Journal of Web Engineering, Vol. 7, No.4 (2008) 299-317

QUALITY AND POTENTIAL FOR ADOPTION OF USABILITY EVALUATION METHODS: AN EMPIRICAL STUDY ON MILE+

DAVIDE BOLCHINI

University of Lugano, Switzerland
bolchind@lu.unisi.ch

FRANCA GARZOTTO

Politecnico di Milano, Italy
garzotto@elet.polimi.it

Received February 15, 2008
Revised June 20, 2008

Web usability evaluation methods are conceptual tools which should enable web designers, web engineers and usability engineers to detect and possibly anticipate usability problems of a web application, and eventually to provide requirements for improving the quality of the user experience. As the number of techniques and methods available grows, practitioners need clear criteria to choose which methods best fit their project needs, resources and organizational goals. Therefore, it becomes more and more important to foster research towards evaluating the quality of the usability evaluation methods, especially in view of their potential adoption among practitioners. Besides focussing on known attributes of *intrinsic* quality of the method (such as coverage, reliability and validity), this paper also explores "perceived" quality attributes related to the potential adoption of the method among practitioners, namely in terms of learnability, perceived difficulty, and cost-effectiveness. We report two empirical studies which have been carried out to measure these quality attributes on a state-of-the-art inspection method for web usability, called MiLE+. The result of this work can be useful to scholars because it provides validation examples and a set of quality attributes to apply to other usability evaluation methods; it also benefits practitioners because it offers a clear guidance about what requirements they should look for when selecting a usability evaluation method for their own project needs.

Key words: web usability, quality, empirical study, inspection, heuristics

1 Introduction

In spite of the large variety of existing usability evaluation methods, both for user interfaces in general, and for web applications in particular [2, 6, 9, 10, 11], the factors that define their quality are seldom discussed in the literature, and relatively few empirical studies exist that attempt to measure them [17, 5, 7, 13]. Consider for example heuristic evaluation, one of the most popular inspection methods for website usability [9, 10]. It is claimed to be "simple" and "cheap", implicitly assuming that these are quality factors, but mainly informal arguments or anecdotic evidence are offered to support these claims (e.g., "few simple heuristics", "no user involvement", "no need of special equipment") with little scientifically documented empirical results on its use in practice.

Understanding the quality factors for evaluation methods, and measuring them, is a challenging research issue in web engineering. Evaluating web usability, in fact, is of paramount importance to improve the quality of the user experience of web applications. Therefore, having at hand reliable and proven usability evaluation methodologies is a vital resource to the web engineering community. Investigating the quality factors of usability evaluation methods should yield practical implications for the industrial acceptability and ultimately for the adoption of these methods in the professionals' community: the empirical evidence of the *usability of the method* itself is a key driving force for having of a methodological "product" accepted and adopted in a real business context.

This paper investigates the above issues and aims at raising a critical reflection on the concept of quality for web usability methods and on the techniques to evaluate it. We extensively report an empirical study in which we explored the quality of an inspection method for web usability called MiLE+. In this work, we decompose the general concept of quality into lower level, more measurable attributes such as learnability, performance, efficiency, coverage, and cost effectiveness and investigate them involving novice evaluators both in a controlled situation (a three hour inspection session) and in the context of a real evaluation project.

The remainder of the paper is as follows. We first provide a high-level overview of MiLE+, the usability inspection method which is evaluated in this paper (section 2). Then we detail the approach, the experimental design of two empirical evaluation studies carried out to assess the quality of MILE+, followed by the presentation and discussion of the results (section 3). Finally, we draw conclusions and directions for future work (section 4).

2 Evaluating Usability with MILE+: an overview

MiLE+ is the evolution of two previous methods that were developed at the respective authors' labs, MiLE (Milano-Lugano Evaluation) [1, 12, 14] and SUE (Systematic Usability Evaluation) [8]. MiLE+ integrates proven techniques and evaluation strategies from various "traditional" usability evaluation methods (heuristic evaluation, scenario driven evaluation, cognitive walkthrough, and task based testing), and distils our ten year experience in applying and teaching usability in educational contexts and at industrial or governmental level. On a general level, MiLE+ is more systematic and structured than other evaluation techniques, thus offering an analytical guidance to carry out the evaluation, focussing specifically on web applications, and being particularly suited for *novice evaluators*.

The basic principle informing the MILE+ evaluation framework is that a web application (and this is true also for interactive applications in general) can be evaluated along *two main orthogonal perspectives* (see figure 1):

1) *Requirements-independent perspective:* the evaluation of the usability can be tackled from a "technical" and "objective" point of view, which considers all those design aspects which are completely (or at least at a large extent) independent from the specific content, domain, goals and users of the application (e.g. information architecture, navigation, interaction mechanisms, graphics and layout).

2) *Requirements-dependent perspective*: this point of view examines the usability in terms of fulfilment of specific needs of specific users in specified contexts of use (ISO-92401). It needs therefore to be strongly informed by the sufficient domain knowledge, and by an understanding of the goals and requirements of the application.

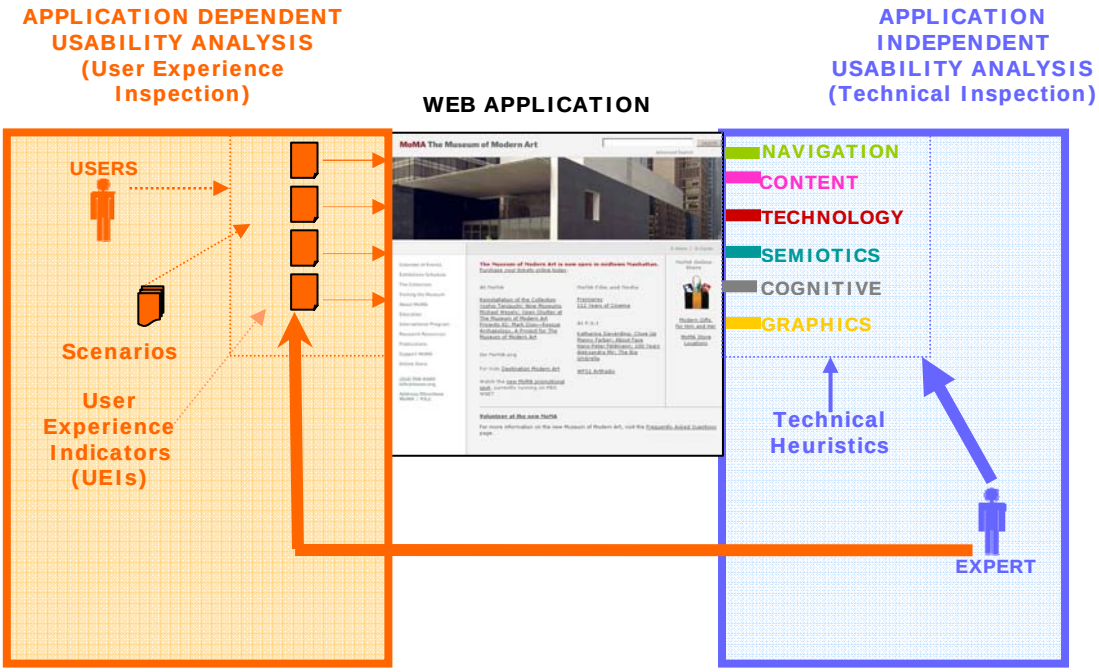


Figure 1 MILE+ evaluation framework at a glance.

From the direct experience in working with usability professionals (both experienced and less experienced) it is clear that this distinction of evaluation perspectives explains and gives reason of a common experience in doing usability studies. On the one hand, web applications which are bullet-proof in terms of navigation and graphics usability (e.g. they score high on all Nielsen's heuristics), can be deeply flawed and disastrous in terms of supporting the intended users in achieving crucial goals (e.g. because the content is not the right one or it is not properly structured *for given user's tasks*). On the other hand, exclusively focussing the effort on validating the support to a set of scenarios does not guarantee to spot general design flaws which can constitute hard obstacles to the user experience in general (outside specific scenarios). The two evaluation perspectives are thus complementary and support one another. They enable evaluators to adopt a flexible and modular strategy to cope with the multi-faceted issues of web usability.

As shown in Figure 1, according to MILE+, the requirements-independent (also called application-independent analysis) is typically carried out through a usability inspection activity (expert review or "technical" inspection). As it will be discussed, MILE+ provides the inspectors

with the necessary conceptual tools (a library of technical heuristics) to carry out this inspection proficiently.

Similarly, the inspector can carry out an expert review on the application-dependent aspects of the application, given a proper gathering of domain knowledge and application requirements, and by following specific MILE+ guidelines (essentially for building scenarios). On the basis of the results of the inspection, to complete and corroborate the application-dependent analysis, users are of course involved through traditional user testing activities. Let us now review in detail, and with a number of application examples, the key features of the evaluation conceptual tools offered by MILE+ in the context of the two evaluation perspectives discussed so far.

2.1 Application Independent Analysis

Even if it may seem against any reasonable usability principle, it is a matter of fact that a number of usability problems in a web application can be spotted and examined independently from the goals of the applications (and its stakeholders), the user requirements, and the context of use of the application. For example, one heuristics offered by MILE+ concerning the "graphics" usability states:

- **Background-Foreground contrast: "**the contrast between the background and the text or images on the foreground should allow the readability of the actual content".

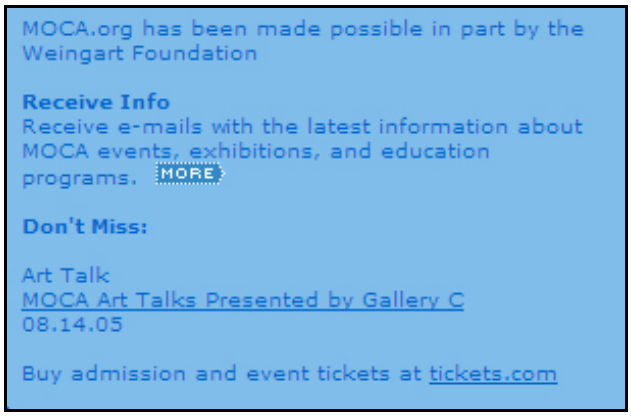


Figure 2 Example of lack of appropriate contrast.

Excerpts from MOCA website (www.moca.org) presents a lack of contrast between background and text, thus causing a low readability, and affecting usability – whatever the user and his needs are. More specific heuristics can be derived from this one, to indicate, for example, a numeric chromatic ratio between background and foreground colour. Some of the technical usability aspects of this type could be even checked by automatic tools (in case these principles gets properly quantified and standardised) and this can help improve the efficiency of this kind of evaluation.

This may be relevant for specific niche of user profiles, such as people with low vision. And this would lead into a more user-dependent kind of analysis. However, the point captured by this

heuristics is that there is a general readability principle to be respected, in order to be able to properly communicate the content to a generic user (even without low vision), independently by the specific tasks and goals to be supported.

With regards to navigation usability, MILE+ also provides the inspectors with a library of heuristics. For instance, we report here the heuristics concerning the design of "back" navigation mechanisms.

- **Backward Navigation in index/list**

  "As the user reaches a list of topics, (s)he should have the control of the navigation both from the starting index to each element and to go back from any of the elements to the index".

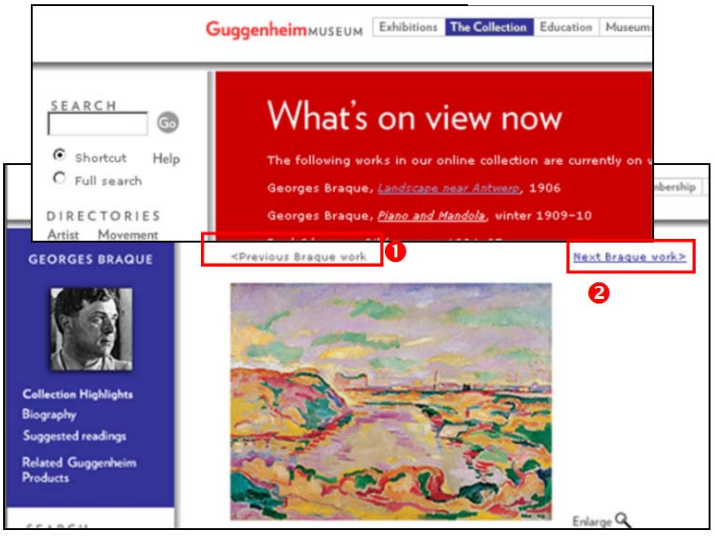


Figure 3 Example of Poor Index Navigation.

In the case of the Guggenheim Museum website (www.guggenheimcollection.org), once the user reaches the list of artworks "now on view" and selects a painting (e.g. Georges Braque - Landscape near Antwerp) he can reach the desired page correctly. When he tries to return to the list of art's works, however, the backward mechanism is absent. The only navigational mechanism are two links called "Previous Braque work" (❶) and "Next Braque Work" (❷) that allow navigating within a guided-tour of Braque's works. There is no efficient way to get back to the index of art works and select from there another painting. This navigational flaw negatively affects the usability of a generic browsing activity of the collection, independently from the specific user profile or user goal.

The activity of performing an application independent analysis is called *Technical Inspection* in MiLE+. During this analysis the evaluator examines the web application assuming the point of view of the designer and not of the end-user and focusing on systematically checking the

compliance with these "technical" heuristics. MILE+ offers a built-in library of (82) usability "technical" heuristics, coupled with a set of operational guidelines that identify the inspection tasks to undertake in order to measure the various usability aspects. These heuristics are organised according to design dimensions:

- *Navigation*: (36) heuristics addressing the usability of the information architecture and navigation mechanisms;
- *Content*: (8) heuristics addressing the general quality of the information offered to the user;
- *Technology/Performance*: (7) heuristics addressing technology-driven features of the application;
- *Interface Design*: (31) heuristics addressing the semiotics of the interface and the graphical layout.

An excerpt from the MILE+ technical heuristics library is shown in the table below.

Dimension		**Examples of Heuristics**
Navigation		Consistency of the overall navigation
		Control of a guided-tour
Content		Text accuracy
		Multimedia consistency
Technology/Performance		System reaction to errors of a user
		Operations management
Interface design		
	Cognitive	Information overload
		Scannability
	Graphics	Font size
		Text layout
	Semiotics	Ambiguity of link labels
		Conventionality of interaction images

Table 1 Excerpt from the library of MILE+Technical Heuristics.

Each MILE+ heuristics embodies a detailed inspection protocol which consists of three basic components (see example in Table 2):

a) the feature or specific portion of the application relevant to the heuristics to be applied;

b) the definition of the potential usability problem to be verified;

c) one or more inspection actions (or tasks) to be carried out by the inspector in order to verify the compliance of the application with the heuristics.

The concepts and vocabulary of MILE+ heuristics are based on common concepts of information architecture and hypermedia design. In particular, the ontology underlying MILE+ borrows from well-known and main stream usability engineering approaches and extensively from IDM (Interactive Dialogue Model) [15,16].

The reliability and strength of the heuristics rely on the fact that they have been collected and iteratively refined over the years as a crystallization of the experience of website usability experts and through a constant alignment to state-of-the-art usability guidelines and patterns.

An interesting aspect of MILE+ application-independent analysis is that it yields insights into usability problems even if the inspector lacks domain knowledge. With MILE+ at hand, the

"domain ignorance" of the inspector (especially common when coping with a new application domain) is not a weakness, but rather a strength, because, thanks to the library of heuristics, the evaluator is able to examine in depth and thoroughly all the design aspects long before spending resources and effort in user research, in understanding the content, and in studying the domain and the specific application goals.

Feature	**Navigation within a Topic**
Heuristics	**Orientation clues**
Action	1. Identify an instance of topic in the website 2. Check whether it is present a *path visibility* (where I can go?): navigate from the home page to the instance of the selected topic and verify whether the path traversed (where have I been?) is communicated. 3. Check whether it is present a *status visibility* (where I am?)*:* navigate randomly within the topic and verify whether the current location within the information architecture is communicated. 4. Check whether it is present a *context visibility:* navigate randomly within the topic and verify that the indication of the information context you are browsing is communicated at every location. 5. Repeat step 1 thru 4 for other 3 instances of the topic.

Table 2 An example of the detail protocol of a navigation heuristics.

2.2 Application Dependent Analysis

A number of aspects of a web application require the evaluator to situate himself within the different possible contexts of use and evaluating how the application actually supports the specific needs of different users to whom it is targeted.

This dimension of usability evaluation is well described by ISO 9241-11 definition of usability, which states that usability is "*the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in particular environments*".

During MILE+ "Application Dependent Analysis", the evaluator has to determine if the user is put in the right conditions for achieving his goals when using the application, answering questions such as: Do the intended target users find the information they need? Are the users effectively and efficiently supported in achieving their goals? Are people properly driven and guided to unexpected content? Is the content relevant for the user(s)? Is interaction enjoyable/entertaining?.

An example of domain dependent usability aspect is *multilinguisticity*: if the application addresses different types of users speaking difference languages, the content should be, obviously, given in more than one language, according to the main application targets. Another example of a requirements-dependent usability aspect is *Predictability***,** i.e., the capability of interactive elements (symbols, icons, textual links, buttons, images…) to clearly anticipate the related content and the effects of the interaction. Some aspects of the semiotics of the interactive elements (e.g. the clarity and comprehensibility of the links labels) are strictly related to the type of users that will use the application, and to their degree of familiarity with the application domain or the specific subject of the application (see figure 4).

Figure 4 Predictability problems in the Armani web site.

For example, in the Armani website (www.armani.com), one of the link labels is called "Armani exchange". If we consider a first time user of the website, with no previous knowledge of the Armani world, and who is willing to get an idea of the offer of the website, it may not be clear at all to this user what is the actual content behind the label. Only a user which knows in-depth Armani, its culture and jargon, can know that Armani Exchange is one of the Armani's Seasonal Collection. Therefore, from a usability point of view, this issue can be considered a usability problem if the intended users of the website are not only "Armani's fan", but also people which are just curious (they do not have the background for understanding this label).

The activity of evaluating application dependent usability aspects is called *User Experience Inspection* in MiLE+. In order to offer a systematic and structured inspection guidance to the evaluator, and to take into account coherently the abovementioned requirements-dependent dimensions, the User Experience Inspection is *scenario-driven*. Scenarios are stories about the use of an interactive application [3, 4] and are widely used and investigated for their ability in supporting various stages of the usability engineering process, from user requirements envisioning to design and evaluation.

MILE+ considers scenarios as key drivers of the user experience inspection and provides a guidance to properly define and use them during the inspection process. In MILE+, the key structural components of a scenario are a context (or scenario setting), a user profile (the description of the relevant characteristics of the potential person making use of the application), a goal (i.e., a general objective to be achieved, defined at the proper level of abstraction) and a set of tasks that are performed to achieve the goal (see figure 5). For MiLE+ User Experience Inspection, the evaluator use scenarios as the primary guide for inspection: s/he performs the tasks envisioned in the previously defined scenarios, tries to anticipate the potential user behaviour (taking into account a specific profile), and progressively comment on them. The definition of scenarios requires an elicitation activity which may involve user research as well as the interaction with different stakeholders: the client, domain experts, and end-users. During User Experience Inspection, the evaluator does not only verify whether or not the application satisfies the scenario. The inspector is guided to assess "how well" a scenario is supported by means of a set of specific attributes, called User Experience Indicators. Thanks to the User Experience Indicators, the inspector is more analytically guided to score the various aspects of the user experience while performing the scenario.

Scenario setting	Well-educated American tourist who knows he will be in town, he wants visit the real museum in three weeks and therefore he would like to know what exhibitions or activities (lectures, guided tours, concerts) will take place on that day.
User profile	Tourist
Goal	Visit the Museum on a specific day
Task(s)	• Find out exhibitions and activities occurring in the museum on March 4th • Get details about museum's location, opening hours and means of transportation.

Figure 5 Example of scenario for a museum website

User Experience Indicators are organized in three categories (see Table 3):

- *Content Experience Indicators*: they focus on the quality of the user interaction with the *content* of the application.
- *Navigation & Cognitive Experience Indicators*: they focus on the quality of the navigation flow and how it meets the cognitive model of the user(s).
- *Operational Flow Experience Indicators:* they focus on the flow of operations (e.g., insertion, update, commit operations) and on how this is natural and easy for the user.

Categories of interaction	Examples of User Experience Indicators
Content Experience	Completeness
	Relevance
	Comprehensibility
Navigation & Cognitive Experience	Predictability of interactive elements
	Learnability
	Memorability
Operational Flow Experience	Naturalness
	Engagement
	Recall

Table 3 Examples of MILE+ User Experience Indicators.

MiLE+ provides a User Experience Indicators library composed of 7 Content Experience Indicators, 7 Navigation&Cognitive Experience Indicators and 6 Operational Flow Experience Indicators, that means a total of *20 Indicators*. Now that we have briefly explained the characteristics of MILE+, we illustrate the approach, the process and the results of an empirical evaluation of the quality of the method.

3 Empirical study on MiLE+

3.1 Quality Attributes and Research Goals

The aim of our empirical study is to evaluate the "quality" of MiLE+ in terms of the following measurable factors: performance, efficiency, coverage, cost-effectiveness, and learnability. All these attributes exploit in different ways the notion of *usability problem*, which must be defined

more precisely. In a complex application, we can identify a number of "page types", representing sets of pages that have similar meaning (e.g., denote topics or functionality of the same "type") and similar lay-out and navigational structure, and a number of "singleton pages", i.e., pages that represent a topic or functionality that cannot be reduced to a class. We define a *problem* as an obstacle to the user experience resulting from a *violation* of a MiLE+ heuristic or user experience indicator, that either occurs in a singleton page or *repetitively* occurs in (at least 3) pages of the same type. For example, in a museum web site the violation of the heuristic "Guided tour control" occurring in three different pages of type "artwork introduction" counts as one problem, but a problem is also the violation of the heuristic "Ambiguity of link labels" occurring in the singleton page "Museum Presentation". Using the above definitions, in the following we precisely define all the indicators for the evaluation study.

Performance indicates the degree at which a method supports the detection of all usability problems for an application, in given inspection conditions. This attribute is also known as completeness or thoroughness, as referred to the activity of an individual inspector. In our study, we have operationalized *performance* as the average rate of the number of different problems found by an inspector (P_i) in given inspection conditions (e.g. time at disposal) against the total number of different problems (P_{tot}).

$$Performance = avrg(\frac{P_i}{P_{tot}})$$

Efficiency indicates the degree at which a method supports a "fast" detection of usability problems. This attribute is also known as *productivity* in similar contexts. For the purpose of our evaluation, *efficiency* is operationalized as the rate of the number of different problems identified by an inspector in relation to the time spent, and then calculating the mean among a set of inspectors:

$$Efficiency = avrg(\frac{P_i}{t_i})$$

where P_i is the number of problems detected by the i-th inspector, and t_i is the time (s)he spent to find the problems.

Coverage denotes the degree at which different inspectors having comparable background and performing an inspection under the same conditions (e.g., time, context, previous knowledge about the system under evaluation) identify *different* problems, and therefore are able to *collectively* identify the overall set of problems. To measure coverage, we consider, for each pair of inspectors, the *average shared problems rate*, i.e., first comparing the number of *common* usability problems found (i.e., the cardinality of the intersection of the problems discovered by the pair of inspectors) against the set of problems individually found by the two inspectors (i.e., the cardinality of the union of the problems discovered by the pair of inspectors); then calculating the mean of the results on all pairs of inspectors:

$$shared\_problems\_rate = avrg_{x<y}(\frac{|Insp_x \cap Insp_y|}{|Insp_x \cup Insp_y|})$$

The lowest is the shared problem rate, the highest is the coverage. On the contrary, the highest is the shared problems rate, the higher is the *reliability* of results, because it indicates how many problems are discovered by more than one inspectors.

Cost-effectiveness denotes the *project effort*, in terms of *person-hours*, needed by a *trained evaluator* to carry on a complete usability 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.

As to *learnability*, we define it as "ease of learn" for a *novice*, i.e., a person having no experience in usability evaluation. Learnability is operazionalized by means of two factors:

- the *learning effort* (in terms of *person-hours*) needed by a *novice* to understand the method and become "reasonably expert", i.e., to be able to carry on an inspection activity and to achieve a reasonable level of performance
- the *perceived difficulty* of both the *learning process* of MiLE+ and the *method's application*, i.e., the use of MiLE+ in practice (in a four values scale 4 = very difficult; 3 = difficult; 2 = easy; 1 = very easy)

3.2 Context, Participants and Experiment Design

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 (19 students) and Milano Campus (26 students). All students had some experience in web development but no exposure to usability.

MiLE+ was presented and discussed through examples and questions&answer sessions in the classroom, during two lessons of approximately *three hours each.* During the first lesson, all students received the following learning material:

- *MiLE+ description:* an article (8 pages) describing the methodology
- *The Libraries of Technical Heuristics and User Experience Indicators (UEIs):* a report describing in detail, all technical heuristics, divided into design dimensions (e.g. content, navigation, interface design, etc.) and all UEIs , including *guidelines* and examples for applying them
- *MiLE+ Examples of use:* two different documents with examples of evaluations
- *Hand-outs of the course:* the slides of the course were given to the inspectors;
- Access to an *Usability Online Course:* a Moodle-based online course on usability foundations and MiLE+.

The study involved two main empirical evaluations, each one focusing on different aspects of MiLE+ quality and performed by different groups of inspectors in different contexts.

3.3. Empirical Evaluation 1: MiLE+ evaluation session

The purpose of this empirical evaluation was to collect data related to *performance*, *efficiency*, *reliability* and *coverage*. We also wanted to test a *hypothesis on learnability:* the *effort* needed by

a novice after the classroom training to study and understand the method and to be able to carry on an inspection activity with a reasonable level of performance is *less than 2 full working days (around 15 persons/hours)*. This experiment involved the *Como group* (*16 students*) and was carried on in the university computer lab *one week after* the MiLE+ classroom lessons, under our supervision. Students were asked to perform a MiLE+ usability inspection of a portion of an assigned web site in a limited span of time (3 hours), and to report *different* problems discovered, according to the definition of problem given above. Considering the weekly lessons schedule, we could assume that the maximum time the students had at disposal to *study* MiLE+ after the classroom lectures was 10-12 hours.

The subject of the evaluation was the section "Collection" of the Cleveland Museum of Art website (www.clevelandart.org), which describes the museum artworks. The artworks are organized in 19 sub-collections organized according to geographical criteria (e.g., African Art, Korean art, …) and works types (e.g., Drawings, Textiles, …). For each artwork an overview and detailed information are provided. Since we did not have enough time for the evaluation of the entire application, we decided to focus students' analysis on a section that is typical of most museum web sites, has an intuitive semantics, and, in the case of the Cleveland museum, is still quite large (approximately 300 pages). To facilitate the reporting activity, and our analysis of the results, every inspector received a report template to fill, structured in terms of:

- *NAME (*of the problem)
- *DIMENSION (*the design dimension the problem refers to, e.g. Content, Navigation, Semiotics…)
- *DESCRIPTION (*of the problem - maximum three lines)
- page *URLs* (the pages – at least three for typed pages, where an instance of the problem was detected).

The inspectors did not know the website they would have to analyze before the inspection day. Ten minutes before the session they received an overview of the application (e.g. application's goals and general structure) and the specification of a *scenario* relevant for the Collection Section.

3.4. Empirical Study 2: MiLE+ evaluation "project"

With the second study we aimed at collecting qualitative and quantitative data about the *effort* needed to perform a *professional* evaluation of an entire, significantly complex web application. In particular, we wanted to explore the distribution of the effort on the different tasks envisioned by a MiLE+ evaluation process, i.e., technical inspection, user experience inspection, negotiation of problems among different evaluators within a team, and production of the final documentation. To this end, with respect to the controlled lab experiment of the first study, we attempted to simulate and investigate an as much as possible *realistic* process of MiLE+ evaluation, as it is

carried on by a team of usability experts in a professional environment. This evaluation involved the *Milano group* (*26 students*) and spanned from the mid of the semester to its end (two months). To ensure that the testers had an acceptable level of knowledge on MiLE+, we started the experiment approximately three weeks after the classroom lessons and involved only students who had successfully passed an intermediate written exam about MiLE+. The students had to perform the MiLE+ evaluation of an entire web site (as part of the exam), working in teams of 3-4 persons, and to produce an evaluation report of professional quality structured according to a format defined by the course teachers. The subject of evaluation could be freely selected among a set of assigned web sites that had been identified and previously evaluated by the course teachers, had comparable complexity and suffered of a comparable amount of usability problems. At the end of the evaluation activity, students were asked to fill a *questionnaire* and to deliver it together with the project documentation. The questionnaire involved closed questions asking the global time needed to learn the different aspects of MiLE+, and to carry on the various activities.

3.5 Results and Discussion

As to the first study, we discuss only the results related to the technical inspection, which are the most relevant to understand the quality attributes to be measured. The average number of usability problems found by the Como students was 14,8. Given that the inspection lasted 3 hours, the average number of problems found in one hour (hourly *efficiency*) is 4.9. Since the total number of problems that were found in three hours by a team of usability experts for the same section of the web site is 41, the resulting *performance* is 14,8, i.e., 36%.

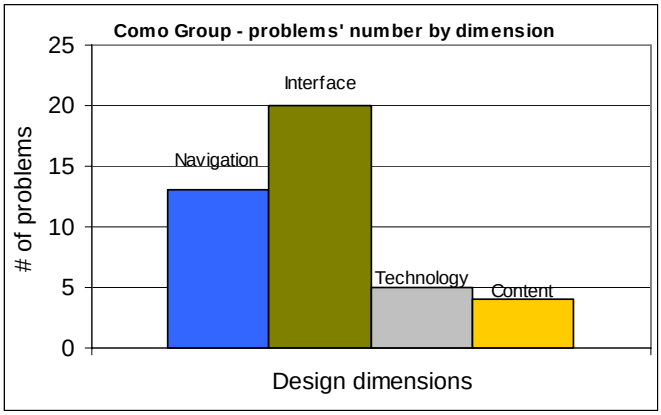


Figure 6 Distribution of discovered problems, by design dimensions.

The results can be read positively. We should consider that in 90% of the cases, the detected problems referred to repeated problem instances, i.e., problems occurring in at least three pages, and we asked students to report different problems only. This means that the average number of problems per testers (14,8) corresponds to an average number of problem instances of 40.

We should read these numbers in light of the profile of the testers and the testing conditions: basically, after 6 hours of training and approximately 10-12 hours of study, in a session of three

hours an average student has been able to identify usability problems in approximately 40 pages and to detect one third of the overall usability problems. Overall, this can be considered a very good result for a novice.

The distribution over the design dimensions (Figure 6) clearly shows a tendency in discovering more usability problems concerning interface design (e.g. labels, layout) and navigation rather than content and technological breakdowns. This does not mean that the application under inspection features fewer problems in the quality of the content than in navigation or interface, but it states that interface and navigation problems are the ones most discovered by the inspector.

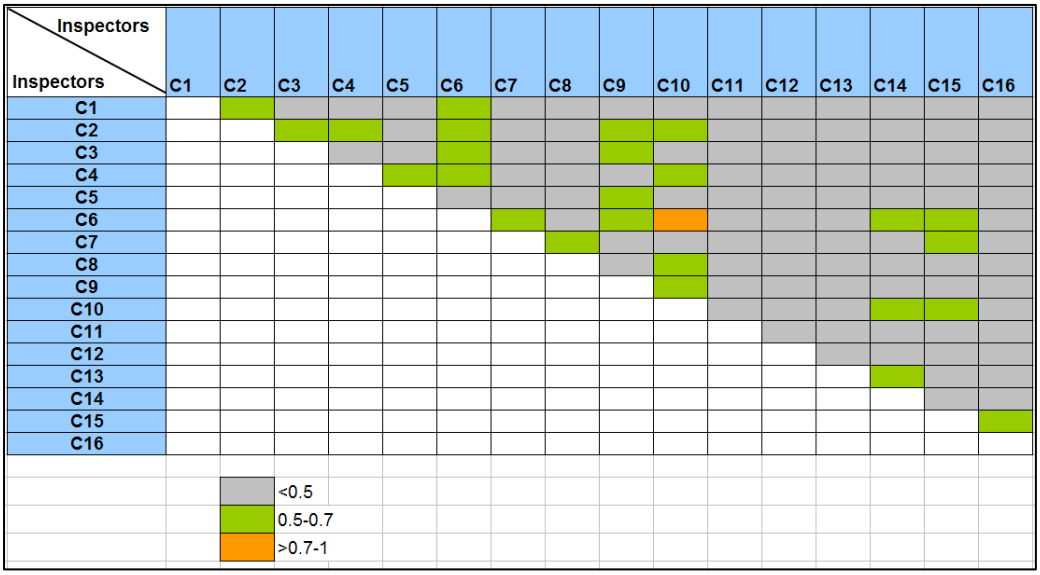


Figure 7. Shared problems rate distribution by pair of inspectors [14].

This result confirms that MILE+ well equips less-experienced practitioners in spotting interface and navigation issues, and this is also reflected in the richness of the library of navigation and interface heuristics at disposal to the inspectors, with respect to the heuristics concerning content and technology (which are far more general and less detailed). Moreover, given the limited time for the inspection (3 hours) and the rush in discovering as many problems as possible, it is reasonable to suspect that navigation and interface problems are easier to detect (with MILE+ at hand) in a limited amount of time with respect, for example, to content issues, which would require carefully analyzing and understanding the content (a more time consuming activity).

Concerning *coverage*, the average shared problems rate is 0.39, which means that 39% of the problems were found by all inspectors. Reading the same result in terms of *reliability*, it means 39% of the problems found are confirmed at least by two different inspectors. Figure 7 reports the shared-problem rate for each pair of inspectors.

For the second study, we shifted the focus from the intrinsic quality of the method to the perceived quality, in terms of cost-effectiveness and learnability.

Figure 8 and 9 highlight that the Milano students invested, on average, the same amount of time (10-13 hours) to study MiLE+ as the Como students involved in the empirical study 1, and that their large majority (73%) found the study of the method per se simple.

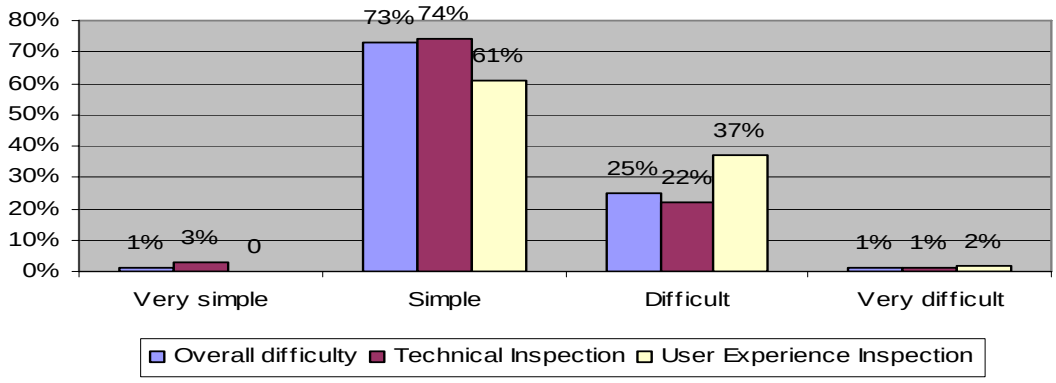


Figure 8 Perceived Difficulty of Learning MiLE+.

Only 47% of the students scored "simple" the use of MiLE+, while 53% judged it difficult or very difficult. It is interesting to notice that a significant amount of persons (83%) estimated difficult or very difficult the activity of defining the scenarios needed for User Experience Evaluation. In abstract terms, during the study, this task may appear simple but in practice it requires a deep reflection on the requirements of the application – an intrinsically complex cognitive work that a novice is oftentimes not used to.

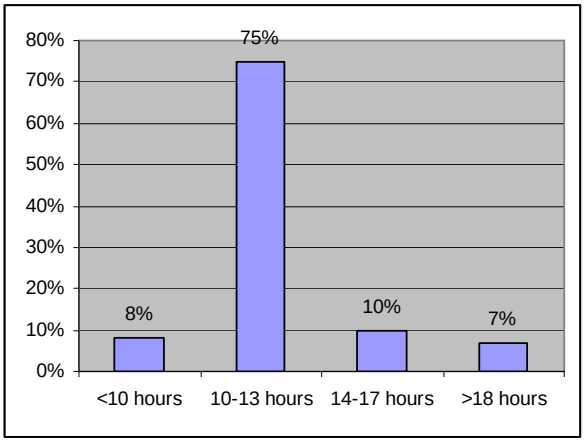


Figure 9 MiLE+ Learning Effort.

In addition (see Figure 10), the user experience evaluation was perceived, during the practice, more difficult than it was expected from the study (when it was judged easier than technical inspection). This may be considered an indication of weakness: although the number of user experience indicators is smaller than the number of heuristics, and they look simpler at a first

glance, their definition is more vague and confuse and measuring them is more difficult for a novice.

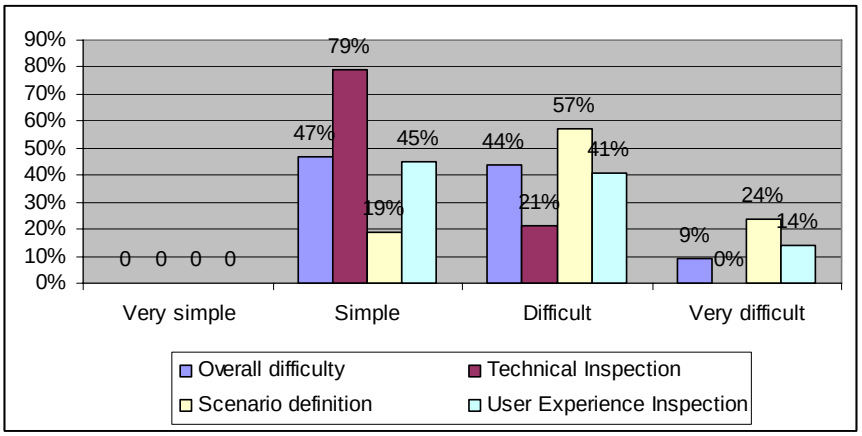


Figure 10 Perceived difficulty of applying MiLE+.

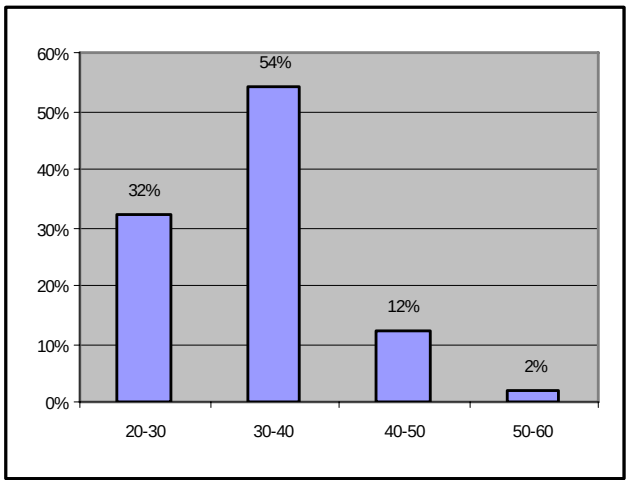
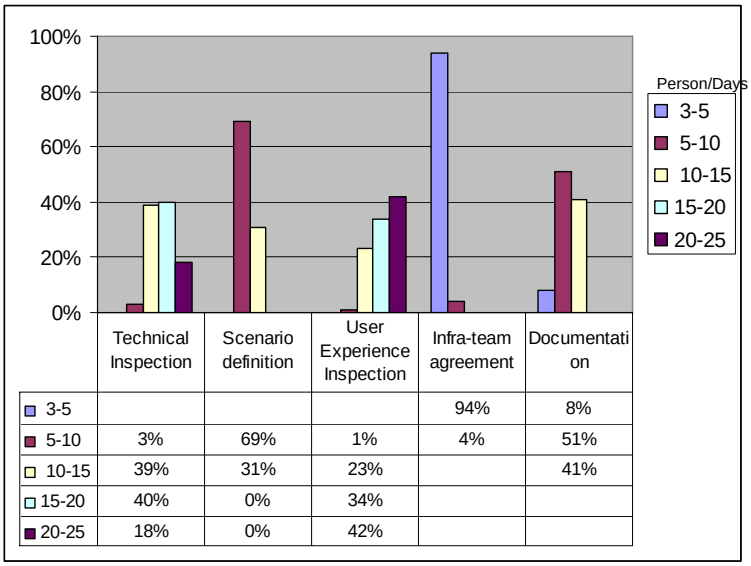


Figure 11 Individual Global Effort for a realistic evaluation process (in terms of number of person-days).

Figure 11 and 12 illustrate the average estimated effort to carry on an entire, realistic evaluation process on a complete, significantly complex application. The numbers refer, in percentage and in average, to the total person/days that a person invested in the evaluation process, either as individual work and as team work. Some interesting aspects of the process emerge from the above data:

- the activities of scenarios definition and the creation of the final documentation are both demanding tasks;
- the "negotiation activity" in order to get an agreement about the final results to be reported, resulted quite fast (3-5 hours for 94% of the persons);

the execution of the technical inspection and the user experience evaluation require a comparable effort (even though the number of user experience indicators is smaller than the number of heuristics): as we have discussed above, this results may highlight a weakness of the model, and suggest a direction of improvement.



	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 12 Individual Effort distribution per task.

In summary, from the analysis of the experimental results, MiLE+ has proved its capability to lead 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.

Our study has also indicated a good learnability of the method, given the proficiency gain by the inspector in using it, after a preliminary short training requiring an acceptable workload (10-13 hours of study). Still, shifting the inspection subject from a (relatively) small-size web site to a full-scale complex application, involves higher levels skills and competence, requires a significant effort by a design team (around 40 person hour per inspector) and involves tasks that novices may find difficult.

4 Related Work and Conclusions

There is an increasing awareness and discussion around the quality attributes of usability evaluation methods. This is due to the growing number of methodologies and techniques which are being developed either in academia or by reflective practitioners. Designers, project managers, and communication experts, especially those with little experience in usability, need clear and possibly objective criteria to choose a given usability evaluation method among the many available.

The quality criteria commonly addressed [17] fall, with few exceptions, in two main families: attributes concerning the quality of the *output* of the usability evaluation, and those concerning the quality of the *process*. As to the output of the evaluation, the number of usability problems has

been traditionally an important criterion that has been considered. Completeness or thoroughness have also been investigated as important factors. Validity has been also claimed to be crucial, as it establishes how well an inspector can predict the actual behaviour of the user [18]. A largely unexplored aspect of the research is the proficiency of the inspectors in using the method with respect to the effort spent in learning it. In this perspective, we have specifically addressed the learnability in our study, an essential precondition to enable the inspectors to yield results with an acceptable expenditure of energy and cost.

Moreover, with respect to the existing research in the area, we have extended the traditional set of quality attributes to be considered, to include the elements of "perceived quality", such as the "perceived difficulty" in using the method and the level of perceived confidence of the inspectors in using it. In fact, as shown by Rogers [19], the attributes of perceived quality are those which play a key role in eventually bring people to decide whether or not to adopt an innovation.

We acknowledge that our effort was devoted mainly in assessing the perception and adoption of MILE+ among novice evaluators (e.g. students who might be future usability professionals). In this study, this choice was based on our experimental conditions (university environment), which favour the opportunity of evaluating MILE+ with a specific target audience. A long-standing experience in teaching and working on MILE+ with more seasoned usability and new media professionals, however, suggests – even if it has not been systematically assessed – that the potential for adoption of the MILE+ philosophy and basic methodological principles is indeed quite high.

As a final remark, instead of just focussing on the quality attributes of the method *per se*, considering its intrinsic characteristics, this work has attempted to provide empirically grounded measures for the *potential for adoption* of a specific evaluation method. A take away message for the paper, besides the actual validation of MILE+ as usability inspection method, is that the abovementioned quality criteria, although expressed by different names and measured in different ways should be ultimately seen in their perspective to foster adoption. They should reasonably motivate practitioners and scholars to confidently adopt a given method in their practice, and proficiently gain, over time, appreciable and visible results. Every adoption of a new usability evaluation method is a significant investment in energy, learning effort, time and human cost. The adoption of the method strongly depends on whether this investment is acceptable in terms on effort to be spent and whether it is likely to be rewarded by the actual relevance and insights yielded by its use. We argue that the result of this work benefits scholars - as it provides validation examples and a set of quality attributes to apply to other usability evaluation methods (to make them more acceptable in the industry) – as well as practitioners, since it provides a clear guidance about the criteria to use when selecting a usability evaluation method for their own project needs.

Acknowledgements

The authors are grateful to all the people who participated in the evaluation study, to the instructors, tutors and students who made the empirical study possible and to Luca Triacca for the precious work on MILE+ during his doctoral research. We also thank all the colleagues who helped us improve the quality of the manuscript.

References

1. Bolchini D., Triacca L., Speroni M. MiLE: a Reuse-oriented Usability Evaluation Method for the Web, Proc. HCI International Conference 2003, Crete, Greece.
2. Brinck, T., Gergle, D., Wood, S.D., Usability for the web, Morgan Kaufmann, 2002.
3. Carroll, J., Making Use – Scenario-based design of Human-Computer Interactions, MIT Press, 2002.
4. Cato, J., User-Centred web Design, Addison Wesley, 2001.
5. 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.
6. Dix, A., Finlay, J., Abowd, G., and Beale, R., Human Computer Interaction, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1998.
7. Lim, K.H., Benbasat, I. and Todd, P.A., An experimental investigation of the interactive effects of interface style, instructions, and task familiarity on user performance, ACM Trans. Comput. Hum. Interact., vol. 3, no 1,pp. 1-37, January 1996.
8. 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.
9. Nielsen, J., Designing Web Usability, New Riders, 1999.
10. Nielsen, J., Mack, R., Usability Inspection Methods, Wiley 1994.
11. Rosson, M.B., Carroll, J., Usability Engineering, Morgan Kaufmann, 2002.
12. Triacca L, Bolchini D., Botturi L., Inversini A, (2004). MiLE: Systematic Usability Evaluation for E-learning Web Applications. ED Media 04, Lugano, Switzerland.
13. 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.
14. Triacca, L., Web Usability – Enhancing the effectiveness of the methodologies and improving their communication features, PhD Thesis, Università della Svizzera italiana, 2005.
15. Bolchini, D., Paolini, P., Interactive Dialogue Model: a Design Technique for Multi-Channel Applications, IEEE Transactions on Multimedia, 8 (3) 2006, 529-541.
16. Bolchini, D., Garzotto, F., Designing Multichannel Web Applications as “Dialogue Systems”: the IDM model, in Rossi, G.; Pastor, O.; Schwabe, D.; Olsina, L. (Eds.), Web Engineering: Modelling and Implementing Web Applications, Series: Human-Computer Interaction Series, Vol. 12, ISBN: 978-1-84628-922-4, Springer, October 2007.
17. Hartson, H.R., Andre, T.S., Williges, R.C., Criteria for Evaluating Usability Evaluation Methods, in International Journal of Human Computer Interaction, 15 (1), 2003, 145-181.
18. Gray,W. D., & Salzman, M. C., Damaged merchandise? Areview of experiments that compare usability evaluation methods. Human–Computer Interaction, 13, 1998, 203–262.
19. Rogers, Everett M. (2003). Diffusion of Innovations, Fifth Edition. New York, NY: Free Press.