. Even though one would expect similar evaluations of an application's usability for same releases, the analysis reveals that the output of this subjective appreciation is context sensitive and therefore mediated by the research design. Our study unmasked a significant interaction between the nature of the tasks used for the usability evaluation and how users evaluate the performance of this application. This interaction challenges the concept of absolute benchmarking in subjective usability evaluation, as some software evaluation methods aspire to provide, since subjective measurement of software quality will be affected most likely by the nature of the testing materials used for the evaluation.

H. Cavallin (✉)
University of Puerto Rico, Rio Piedras, Escuela de Arquitectura, San Juan, PR 00901, USA
e-mail: hcavallin@uprrp.edu

W. M. Martin
University of California at Berkeley, Office 384, Wurster Hall UC Berkeley,
Berkeley, CA 94720, USA
e-mail: wmmartin@socrates.berkeley.edu

A. Heylighen
Department of Architecture, Urbanism and Planning, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 1, 3001 Leuven, Belgium
e-mail: ann.heylighen@asro.kuleuven.be

Introduction

Evaluating software during development has become a very useful instrument for gaining insights into how it will be received by potentially interested users. In order to produce results that can inform the applications' development, the measurement tools used in these evaluations have to offer both reliability and validity to enable fairly comparing evaluations across applications.

This study investigates how subjective usability evaluation across applications can be affected by the conditions in which the evaluation takes place, and challenges the concept of absolute benchmarking in this type of evaluations. After briefly providing some background on comparative usability evaluation (Sect. 2) and software evaluation methods (Sect. 3), the overall set-up of our research is outlined (Sect. 4), followed by a synthesis of the major findings (Sect. 5). The paper closes with a discussion and implications for future research (Sect. 6).

Comparative usability evaluation

Testing and evaluation form an integral part of the life cycle of software. Developing computer technologies is an interactive process of which evaluation is a critical component (Sears 2003). In case of comparative software evaluations, the goal is to test software products side by side by focusing on issues related to their relative performance (Dumas 2003). When comparing the usability between different releases of the same application, the evaluation must determine relative measurements based on the usability of the features and afforqdances for each of the releases, in a way that allows weighing this output against their individual performances.

This situational study of software performance can be approached from a problem solving perspective. Completing the type of tasks presented to the participants for evaluating an application's usability clearly fits the definition of problem solving, an activity that confronts people with situations in which they want to achieve some desired state, yet do not know immediately what actions to perform in order to reach that state (Newell and Simon 1972). In usability evaluation, the impact of changes in an application is usually measured by presenting a particular problematic situation connected with the application to the participants, who are asked to use the software in solving a particular task. This research approach based on task scenarios attempts to deal with the problem of the artificiality of the testing itself. In the particular case of software applications, the participants in the evaluation are asked to solve a task similar to what they would face if using the applications in their real work environment. If the evaluation focuses on comparing different releases of the same application, participants are asked to use each release in solving equivalent sets of tasks.

Software evaluation methods

Types and objectives

Dumas (2003) differentiates between three kinds of user-based software evaluation methods: user-administered questionnaires, user observation, and empirical usability testing. The latter started during the early 1980s, a period in which computer software started to reach a user group beyond computer professionals. The first usability tests described at the first Human Computer Interaction Conference were presented in the style of experimental psychology reports, and were loaded on the experimental side (Ledgard in Dumas 2003). This approach was later challenged by the idea of usability research based on scenarios instead of experiments. User observation, on the other hand, is more suited when the evaluation takes place in situations where little control can be achieved, as is the case for applications that are better evaluated in their use environment. Finally, user-administrated questionnaires can be used either as a stand-alone measure of usability or along with any other method. Dumas points out that in the recent history of user-administrated questionnaires, there have been two objectives regarding usability measurement: (a) to develop short questionnaires that can be used to measure users' subjective evaluation of a product, and (b) to create a questionnaire to provide an absolute measurement of the subjective usability of a product. Having available an absolute measuring system for the subjective usability of a product is a particularly important aspiration in comparative usability evaluation. Crucial in achieving this purpose is the assessment of two aspects: the validity and the reliability of such questionnaires.

During the past 20 years, researchers in the area of usability measurement have proposed different questionnaires intended to measure subjective impressions of applications' performance. The best known examples of this type of questionnaires are: the Software Usability Scale (SUS) (Brooke 2004); the Computer User Satisfaction Inventory (CUSI) (Kirakiwski and Corbett 1988); and the Questionnaire for User Interaction Satisfaction (QUIS) (Chin, Diehl, and Norman 1988). These questionnaires have been designed to collect relevant data regarding user experience with applications, via various numbers of questions in a range from 10 to 71, using Likert scales ranging from five to nine levels of evaluation. A fourth type of user-administrated questionnaire, and the focus of our study, is the Software Usability Measurement Inventory (SUMI), developed by the Human Factors Research Group (HRFG) at the University College Cork.

SUMI

SUMI provides a scale that measures general satisfaction or "Global usability" for software applications. Work on SUMI started in late 1990 as one of the work packages entrusted to the HFRG within the MUSiC[1] project (Dybkjaer and Bernsen 2000). MUSiC was an European project to develop a set of metrics-based methods that can be used individually or together both to specify formal requirements for the

[1] Metrics for Usability Standards in Computing (CEC ESPRIT project number 5429).

usability of a product, and to assess whether a product meets them (Dybkjaer and Bernsen 2000). The project defined usability in terms of the quality of use of a product, and developed tools and procedures for measuring that quality. The SUMI questionnaire was first published in 1993. This survey divides user perceptions into five dimensions:

1. *efficiency* refers to the user's perception that the software is enabling the task(s) to be performed in a quick, effective and economical manner;
2. *effect* denotes whether the user feels mentally stimulated and pleasant as a result of interacting with the software;
3. *helpfulness* reflects the user's perception that the software communicates in a helpful way and assists in the resolution of operational problems;
4. *control* refers to the user's feeling that the software is responding in a normal and consistent way to input and commands;
5. *learnability* represents the user's feeling that becoming familiar with the software is relatively straightforward and that its "help" information, manuals, etc. are readable and instructive.

Additionally, SUMI provides an Item Consensual Analysis (ICA), a feature that enables researchers to calculate the actual proportion of responses from the evaluation sample and the statistically expected proportions based on the standardization data for that item. This evaluation allows discriminating the dimensions of the software that are significantly better or worse than the global software standard. The goodness of fit between the observed and expected values is summarized using Chi Square.

The raw scores obtained from the participants are collated and compared with the appropriate normative tables by an application especially designed for processing SUMI data called SUMISCO. The output is then standardized using the z-transform so that the mean score for the population is 50 and the population standard deviation is 10. According to these values, SUMI's developers set the standard for state of the art commercial software to a value of 50. Software above 50 is ahead of the state of the art for quality of use; software below 50 is behind. Based on the items and their corresponding dimensions, it is possible to report the points emphasized by the participants in their evaluations. SUMI was designed to enable researchers to evaluate the perceived usability of software systems based on an absolute benchmark instead of comparatively.

Research set-up

The data used in this paper were collected as part of a bigger set of researches designed to evaluate productivity variations between three different versions of a market leader's drafting application.

As part of the research, we measured changes in ease of use, performance, and perceived usability between the three versions of the application (V1, V2, and V2). From the beginning of the study, we defined that, in order to accomplish this goal, the study was to use both qualitative and quantitative methods for both collecting

and processing the data. As quantitative measurements, we used time tracking of the participants' performance, as well as the level of completeness of each task. For the qualitative measurements, we used SUMI and focus groups. This multi methods approach provided us with a broader understanding of the context and relevant dimensions involved in the resolution of the tasks, while operating as a mechanism for triangulating the different findings.

The comparisons took place in two rounds (Group 1 and Group 2). During spring 2003, we measured the differences in performance and subjective evaluation for Versions 1 and 2 of the application with Group 1 (G1). During spring 2004, Group 2 (G2) participated in a comparison between Versions 1 and 3. Each group was subdivided into two sub-groups (SG) (Table 1). These sub-groups allowed to implement a 2×2 design for the tests, i.e., exposing one sub-group to V1 first and then after to V2, while exposing the second sub-group to V2 first and then to V1. The purpose of this design was to neutralize the effect that order of exposure to the version could have on the results.

Participants

All subjects were current users of V1 of the application. The reason for selecting current users was to simulate an upgrade scenario to both Versions 2 and 3: V1 users had to solve problems based on their present expertise of the application using V1, and then learn and use either V2 or V3 in order to solve similar tasks as those solved with V1 (or vice-versa depending on the sub-group).

Our research involved a total of 54 users. For G1, SG 1.1 included 20 users and SG 1.2 14. The level of expertise among the users varied. A screening questionnaire, which turned out to be consistent with a post-factual evaluation of user performance, revealed three user-types: experts (15%), average users (73%), and novices (12%). In G2, there were a total of 19 subjects (9 in SG 2.1 and 10 in SG 2.2). Also in this group, the level of expertise among the users varied. Through a screening questionnaire the users identified themselves as expert (61%), average (33%), or novice users (5%) of the software. In order to make the post hoc comparisons across homogeneous sub-groups, we used only those responses coming from participants identified as experts and as average users.

All subjects were recruited by the researchers from a list of current users provided by the company that produces the application. Participants were contacted via E-mail, and agreed to participate in the research voluntarily. The company provided each participant with a free copy of Version 2 or 3 of the application (or

Table 1 Research design showing the 2 user groups, 4 sub-groups, and 3 versions evaluated

	Sub-groups (SG)	Order of tests
G1	1.1	V1 vs. V2
	1.2	V2 vs. V1
G2	2.1	V1 vs. V3
	2.2	V3 vs. V1

other equivalent application produced by the company) as compensation at the end of their participation in the study.

According to SUMI's Background notes (Background notes on the SUMI questionnaire), an analysis with tolerable precision using SUMI requires minimum 10–12 users, although evaluations have been carried out successfully with smaller sample sizes. However, the notes remark, the generalisability of the SUMI results depends not so much on the sample size itself, but on the care with which the software's context of use has been studied and the design plan has been made.[2]

Round 1 of the study (involving G1) took place at the company's headquarters, Round 2 (involving G2) in a special facility in Los Angeles (Ca). In both cases, the premises were configured to support groups of participants, working simultaneously in solving the task. The equipment used for the tests included a combination of desktop and notebook computers for G1 and desktop computers for G2. In both cases, the company determined the availability of the equipment, and freshly installed out-of-the-box versions of the applications were provided to the participants at the beginning of each set of tests.

Data collection

In order to measure productivity gains, as the original study intended, a series of tasks was designed to simulate the production-drafting environment that exists in many day-to-day business operations in professional practice. The study evaluated several newly developed features in each release of the application.

Each feature was incorporated into either an individual or a team-based exercise, depending on the type of performance to be evaluated. Every feature in the selected version of the application was then measured to determine the general ease of use and the variations in raw speed when solving the given tasks. The goal of the tasks was to compare the performance and ease of use of the new features in both V1 versus V2, and V1 versus V3.

In G1, the tasks used for each sub-group and software version stressed aspects of the application concerning changes in the Graphical User Interface (GUI) to optimize procedures involving routine actions for creating entities in files. In G2, the tasks emphasized aspects related to creating and manipulating files via a new feature that allows configuring file characteristics automatically, as well as linking and indexing files.

After ending each set of tasks, subjects completed a questionnaire designed to collect qualitative and quantitative measures of their perception regarding satisfaction and experience with the application, and regarding its global performance.

[2] The SUMI's Background notes read as follows: "the minimum user sample size needed for an analysis with tolerable precision using SUMI is on the order of 10–12 users, although evaluations have been carried out successfully with smaller sample sizes. However, the generalisability of the SUMI results depends not so much on the sample size itself, but the care with which the context of use of the software has been studied and the design plan has been made. As summarised by Macleod (see elsewhere in this volume) this involves identifying the typical users of the software, the goals which they typically wish to achieve, and the technical, physical and organisational environments in which the work is carried out (for prototype systems, this involves determining the future context of use). The design plan requires an adequate sampling of the context of use."

In addition, subjects were asked to rate the importance of specific functions of the application and their perception of these functions' performance using SUMI. The data collected at this stage constitute the basis for this analysis, and are further explored in the next section.

We expected that, when analyzing post hoc how the participants had evaluated V1 of the application, their subjective evaluations for this version would be similar across all groups and subgroups since all subjects were current users of V1.

Results

When comparing the data obtained using SUMI in G1 and G2 for both sub-groups, we found differences in how V1 is evaluated both globally and for each of the dimensions. Most surprisingly, closer resemblances exist between the means for each of the SUMISCO dimensions for V1 and the corresponding means for the compared version (V2 and V3) in both G1 and G2, than between the evaluations by G1 and G2 of the same V1 of the application.

Comparison between V1 and V2 for the expert users reveals small variations in the resulting means for each dimension evaluated. When using an Analysis of Variance, however, these variations do not show any statistical significance. The same holds for the comparison between V1 and V2 for the average users. Comparing the values that experts and average users in G1 gave for V1 does not reveal any statistically significant differences between them either.

When using SUMI to compare V1 and V3 as evaluated by the experts, we can see also small differences between the mean values obtained for each version, for each of the dimensions; the same is true when comparing the evaluations by the average users. As for G1, there are no statistically significant differences between the means for the different dimensions when comparing the evaluations by both expert and average users exposed to V1 in Group 2.

However, a major difference can be observed when we compare the results for V1 obtained from all average users, i.e., both in G1 and G2 in four of the six SUMI categories. In this comparison, the means for the SUMI evaluation of V1 produced by subjects in G1 significantly differ from the means for V1's SUMI evaluation by subjects in G2 (Table 2) in the cases of *Global evaluation* (F [1,13] = 4.291, $p < 0.1$), *Efficiency* (F [1,13] = 5.717, $p < 0.05$), *Control* (F [1,13] = 8.978, $p < 0.05$), and *Learnability* (F [1,13] = 3.665, $p < 0.10$).

Differences are also found between the means for the evaluations of V1 by experts in G1 and G2. In this comparison, the means for V1's SUMI evaluation obtained from subjects in G1 show significantly differ from the means for the same evaluation obtained from subjects in G2, be it only for the dimension *Learnability* (F [1,13] = 6.157, $p < 0.05$).

In other words, the evaluations of different versions of the application that were used for performing similar tasks are closer to each other than the evaluations obtained for the same version when performing different types of tasks. This is particularly true for the group of average users that participated in this study.

Table 2 ANOVA for both average users employing V1 in Groups 1 and 2

		Sum of squares	*df*	Mean square	*F*	Sig.
Global	Between groups	338.201	1	338.201	4.291	0.050
	Within groups	1024.732	13	78.826		
	Total	1362.933	14			
Efficiency	Between groups	247.543	1	247.543	5.717	0.033
	Within groups	562.857	13	43.297		
	Total	810.400	14			
Affective	Between groups	65.744	1	65.744	1.152	0.303
	Within groups	741.589	13	57.045		
	Total	807.333	14			
Helpfulness	Between groups	224.233	1	224.233	1.069	0.320
	Within groups	2727.500	13	60.593		
	Total	2951.733	14			
Control	Between groups	544.019	1	544.019	8.978	0.010
	Within Groups	787.714	13	60.593		
	Total	1331.733	14			
Learnability	Between groups	297.619	1	297.619	3.665	0.078
	Within groups	1055.714	13	81.209		
	Total	1353.333	14			

Taken together all results reported above, two factors could account for the differences in the evaluations of V1 when comparing G1 and G2: the variation in the subjects' level of expertise and the variation in the conditions under which the measurements took place, c.q. the different nature of the tasks that both user groups performed before each of the SUMI measurements. Experts seem to have more stable opinions than average users, whose subjective impression of an application seems to be more affected by the nature of the task they were solving.

Discussion

Our findings raise serious questions about the measurability of software usability. Considering that the SUMI method was explicitly developed to provide an absolute benchmark for measuring subjective evaluations, the results presented above came to us as a surprise. Dybkjaer and Bernsen (2000) alert evaluators that, when conducting usability evaluations, scenarios should be designed to avoid priming the users on how to interact with the system. In our study, we have found that scenarios cannot only affect the task solving level, but also prime the subjective evaluation of an application that users produce after performing tasks of a particular nature.

In our research, the differences in the nature of the tasks seem to have affected the outcome of the application's evaluations. Even though that is something to be expected when appraising applications' qualities, one would assume an instrument

intended to set an absolute benchmark as SUMI to provide evaluators with a closer measurement for the quality of applications that are identical, as in the case of V1 for this study. Judging from our results, it seems that the goal of having an absolute system for measuring perceived quality in applications is difficult to obtain, and that subjective measurement of software quality will be affected most likely by the nature of the testing materials used for the evaluation. In particular, our results suggest that the level of usability—and, concurrently, of productivity—is not so much an inherent attribute of an application or release, but rather an attribute of that application/release in combination with a certain user, the task s/he is performing, and most probably other aspects of the context of the evaluation. If we are to assess the benefits of software and technology for our ability to relate to, understand and interact effectively with others, then these assessments should acknowledge the role that the context of evaluation (in our case: the nature of the tasks) plays in the measurement of the effectiveness of that technology.

Acknowledgments This research project was sponsored by a grant provided by Autodesk Inc.

References

Background notes on the SUMI questionnaire. Retrieved. from http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html

Brooke J (2004, Sep. 30, 2004) SUS—A quick and dirty usability scale. Retrieved October 3rd, 2004, from http://www.usability.serco.com/trump/documents/Suschapt.doc

Chin JP, Diehl VA, Norman KL (1988) Development of an instrument measuring user satisfaction of the human-computer interface. Paper presented at the SIGCHI

Dumas JS (2003) User-based evaluations. In: Jacko J, Sears A (eds) The human computer interaction handbook. Lawrence Earlbaum Associates, New Jersey

Dybkjaer L, Bernsen NO (2000) Usability issues in spoken dialogue systems. Nat Lang Eng 6(3/4):243–271

Kirakiwski J, Corbett M (1988) Measuring user satisfaction. In: Jones DM Winder R (eds) People and computers IV. Cambridge University, Cambridge

Newell A, Simon H (1972) Human problem solving. Prentice Hall, New Jersey

Sears A (2003) Testing and Evaluation. In: Jacko JA, Sears A (Eds) The human–computer interaction handbook: fundamentals, evolving technologies and emerging applications. LEA, New Jersey