Mind-ing the Task
The role of context in usability research

Humberto Cavallin
University of California
Berkeley
Office 384, Wurster Hall
UC Berkeley, Berkeley, CA 94720
hcavalli@uclink4.berkeley.edu

W. Mike Martin
University of California
Berkeley
Office 384, Wurster Hall
UC Berkeley, Berkeley, CA 94720
wmmartin@socrates.berkeley.edu

Ann Heylighen
Katholieke Universiteit Leuven
Dept. ASRO
Kasteelpark Arenberg 1
B-3001 Leuven, Belgium
ann.heylighen@asro.kuleuven.ac.be

ABSTRACT
In this paper we describe our findings regarding the role of context in usability evaluation, particularly how the nature of the tasks can affect the users' perception of the performance of a particular application. Our findings show a relationship between the variation in the nature of the tasks used for usability evaluation on the one hand, and the way in which subjects evaluated these applications afterwards by using user-administered questionnaires on the other hand. These findings contradict the absolute benchmarking goal of some of these tools, thus raising questions about the possibility of achieving that kind of benchmarks in software usability evaluation, and about how comparative measurements of the benefits of software and technology take place in laboratory conditions.

KEYWORDS
Software evaluation, problem solving, user-administered questionnaires, reliability, absolute benchmarking, nature of the task

INTRODUCTION
As software applications progress, the distinction between design and evaluation is becoming more blurred, because of the involvement of users in testing prototypes of the applications very early in the software design process.
In order to produce results that can be utilized to inform the development of the applications, however, the data collected by the measurement tool used in these evaluations have to offer both reliability and validity to enable fairly comparing the evaluations across applications.
This study investigates how usability evaluation across applications can be affected by the conditions in which the evaluation takes place, and challenges the concept of absolute benchmarking in subjective usability evaluation.

COMPARATIVE USABILITY EVALUATION

Testing and evaluation is an integral part of life cycle of the software. As Sears (2003) points out, developing computer technologies is part of an interactive process of which evaluation is a critical component.

In the case of comparative software evaluations, the desired goal is to compare software products by focusing on issues related to their comparative performance (Dumas, 2003).When comparing the usability between different releases of the same application, the metrics used in the comparative evaluation must determine relative measurements based on the usability of the features and affordances for each of the releases, in a way that allows using this output to compare their individual performances.

This situational study of software performance can be approached from a problem solving perspective. The processes of completing the type of tasks presented to the participants by evaluating usability/productivity of applications, clearly fit the definition of problem solving as an activity that confronts people with situations in which they want to achieve some desired state, yet do not know immediately what series of actions have to be performed in order to reach that state (Newell & Simon, 1972).

In usability evaluation, the measurement of these aspects usually takes place in situations in which a particular problematic situation connected to the application is given to the participant in the study, who is asked to use the software in order to solve a particular task. This research approach based on task scenarios is an attempt to deal with the problem of the artificiality of the testing itself.

In the particular case of software applications, the participant in the evaluation study will be asked to solve a task similar to the one they would face if they used the evaluated applications in their real work environment. If the evaluation focuses on the comparison of different releases of the same application, then participants are asked to use each application in solving equivalent sets of tasks.

SOFTWARE EVALUATION METHODS

Dumas (2003) differentiates between three kinds of user-based software evaluations methods: user-administered questionnaires, users' observation, and empirical usability testing.

According to Dumas, empirical usability testing started during the early 1980s, a period in which computer software started to reach a user group beyond that of 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 as a method when the evaluation has to take place in situations where little control can be achieved, as in the case of applications that are better evaluated in their use environment.

Finally, user-administrated questionnaires can be used, according to Dumas, as a stand-alone measure of usability or along with any of the other methods described before. Dumas points out that in the recent history of user-administrated

questionnaires, there have been two objectives regarding usability measurement: (a) development of short questionnaires that can be used to measure users' subjective evaluation of a product, and (b) creation of a questionnaire to provide an absolute measurement of the subjective usability of a product.
Having available an absolute measuring system for subjective usability of a product is a particularly important aspiration in comparative usability evaluation. The assessment of two aspects is crucial in achieving this purpose: the validity and the reliability of such questionnaires.
During the last twenty years, researchers in the area of usability measurement have proposed different questionnaires intended to measure subjective impressions of the performance of applications. Among them, the most known examples of this type of questionnaires are:

1. Software Usability Scale (SUS) (Brooke, 2004)
2. Computer User Satisfaction Inventory (CUSI) (Kirakiwski & Corbett, 1988)
3. Questionnaire for User Interaction Satisfaction (QUIS) (Chin et al., 1988).

These questionnaires have been designed to collect relevant data regarding user experience with applications, via different 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 (*Background notes on the SUMI questionnaire*). MUSiC was an European project to develop a set of metrics-based methods that can be used individually or together to both specify formal requirements for the usability of a product, and to assess whether a product meets those requirements (*Metrics for Usability Standards in Computing*). The project defined usability in terms of the quality of use of a product, and developed tools and procedures for measuring that quality.
The purpose of SUMI in the context of MUSiC was to develop questionnaire methods for assessing usability. The objectives of this work package were:

1. to examine the CUSI[2] Competence scale, and to expand it and/or extract further subscales if warranted by the evidence;
2. to achieve an international standardization database for the new questionnaire and to validate its use in commercial environments.

Both these objectives were achieved according to the participant researchers by the end of the project.

[1] Metrics for Usability Standards in Computing (CEC ESPRIT project number 5429)
[2] Computer User Satisfaction Inventory

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. Affect: refers to whether the user feels mentally stimulated and pleasant as a result of interacting with the software;
3. Helpfulness: refers to 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: refers to 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 researchers to discriminate the dimensions of the software that are significantly better or worse than the global software standard. These dimensions are the same five dimensions listed earlier. 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, the developers of SUMI 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 the state of the art. Based on the items and their corresponding dimensions, it is possible to report the points emphasized by the participants in their evaluations.

Unlike the other questionnaires previously mentioned, SUMI was designed to enable researchers to evaluate software systems based on an **absolute benchmark** for measuring perceived usability in applications, and not comparatively.

It is precisely because of this reason that we decided to use SUMI as the questionnaire for measuring the perceived variations in usability for our study, hoping to find in SUMI the absolute benchmarking we wanted for comparing the different subjective evaluations.

METHODOLOGY

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

As part of the research design, we measured changes in both ease of use and performance between applications' versions. We defined from the beginning of the study that in order to accomplish this goal, the study was to use both qualitative and quantitative methods for both collecting and processing the data. We observed a

posteriori that this multi methods approach provided us with a broader understanding of the context and relevant dimensions, while operating as a mechanism for triangulating the different findings.
The research design involved three different versions of the application, and was developed in two stages. Each stage involved the participation of two groups (Table 1). Stage 1, which produced the measurements for comparing Versions 1 and 2 of the application, took place during the Spring of 2003. Stage 2, involving a comparison between Versions 1 and 3, took place during Spring 2004.

Table 1: Research design showing the two Stages, the four users' groups, and the combination of versions evaluated.

	Groups	Versions
Stage 1	Group 1	V1 vs. V2
	Group 2	V1 vs. V2
Stage 2	Group 3	V1 vs. V3
	Group 4	V1 vs. V3

Participants

As mentioned before, each of the studies was conducted using two groups of subjects, all of them already users of the application, usually trained on Version 1. The reason for selecting current users of the application was to be able to simulate an upgrade scenario to both Versions 2 and 3.
According to Nielsen (2000), when testing usability a number of users equal or bigger than 15 can account for 100% of the usability issues.[3] A similar consideration is presented in SUMI's "Background notes" (*Background notes on the SUMI questionnaire*), which state that the minimum user sample size needed for an analysis with tolerable precision using SUMI is in the order of 10 to 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.

In our research we worked with a total of 54 users. For Stage 1, the first group included 20 users, the second 14. The sample was composed of architects (47%), civil engineers (18%), mechanical engineers (14%), and other professionals (21% – electrical engineers, interior designers, landscape designers, and piping engineers); all typical professions the application has been designed for.
For this Stage, subjects were recruited by UC Berkeley 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

[3] Nielsen and Landauer (1993) presented the results supporting this issue at the ACM INTERCHI'93 Conference. Based on empirical data, they showed that the number of usability problems found in a usability test with n users is equal to N(1-(1-L)n), in which N stands for the total number of usability problems in the design and L is the proportion of usability problems discovered while testing a single user. The typical value of L found by Nielsen and Landauer is 31%.

company provided each participant with a free copy of Version 2 of the application (or other equivalent application produced by the company) as compensation for their participation in the study.
The level of expertise among the users varied. A screening questionnaire that was consistent with a post-factual evaluation of user performance, revealed three user-types: experts (15%), average users (73%), and novices (12%).
For Stage 2 of the study, there were also two subjects groups. The groups included a total of 19 subjects (9 in Group 1 and 10 in Group 2). The sample was composed of building design, construction, management professionals (61%), as well as individuals from various other fields including entertainment systems design, entertainment/museum exhibit design, mapping, civil, and infrastructure management, manufacturing design and process management, and utility engineering (totalizing 39%).
Also in this stage, 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.
Stage 1 of the study took place at the company's headquarters in San Francisco's Bay Area, Stage 2 in a special facility in Los Angeles, California. 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 Stage 1 and desktop computers for Stage 2. 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.

Research design

In order to measure productivity gains, we used a series of tasks designed to simulate the production-drafting environment that exists in many of the day-to-day business operations in professional practice. The study evaluated several of the newly developed features in each release of the application.
Each of these features was incorporated into either an individual or a team-based exercise, depending on the type of performance to be evaluated. Each 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 used for solving the given tasks.
Both studies were divided into three separate phases, in which users performed various tasks. The goal of the tasks was to compare the performance and ease of use of the new features in both Version 1 vs. 2, and Version 1 vs. 3.
In Stage 1, the tasks used for each group and software version, stressed aspects of the application related to changes in the Graphical User Interface (GUI) directed at optimizing procedures involving routine actions for creating entities in the files. In Stage 2, the tasks emphasized aspects related to the creation and manipulation of files via a new feature of the application that allows configuring file characteristics automatically, as well as linking and indexing files.
The main source of quantitative data for this research came from timing the performance on solving the tasks during the different phases. These data include two primary metrics. The first is a global performance measure that compares the amount

of time required to complete each task. The second measure examines the amount of time it took each user to perform a specific activity towards the completion of the overall task. Comparing the outcomes of the two phases provided a mechanism for evaluating the performance changes between versions. These data are not relevant for this paper.

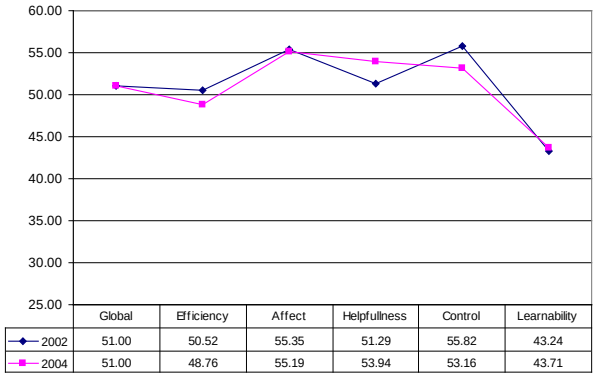
After ending each set of tasks, subjects completed a questionnaire designed to collect both qualitative and quantitative measures of their perception regarding satisfaction and experience with the application, as well as regarding the global performance of the application. In addition, subjects were 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 are further explored in this paper.

RESULTS

When comparing the data obtained using SUMI in Stages 1 and 2 for both groups, analysis shows differences for the evaluations of Version 1, both globally and for each of the five dimensions,

We found, however, closer resemblances in the means for each of the six SUMISCO dimensions when pairing results by group using Version 1 with their correspondent comparison's version (Version 2 and 3) for both Stages 1 and 2.

In the comparison between Version 1 and 2 (Fig. 1), we can see small variations in the resulting means for each of the dimensions evaluated. These variations, when analyzed using an Analysis of Variance, do not show any statistical significance (Table 2).



	Global	Efficiency	Affect	Helpfullness	Control	Learnability
2002	51.00	50.52	55.35	51.29	55.82	43.24
2004	51.00	48.76	55.19	53.94	53.16	43.71

Figure 1: Means for the values obtained with SUMI for Stage 1 (V1 vs. V2)

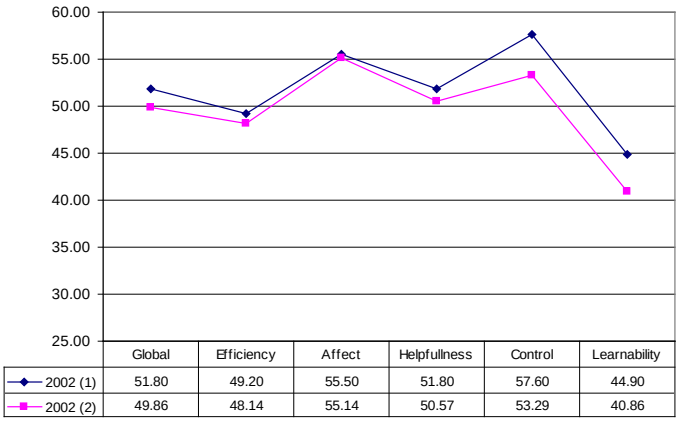
Table 2: ANOVA for the different dimensions measured using SUMI for V1 and V2 (Stage 1, both Groups).

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	6.044	1	6.044	.087	.769
	Within Groups	3187.935	46	69.303		
	Total	3193.979	47			
EFF	Between Groups	125.253	1	125.253	1.433	.237
	Within Groups	4020.414	46	87.400		
	Total	4145.667	47			
AFFECT	Between Groups	.279	1	.279	.004	.948
	Within Groups	3018.721	46	65.624		
	Total	3019.000	47			
HELP	Between Groups	76.600	1	76.600	.857	.359
	Within Groups	4109.400	46	89.335		
	Total	4186.000	47			
CONT	Between Groups	77.815	1	77.815	1.123	.295
	Within Groups	3188.664	46	69.319		
	Total	3266.479	47			
LEARN	Between Groups	2.471	1	2.471	.022	.883
	Within Groups	5187.446	46	112.771		
	Total	5189.917	47			

When we compare the two subgroups of participants who used Version 1 during Stage 1 (Fig. 2), we cannot find any statistically significant differences either (Table 3). This result is expected, considering that both subgroups are evaluating Version 1 under similar testing conditions.

When SUMI is used to compare Version 1 and 3 (Fig. 3), we can see also differences between the mean values obtained for each version, for each of the dimensions. However, only Learnability shows a means difference that is statistically significant ($F[1,37]=14.054$, $p<0.01$).

As in the previous case, there are no statistically significant differences between the means for the different dimensions when comparing the two subgroups exposed to Version 1 (Fig.4 and Table 5).

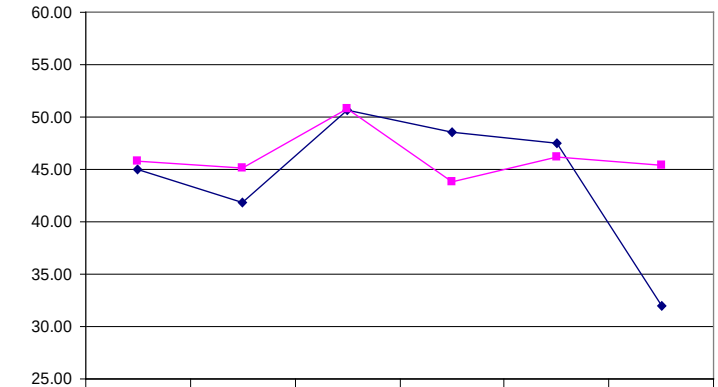


	Global	Efficiency	Affect	Helpfullness	Control	Learnability
2002 (1)	51.80	49.20	55.50	51.80	57.60	44.90
2002 (2)	49.86	48.14	55.14	50.57	53.29	40.86

Figure 2: Means for the values obtained with SUMI for the evaluation of V1 by both Groups during Stage 1.

Table 3: ANOVA for the different dimensions measured using SUMI for the evaluation of V1 during Stage 1.

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	15.543	1	15.543	.216	.649
	Within Groups	1078.457	15	71.897		
	Total	1094.000	16			
EFF	Between Groups	4.602	1	4.602	.051	.824
	Within Groups	1354.457	15	90.297		
	Total	1359.059	16			
AFFECT	Between Groups	.525	1	.525	.008	.931
	Within Groups	1025.357	15	68.357		
	Total	1025.882	16			
HELP	Between Groups	6.215	1	6.215	.044	.837
	Within Groups	2137.314	15	142.488		
	Total	2143.529	16			
CONT	Between Groups	76.642	1	76.642	1.567	.230
	Within Groups	733.829	15	48.922		
	Total	810.471	16			
LEARN	Between Groups	67.302	1	67.302	.674	.425
	Within Groups	1497.757	15	99.850		
	Total	1565.059	16			

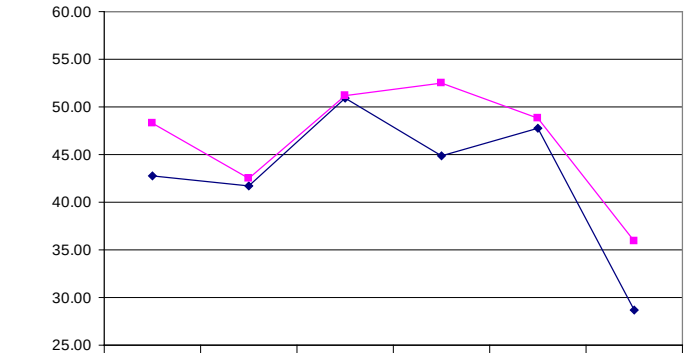


	Global	Efficiency	Affect	Helpfullness	Control	Learnability
2002	45.05	41.80	50.60	48.60	47.55	32.00
2005	45.74	45.16	50.74	43.79	46.21	45.37

Figure 3: Means for the values obtained with SUMI for both V1 and V5 (Stage 2, both groups).

Table 4: ANOVA for the different dimensions measured using SUMI for V1 and V3 (Groups 3 and 4 together).

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	4.597	1	4.597	.060	.808
	Within Groups	2832.634	37	76.558		
	Total	2837.231	38			
EFFICIEN	Between Groups	109.863	1	109.863	1.401	.244
	Within Groups	2901.726	37	78.425		
	Total	3011.590	38			
AFFECT	Between Groups	.182	1	.182	.002	.967
	Within Groups	3800.484	37	102.716		
	Total	3800.667	38			
HELPF	Between Groups	225.478	1	225.478	2.794	.103
	Within Groups	2985.958	37	80.702		
	Total	3211.436	38			
CONTROL	Between Groups	17.482	1	17.482	.289	.594
	Within Groups	2236.108	37	60.435		
	Total	2253.590	38			
LEARNA	Between Groups	1741.323	1	1741.323	14.054	.001
	Within Groups	4584.421	37	123.903		
	Total	6325.744	38			

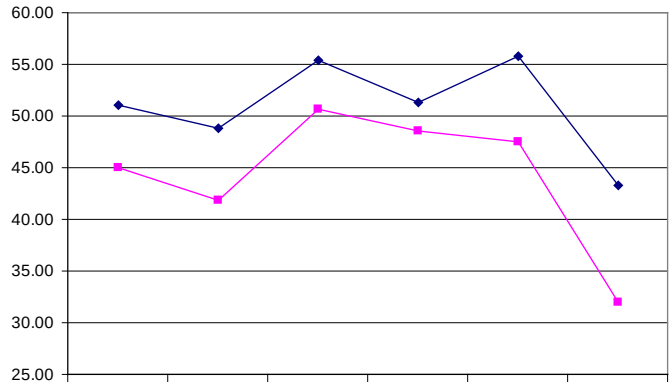


	Global	Efficiency	Affect	Helpfullness	Control	Learnability
2002 (1)	42.70	41.70	50.90	44.90	47.80	28.70
2002 (2)	48.33	42.44	51.22	52.44	48.78	35.89

Figure 4: Means for the values obtained with SUMI for both groups using of V1 during Stage 2.

Table 5: ANOVA for the different dimensions measured using SUMI for V1 comparing Groups 3 and 4 during Stage 2.

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	150.321	1	150.321	2.173	.159
	Within Groups	1176.100	17	69.182		
	Total	1326.421	18			
EFICIEN	Between Groups	2.625	1	2.625	.027	.871
	Within Groups	1630.322	17	95.901		
	Total	1632.947	18			
AFFECT	Between Groups	.492	1	.492	.006	.939
	Within Groups	1408.456	17	82.850		
	Total	1408.947	18			
HELPFU	Between Groups	269.615	1	269.615	3.894	.065
	Within Groups	1177.122	17	69.242		
	Total	1446.737	18			
CONTROL	Between Groups	4.529	1	4.529	.075	.788
	Within Groups	1029.156	17	60.539		
	Total	1033.684	18			
LEARNAB	Between Groups	244.801	1	244.801	2.248	.152
	Within Groups	1850.989	17	108.882		
	Total	2095.789	18			

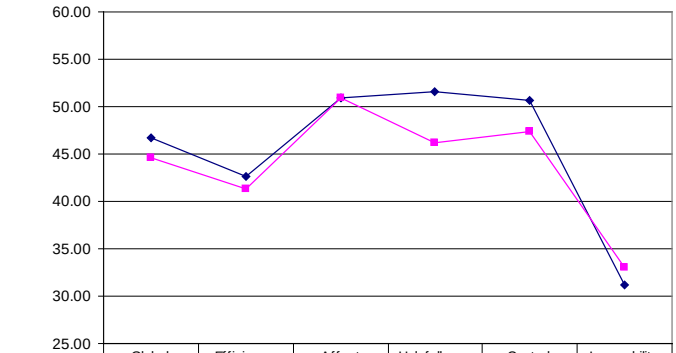


	Global	Efficiency	Affect	Helpfullness	Control	Learnability
2002 (2004)	51.00	48.76	55.35	51.29	55.82	43.24
2002 (2005)	45.05	41.80	50.60	48.60	47.55	32.00

Figure 5: Means for the values obtained with SUMI for all groups of V1 for Stages 1 and 2, grouped by Stage.

Table 6: ANOVA for both V1 groups for Stages 1 and 2, grouped by Stage.

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	325.320	1	325.320	4.631	.038
	Within Groups	2458.950	35	70.256		
	Total	2784.270	36			
EFFICIEN	Between Groups	445.741	1	445.741	5.172	.029
	Within Groups	3016.259	35	86.179		
	Total	3462.000	36			
AFFECT	Between Groups	207.588	1	207.588	2.892	.098
	Within Groups	2512.682	35	71.791		
	Total	2720.270	36			
HELPF	Between Groups	66.698	1	66.698	.649	.426
	Within Groups	3596.329	35	102.752		
	Total	3663.027	36			
CONTROL	Between Groups	629.012	1	629.012	10.806	.002
	Within Groups	2037.421	35	58.212		
	Total	2666.432	36			
LEARNA	Between Groups	1159.968	1	1159.968	11.077	.002
	Within Groups	3665.059	35	104.716		
	Total	4825.027	36			

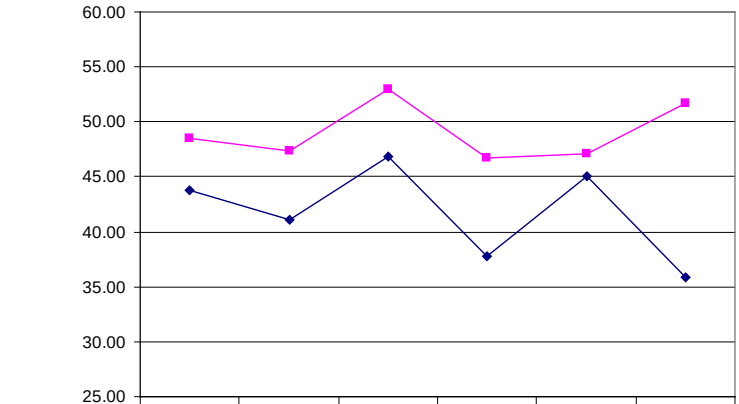


	Global	Efficiency	Affect	Helpfullness	Control	Learnability
Average	46.71	42.57	50.86	51.57	50.71	31.14
Expert	44.64	41.27	50.91	46.18	47.36	33.00

Figure 6: Means for the values obtained with SUMI for V1 for Stage 2, grouped by expertise (Average and Expert).

Table 7: ANOVA for the values obtained with SUMI for V1 for Stage 2, grouped by expertise (Average and Expert).

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	18.470	1	18.470	.226	.641
	Within Groups	1305.974	16	81.623		
	Total	1324.444	17			
Efficiency	Between Groups	7.215	1	7.215	.072	.792
	Within Groups	1599.896	16	99.994		
	Total	1607.111	17			
AFFECT	Between Groups	1.154E-02	1	1.154E-02	.000	.991
	Within Groups	1399.766	16	87.485		
	Total	1399.778	17			
Helpfulness	Between Groups	124.260	1	124.260	1.518	.236
	Within Groups	1309.351	16	81.834		
	Total	1433.611	17			
CONTROL	Between Groups	48.026	1	48.026	.826	.377
	Within Groups	929.974	16	58.123		
	Total	978.000	17			
Learnability	Between Groups	14.754	1	14.754	.114	.740
	Within Groups	2070.857	16	129.429		
	Total	2085.611	17			



	Global	Efficiency	Affect	Helpfullness	Control	Learnability
Average V5	43.83	41.14	46.86	37.71	45.00	35.86
Expert V5	48.45	47.36	53.00	46.73	47.09	51.73

Figure 7: Means for the values obtained with SUMI for V3 for Stage 2, grouped by expertise (Average and Expert).

Table 8: ANOVA for the values obtained with SUMI for V3 for Stage 2, grouped by expertise (Average and Expert).

		Sum of Squares	df	Mean Square	F	Sig.
GLOBAL	Between Groups	118.008	3	39.336	.545	.655
	Within Groups	2235.535	31	72.114		
	Total	2353.543	34			
Efficiency	Between Groups	221.825	3	73.942	.931	.437
	Within Groups	2461.775	31	79.412		
	Total	2683.600	34			
AFFECT	Between Groups	50.786	3	16.929	.163	.920
	Within Groups	3215.100	31	103.713		
	Total	3265.886	34			
Helpfulness	Between Groups	374.377	3	124.792	1.957	.141
	Within Groups	1976.366	31	63.754		
	Total	2350.743	34			
CONTROL	Between Groups	70.660	3	23.553	.417	.742
	Within Groups	1750.883	31	56.480		
	Total	1821.543	34			
Learnability	Between Groups	2602.371	3	867.457	8.579	.000
	Within Groups	3134.372	31	101.109		
	Total	5736.743	34			

A major difference however, can be observed when we compare the results obtained for all the groups that used Version 1, both for the Stages 1 and 2 (Fig. 5). In this comparison, the means for the SUMI value show statistically significant differences (Table 6), in the cases of Global evaluation ($F[1,35]=4.631$, $p<0.05$), Efficiency ($F[1,35]=5.172$, $p<0.05$), Control ($F[1,35]=10.806$, $p<0.01$), and Learnability ($F[1,35]=11.077$, $p<0.01$).

When we analyze for Stage 2 the data of the participants using Version 1 by expertise, we find that the ANOVA for the means obtained by Average and Expert users for each of the dimensions show no statistically significant differences (Fig. 6 and Table 7).

When we analyze for Stage 2 the dataset for the participants using Version 3 by expertise, the ANOVA for the means obtained by Average and Expert users show no statistically significant differences for any dimension except Learnability ($F[3,31]=8.579$, $p<0.01$, in Fig. 7 and Table 8).

In summary, this comparison shows that there are no statistically significant differences in 91% of the data points evaluated when we group them by expertise. Being that the case, there seems little reason to think that the different ratios of experts, average and novices in the different groups could account for the differences between evaluations in Stages 1 and 2 of Version 1 of the application.
By consequence, the only factor that could account for the differences obtained in the evaluations of Version 1 in Stages 1 and 2, is the variation in the conditions under which the measurements took place, that is the different nature of the tasks the two user groups were asked to perform before each of the SUMI measurements.
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.

DISCUSSION

Considering that one of the initial reasons for selecting SUMI in our research was that this tool could provide an absolute benchmark for the measurement of the different subjective evaluations, the results presented above came to us as a surprise.
Dybkjaer and Bernsen (2000) alert evaluators that when conducting usability evaluations, one of the aspects to be aware of is that scenarios should be designed to avoid priming the users on how to interact with the system. In the case of 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 the qualities of applications, one would assume an instrument intended to set an absolute benchmarking as SUMI to provide evaluators with a closer measurement for the quality of applications that are identical, as in the case of Version 1 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 evaluations. In particular, our results suggest that the level 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 (Cavallin et al. 2005), the task he or she is performing, and most probably other aspects of the context in which the evaluation takes place. 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 have to 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.

ACKNOWLEDGEMENTS

Ann Heylighen is a Postdoctoral Research Fellow of the Fund for Scientific Research-Flanders (F.W.O.-Flanders). The authors would like to thank Autodesk Inc.

for the support provided to our research, and Stefan Boeykens for his comments on earlier versions of this paper.

REFERENCES

Background notes on the SUMI questionnaire. Retrieved September 25th, 2004, from http://www.ucc.ie/hfrg/questionnaires/sumi/sumipapp.html

Cavallin, H., Martin, W.M, & Heylighen, A. (2005). This is not a Caucus-Race, accepted for presentation at *Social Intelligence Design 2005*, March 2005, Stanford (Ca).

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). *Development of an instrument measuring user satisfaction of the human-computer interface.* Paper presented at the SIGCHI.

Dumas, J. S. (2003). User-based evaluations. In J. Jacko & A. Sears (Eds.), *The Human Computer Interaction Handbook.* New Jersey: Lawrence Earlbaum Associates.

Dybkjaer, L., & Bernsen, N. O. (2000). Usability issues in spoken dialogue systems. *Natural Language Ingineering, 6*(3/4), 243-271.

Kirakiwski, J., & Corbett, M. (1988). Measuring User Satisfaction. In D. M. Jones & R. Winder (Eds.), *People and Computers IV*. Cambridge: Cambridge University Press.

Metrics for Usability Standards in Computing. Retrieved October 3rd, 2004, from http://www.newcastle.research.ec.org/esp-syn/text/5429.html

Newell, A., & Simon, H. (1972). *Human Problem Solving*. New Jersey: Prentice Hall.

Nielsen, J. (2000). *Why you only need to test with 5 users*. Retrieved September 23, 2004, from http://www.useit.com/alertbox/

Nielsen, J., & Landauer, T. K. (1993) A mathematical model of the finding of usability problems, P*roceedings of ACM INTERCHI'93 Conference* (Amsterdam, The Netherlands, 24-29 April 1993), pp. 206-213