EPOCH conference on Open Digital Cultural HeritageSystems (2008)
David Arnold, Franco Niccolucci, Daniel Pletinckx, LucVanGool (Editors)

EVALUATING USABILITY ASSESSMENT METHODS FOR WEB BASED CULTURAL HERITAGE APPLICATIONS

Davide Bolchini[2], Nicoletta Di Blas[1], Franca Garzotto[1], Paolo Paolini[1,2], Lorenzo Cantoni[2], Elisa Rubegni[2]

[1] HOC-LAB Politecnico di Milano, IT;
[2] TEC-LAB USI, University of Lugano, CH

Abstract

In spite of the variety of existing methods relatively for usability evaluation, a surprisingly limited amount of research has investigated their "quality" from a conceptual and empirical perspective Our work investigates the concept of quality for usability evaluation methods in the specific domain of web applications for cultural heritage. We define a conceptual framework for quality measurement based on attributes such as performance, efficiency, cost effectiveness, and learnability. An empirical study has been carried on to exemplify our methodological proposal, which measured the above factors for a usability inspection method called CH-MiLE, specifically devoted to the domain of cultural heritage web sites. The study involved a sample of 42 participants who used CH-MiLE to inspect a number of cultural heritage web sites under different experimental conditions. Our work is the result of a joint effort of HOC – Lab, Politecnico di Milano, and TEC-Lab, University of Lugano, carried on within Workpackage 4 of the Project EPOCH - European Research Network on Excellence in Processing Open Cultural Heritage, Proposal/Contract n° IST-2002- 507382.

Categories and Subject Description: H5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/Methodology; H5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia;

1. Introduction

In spite of the large variety of existing usability evaluation methods [BGW02] [WBH88] [Nie99] [NiM94] [Roc02], both for interactive systems in general, and for web applications in particular, relatively few studies exist that attempt to validate them, and to explore the factors that define their *quality*. Understanding the factors that contribute to the quality of a usability evaluation method, defining proper quality measurement procedures, and providing some empirical measures, not only represent a research challenge, but also pave the ground towards the *acceptability* and *adoption* of the method in real-life settings.

Our work, carried on as a joint effort of HOC – Lab, Politecnico di Milano, and TEC-Lab, University of Lugano, within Workpackage 4 of the EPOCH Project, has investigated the concept of *quality* for usability evaluation methods in the specific domain of *web applications for cultural heritage*. In this paper, we discuss some results of our work that can be useful to cultural heritage professionals and may help them to choose the proper method to adopt when evaluating their web sites for usability. In the rest of the paper, we propose a *conceptual framework* that define a set of quality attributes for a usability evaluation methods in general (section 2). Then we provide a brief overview of CH-MiLE (section 2), a web usability evaluation method that has been defined specifically for the Cultural Heritage domain. In section 3, we discuss an empirical study that has measured the quality of CH-MiLE against the quality attributes proposed in our framework. The study involved a sample of *42* participants who used CH-MiLE to inspect a number of museum web sites under different experimental conditions. Section 5 draws the main conclusions.

2. A Quality Framework for Usability Evaluation Methods

Quality is a very broad and subjective concept, oftentimes defined in terms of "fitness to requirements" [Fen91], and should to be decomposed into lower level factors in order to be measured. For usability evaluation methods, a possible criterion to identify such factors is to consider the *requirements of usability practitioners* and to focus on the attributes that may contribute to *acceptance* and *adoption* of a method in the practitioners' world [GaP07]. Our experience in academic teaching and industrial training and consulting heuristically indicates that "practitioners" want to become able to use a method after an "acceptable" time (1-3 person-days) of "study"; they want to detect the largest amount of

usability "problems" with the minimum effort, producing a first set of results in few hours, and a complete analysis in few days. We operationalize such requirements in terms of the following factors: performance, efficiency, cost-effectiveness, and learnability, defined as follows.

Performance: it indicates the degree at which a method supports the detection of all existing usability problems for an application. It is operationalized as the average rate of the number of different problems found by an inspector in given inspection conditions (e.g. time at disposal) against the total number of existing problems.

Efficiency: It indicates the degree at which a method supports a "fast" detection of usability problems. This attributes is operationalized as the rate of the number of different problems identified by an inspector in relation to the time spent [5], and then calculating the mean among a set of inspectors:

Cost-effectiveness . It denotes the *effort* - measured in terms of *person-hours* - needed by an *evaluator* to carry on a complete evaluation of a significantly complex web application and to produce an evaluation documentation that meets professional standards, i.e., a report that can be proficiently used by a (re)design team to fix the usability problems.

Learnability. Learnability denotes the ease of learning a method. We operazionalize it by means of the following factors:

- the *effort*, in terms of *person-hours*, needed by a *novice*, i.e., a person having no experience in usability evaluation, to become "reasonably expert" and to be able to carry on an inspection activity with a reasonable level of performance
- the novice's *perceived difficulty of learning*, i.e., of moving from "knowing nothing" to "feeling reasonably comfortable" with the method and "ready to undertake an evaluation"
- the novice's *perceived difficulty* of *applying application*, i.e., of using the method in a real case.

The following of this paper will exemplify how the above conceptual framework can be used to empirically measure the quality of a usability inspection method for cultural heritage web sites, called CH-MiLE, shortly introduced in the next section.

3. CH-MiLE in a Nutshell

CH-MiLE is the "customization" for Cultural Heritage Web Applications of a general methodology called MiLE (Milano Lugano Evaluation Method) defined by a joint research group at Politecnico di Milano (I) and University of Lugano (CH). MiLe [BTS03] [TBB*04] is in turn the evolution of a previous inspection technique for the usability of hypermedia and web applications named SUE [MCG*02], [DCM*03] which integrates in a novel framework concepts from various "general" usability evaluation methods (heuristic evaluation, scenario driven evaluation, cognitive walkthrough, task based testing) and some solutions for usability problems detection. The main purpose of MiLE is to prove a conceptual framework to support a *systematic* and *structured* evaluation process, especially for *novice* evaluators. A key concept of MiLE is that an interactive application can be evaluated along *two main perspectives* (see figure 1): from a "technical", "neutral", "application independent" perspective, and from a "user experience", "application dependent" perspective.

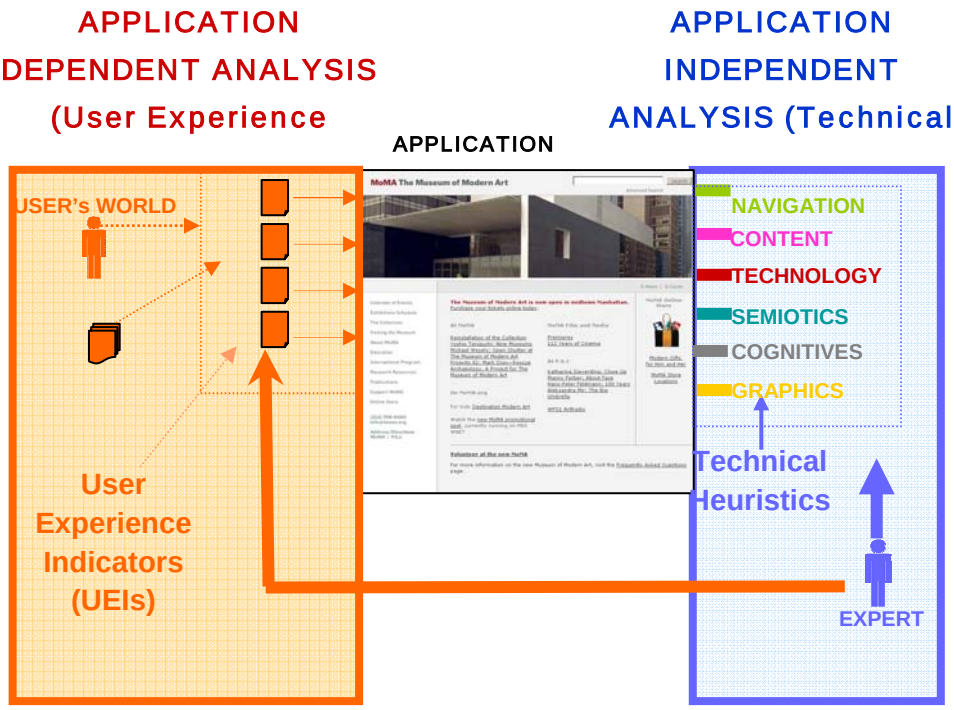


Figure 1: The MiLE Usability Method at a glance

An application independent evaluation is called *Technical Inspection* in MiLE; it considers the design aspects that are typical of the web and can be evaluated independently from the application's domain, its stakeholders, user requirements, and

contexts of use. A Technical Inspection exploits a built-in *library of* (82) *Technical Heuristics*, coupled by a set of operational *guidelines* that suggest the inspection tasks to undertake in order to measure the various heuristics. These are organized according to various *design dimensions:*

- *Navigation*: (36) heuristics addressing the website's navigational structure
- *Content*: (8) heuristics addressing the information provided by the application,
- *Technology/Performance*: (7) heuristics addressing technology-driven features of the application
- *Interface Design*: (31) heuristics that address the *semiotics* of the interface, the *graphical layout*, and the "*cognitive*" aspects (i.e., what the user understands about the application and its content or functionality)

For example, the Interface Design /Graphics heuristic "Background contrast", states a general principle of web visual design **"**The contrast between the page background and the text should promote the legibility of the textual content". The Navigation heuristic "Index Backward Navigation") claims that "When a user reaches a topic page from a list of topics ("index page"), (s)he should be able to move back to the index page without resorting on the back button of the browser".

An application dependent evaluation is called *User Experience Evaluation in MiLE*. It focuses on the aspects of the user experience that can be assessed only considering the actual domain of the application, the profiles of the intended users, the goals of the various stakeholders, or the context of use. The usability attributes that are evaluated during this activity are called *User Experience Indicators (UEIs).* MiLE provides a library of 20 UEIs, organized in three categories (see Table 2):

- *Content Experience Indicators*: 7 UEIs focusing on the quality of the *content*
- *Navigation & Cognitive Experience Indicators*: 7 UEIs focusing on the naturalness of the navigation flow and how it meets the user cognitive model
- *Operational Flow Experience Indicators:* 6 UEIS considering the naturalness of single user operations (e.g., data insert or update) and their flow

Consider for example the Content Experience UEI *Multilinguism*, which states that "the main contents of the web site should be given in the various languages of the main application targets". Obviously, there is no way to assess if *Multilinguism* is violated or not, without knowing the characteristics of the application targets. A similar argument holds for *Predictability***,** which refers to the capability of interactive elements (symbols, icons, textual links, images…) to help user anticipate the related content or the effects of an interaction [6]. Being predictable or not depends at large degree on the user familiarity with the application domain, with the specific subject of the application, and with the application general behaviour.

MiLE adopts a *scenario-based approach* [Car02] to guide User Experience Evaluation. In general terms, scenarios are "stories of use". In MiLE, they are structured in terms of a "general description", a user profile, a goal (i.e., a general objective to be achieved) and a set of tasks that are performed to achieve the goal. During User Experience Inspection, the evaluator behaves as the users of the scenarios that are relevant for the application under evaluation; he performs the tasks envisioned in these "stories", tries to image the user thoughts and reactions, and progressively scores the various UEIS on the basis of the degree of user satisfaction and fulfillment of scenarios goals and tasks. .

MiLE can be regarded as a *meta-methods* that can be specialized for each specific web site or application domain under evaluation.

CH-MiLE is an example of such specialization in the domain of Cultural Heritage. It provides an *evaluation kit* composed of: all domain independent heuristics of MiLE, a subset of domain independent User Experience Indicators that are relevant for the CH domain; a set of *scenarios* that has been defined in cooperation with museum experts and cultural heritage professionals, and capture a wide set of situation of use that are relevant for different profiles of users of Cultural Heritage web sites. An example of such a scenario is reported in figure 2.

Scenario description	A well-educated tourist knows he/she will be in town, he wants visit the real museum on December 6th 2004 and therefore he/she would like to know what special exhibitions or activities of any kind (lectures, guided tours, concerts) will take place in that day.
User profile	Tourist
Goal	Visit the Museum in a specific day
Task(s)	• Find the exhibitions occurring on December 6th 2004 in the real museum • Find information about the museum's location

Figure 2: *a CH- MiLE scenario*

4. An Empirical Study on CH-MiLE

The purpose of our empirical study was to measure the "quality" of a CH-MiLE evaluation process in terms of the factors defined in the section 2: performance, efficiency, cost-effectiveness, and learnability. The study involved *two* sub-studies – hereinafter referred as *Study 1* and *Study 2* - that focused on different quality aspects and used different procedures.

4.1 Participants

The overall study involved *42* participants, selected among the students attending two Human Computer Interaction classes of the Master Program in Computer Science Engineering at Politecnico di Milano, hold respectively in the Como Campus and in the Milano Campus. The participant profile was homogeneous in term of age and technical or methodological background. All students had some experience in web development but no prior exposure to usability. They received a classroom training on usability and CH-MiLE during the course, for approximately *5 hours* consisting of an introduction to the method, a discussion of 2 evaluation case studies, and Q&A sessions. All students were provided with the same learning material, composed of: a CH-MiLE overview report, the Library of Technical Heuristics and User Experience Indicators, CH-MiLE scenarios, including guidelines and examples, and an Online Usability Course developed by the University of Lugano (http://athena.virtualcampus.ch/webct/logonDisplay.dowebct).

4.2 Procedure of Study 1: CH-MiLE "Quick Evaluation"

The purpose of Study 1 was to measure the *efficiency* and *performance* of our method. We also wanted to test a *hypothesis on learnability:* the *effort* needed by a novice to study the method (besides the 5 hours classroom training) and to become able to carry on an inspection activity with a reasonable level of performance is *less than 15 persons/hours*.

Study 1 involved the *Como group* (*16 students*), who were asked to use CH-MiLE to evaluate a portion of an assigned museum web site (Cleveland Museum of Art website - www.clevelandart.org/index.html) and to report the discovered usability problems, working individually in the university computer lab for *three hours*. The scope of the evaluation comprised the pages from "home" to the section "Collection", which describes the museum artworks, and the whole "Collection" section, for a total of approximately 300 pages (singletons or of different types). Students did not know the assigned website in advance. Before starting the evaluation session, they received a brief explanation of the application's goals and of the general information structure of the web site, and a written specification of two relevant scenarios. Students were asked to report one "problem" (as defined in the previous section) for the same heuristic or UEI, to force them to experiment different heuristics and UEIs. They used a reporting template composed of: *Name* and *Dimension* (of the violated heuristic or UEI), *Problem Description* (maximum three lines), *url* (of a sample page where the violation occurred). The students' evaluation sessions took place one week after CH-MiLE classroom training, so that, considering the intense weekly schedule of our courses, we could assume that the students had at disposal a maximum of 15 hours to study CH-MiLE.

4.3 Procedure of Study 2: CH-MiLE Evaluation "Project"

The purpose of Study 2 was to investigate the *perceived difficulty* of *learning* and *using CH- MiLE*, and the *effort* needed to perform a *professional* evaluation. We also wanted to explore the effort needed for the different CH-MiLE activities, i.e., technical inspection, user experience inspection, scenario definition, "negotiation" of problems within a team, and production of the final documentation. Study 2 involved the *Milano group* (26 students) for a period of two months. Since we wanted to investigate an as much as possible *realistic* evaluation process using CH-MiLE, i.e., similar to the one carried on by a team of usability experts in a professional environment; participants had to evaluate an entire, significantly complex web site, to work in team (of 3-4 persons), and to deliver an evaluation report of professional quality. The subject of evaluation was freely selected by the teams within a set of assigned *museum* web sites that had comparable complexity and suffered of a comparable amount of usability problems (detected by means of a preliminary professional evaluation). To ensure an acceptable and homogeneous level of knowledge on CH-MiLE in all participants, study 2 involved only students who had successfully passed an intermediate written exam about the method. The evaluation documentation delivered by the study participants was acknowledged as a course "project" and considered for exam purposes. All teams were scored quite high (A or B), meaning that they produced a complete evaluation report of good or excellent quality.

The data collection technique for measuring the different attributes was an online *questionnaire*. It comprised closed questions about the degree of *difficulty* of studying and using CH-MiLE and about the *effort* needed to learn the method and to carry on the various evaluation activities. The questionnaire was explained to the students before they started their project and was delivered at the course exam together with the final project documentation.

4.3 Results

For lack of space, we discuss here only the main results of the two empirical studies. The analysis of the 16 problem reports produced by Como students in study 1 shows that the average number of problems was *14,8***,** with an *hourly efficiency* of *4,9* (average number of problems found in one hour). Since the total number of existing problems (discovered by a team of usability experts) is *41*, the *performance* is *36%*. If we consider the profile of the testers and the testing conditions, these results can be read positively. They confirm our hypothesis on learnability and indicated that after 6 hours of training and a maximum of 15 hours of study, a novice can become able to detect more than one third of the existing usability problems!

Some key results of the analysis of the *questionnaire* data collected during study 2 are presented in the following figures.

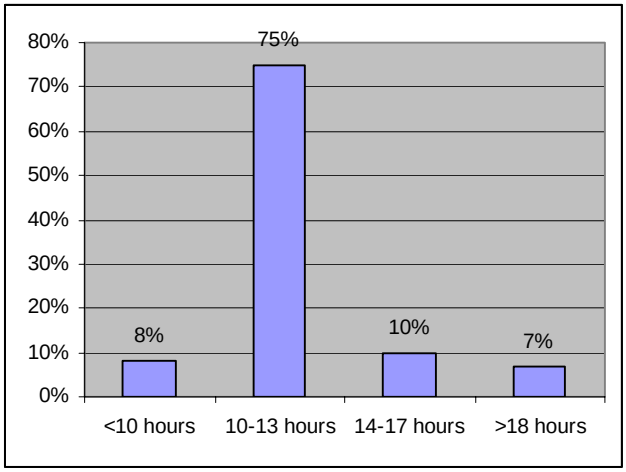


Figure 3: CH-MiLE Learning Effort

Concerning the *learning effort*, participants invested in the preliminary study of CH-MiLE an average amount of time of *10-13 hours* (see Figure 3), which is comparable with the estimated effort of Como students. Concerning *learning difficulty*, a large majority of participants (*73%*) found CH-MiLE study activity *rather simple*- see Figure 4.

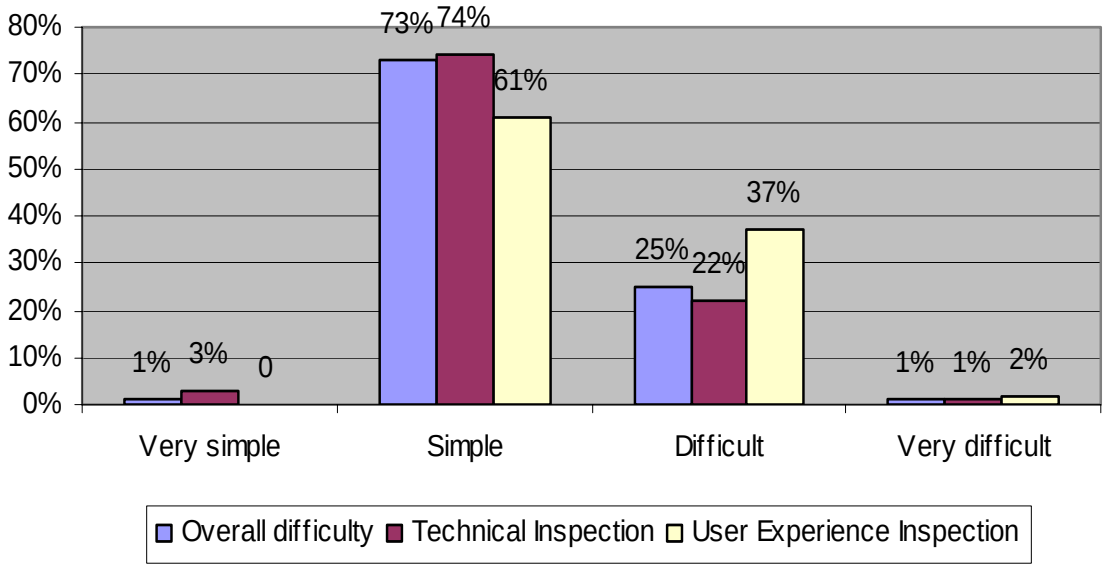


Figure 4: Perceived Difficulty of Learning CH-MiLE

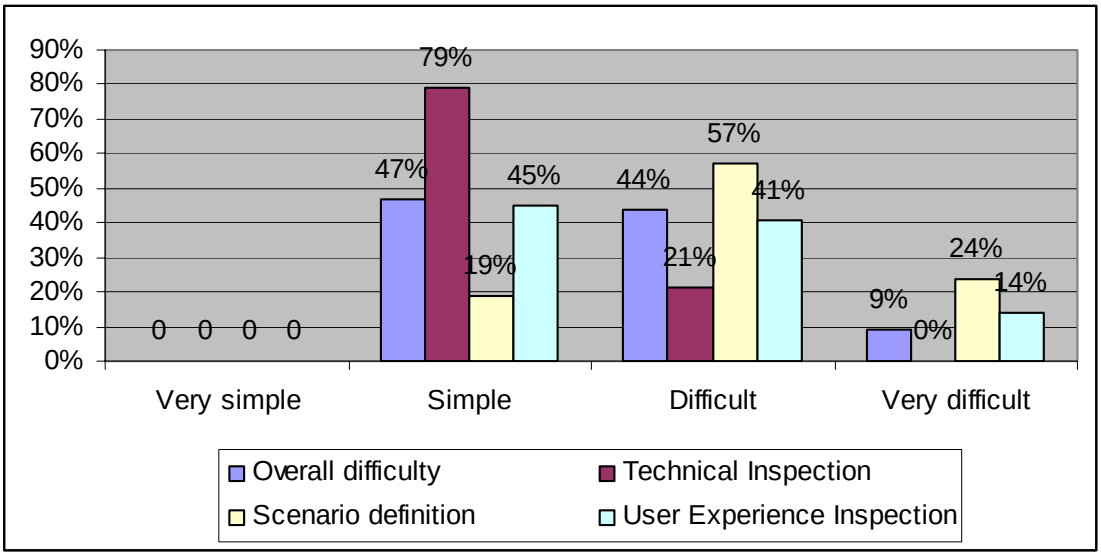


Figure 5: **Perceived Difficulty of Using CH-MiLE**

Figure 5 highlights that students perceived the *use* of CH-MiLE in a real project as more complex than studying it. Only *47%* of the students scored "*simple*" the use of CH-MiLE , while *53%* judged it *difficult* or *very difficult*.

Figure 5 also shows that the User Experience Evaluation was perceived slightly more difficult than it was expected from the study (compare Figure 5 with Figure 4). These data may indicate a weakness of the CH-MiLE method: although the number of User Experience Indicators is smaller than the number of Technical Heuristics, the definition of the former is more vague and confused, and their measurement may result more difficult for a novice. Another reason for the difficulty of performing User Experience Inspection might be related to the difficulty of defining "good" *scenarios*. Figure 5 pinpoints that a significant amount of participants (*81%)* estimated this activity *difficult* or *very difficult*. Indeed, if the concept of scenario is simple and intuitive, defining appropriate scenarios requires the capability - that a novice oftentimes does not possess - of eliciting requirements and reflecting on users profiles and application goals.

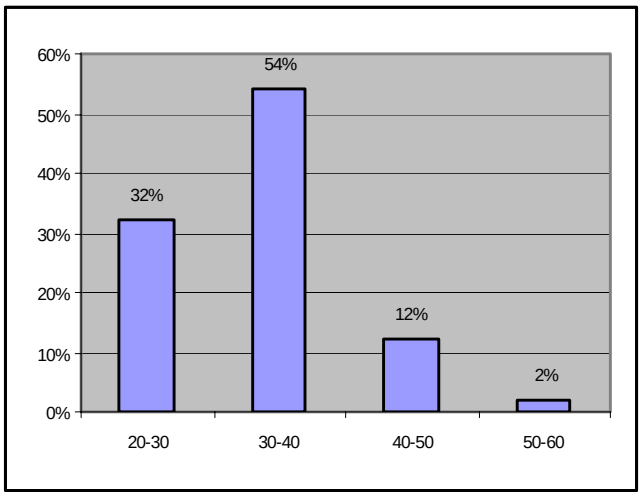


Figure 6: Individual Effort for a Professional Evaluation Process (in person/hours)

Concerning *cost-effectiveness*, Figures 6 and 7 highlight the *average effort* to perform a professional evaluation process of an entire application, and the effort allocation on the various activities. The effort is calculated in person/hours, by each single evaluator, considering the time spent working both individually and in team.

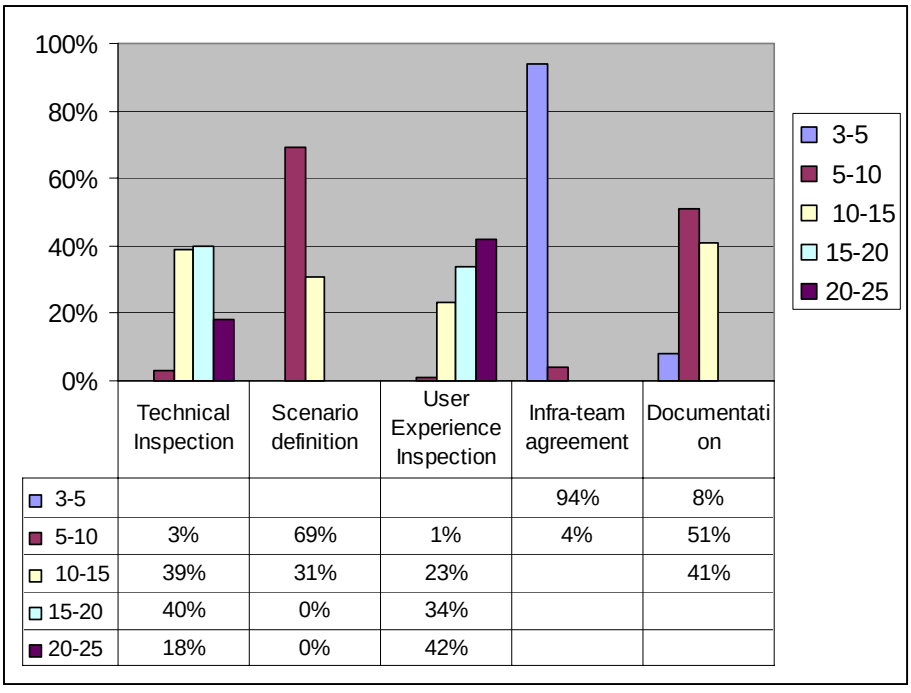
Some interesting aspects emerge from the data on cost effectiveness:

- *54%* of the participants invested from *30 to 40 hours* in the *overall* evaluation process; this means that a team of 2-3 evaluators can deliver a professional report of a significantly complex web application in one week at a total cost of 0.5-0.75 person/month, which is a reasonable timing and economic scale in a real life organizational context

- consistently with the results in Figure 5, the activity of *scenarios* definition is an effort demanding task: *69%* of the participants invested *5-10 hours* in this work

- *5-10 hours* is also the effort invested by 41% of the students in *reporting*; if we consider that all team declared that the reporting work was shared among team members, we can estimate as approximately 1,5-1 person-week the global team effort for the reporting task

- the "negotiation activity" (i.e., getting a team agreement about the final results to be reported) resulted quite fast (3-5 hours for 94% of the persons), which suggest that CH-MiLE supports the standardization of the inspection process and the homogenization of results



	Technical Inspection	Scenario definition	User Experience Inspection	Infra-team agreement	Documentation
3-5				94%	8%
5-10	3%	69%	1%	4%	51%
10-15	39%	31%	23%		41%
15-20	40%	0%	34%		
20-25	18%	0%	42%		

Figure 7: Individual Effort distribution per Task in a Professional Evaluation Process

In summary, the analysis of the experimental results has proved that CH-MiLE meets the need of "practitioners" stated in section 3. Our empirical has proved that the learnability of the method is good, since after a short training (5 hours), understanding CH-MiLE basics requires an acceptable workload of study (10-15 hours). The method has also proved to support inexperienced inspectors in performing an efficient and effective inspection both in the context of a short term, quick evaluation (3 hours) and in the context of a real project. Still, our study has also shown that shifting the inspection scope from a (relatively) small-size web site to a full-scale complex application, requires higher levels skills and competence (e.g., for scenario definition) that go beyond usability

know how in a strict sense, and can only be gained through experience.

5 Conclusions

Quality is a very broad and generic term, especially if applied to methodological products, and can be defined along many different perspectives. Our work is only a first step towards the definition of a quality assessment framework for web usability evaluation methods, in particular to address the evaluation requirements of cultural heritage institutions. In this paper, we have suggested that *learnability*, *performance*, *efficiency*, and *cost effectiveness* are possible measurable attributes for methodological quality of web usability evaluation techniques that should be considered as important element to consider by cultural organizations and professionals when considering the conceptual tool to use for the usability evaluation of their web sites. We have discussed how the above factors can be measured, presenting an empirical study that evaluated the quality of a specific a method - CH-MiLE - explicitly conceived for usability inspection of cultural heritage web sites.

Acknowledgments

This work has been funded by Project EPOCH - European Research Network on Excellence in Processing Open Cultural Heritage, Proposal/Contract n° IST-2002- 507382. The authors are grateful to Maria Pia Guermandi – IBC Emilia Romagna (I), for her contribution in the definition of CH-MiLE. Special thanks go to the participants to the empirical study, and to Luca Triacca for his collaboration during study 1.

References

[BTS03] Bolchini D., Triacca L., Speroni M. MiLE: a Reuse-oriented Usability Evaluation Method for the Web, Proc. HCI International Conference 2003, Crete, Greece, 2003.

[BGW02] Brinck, T., Gergle, D., Wood, S.D., Usability for the web, Morgan Kaufmann, 2002

[Car02] Carroll, J., Making Use – Scenario-based design of Human-Computer Interactions, MIT Press, 2002

[DCM*03]De Angeli A., M.F. Costabile, M. Matera, F. Garzotto, P. Paolini. On the advantages of a Systematic Inspection for Evaluating Hypermedia Usability. In International Journal of Human Computer Interaction, Erlbaoum Publ. Vol. 15 (3), June 2003, pp. 315-336.

[Fen91] Fenton, N. E. (1991) *Software Metrics: A Rigorous Approach*, 2nd edn. Chapman & Hall, 2002

[GaP07] Garzotto F., Perrone V.: Industial Acceptability of Web Design Methods: an Empirical Study**,** Journal of Web Engineering, Vol. 6, No.1 (2007) pp. 073-096

[MCG*02] Matera, M., Costable M.F., Garzotto F., Paolini P., SUE Inspection: An Effective Method for Systematic Usability Evaluation of Hypermedia, IEEE Transactions on Systems, Men, and Cybernetics, Vol.32, No. 1, January 2002

[Nie99] Nielsen, J., Designing Web Usability, New Riders, 1999

[NiM94] Nielsen, J., Mack, R., Usability Inspection Methods, Wiley 1994

[Roc02] Rosson, M.B., Carroll, J., Usability Engineering, Morgan Kaufmann, 2002

[TBB*04] Triacca L, Bolchini D., Botturi L., Inversini A, MiLE: Systematic Usability Evaluation for E-learning Web Applications. ED Media 04, Lugano, Switzerland

[WBH88] Whiteside J., Bennet J., and Holtzblatt K., Usability engineering: Our experience and evolution, in Handbook of Human-Computer Interaction, M.Helander, Ed. Amsterdam, The Netherlands, North-Holland, 1988, pp.791-817