Quality of Web Usability Evaluation Methods: An Empirical Study on MiLE+

Davide Bolchini[1] and Franca Garzotto[2]

[1] TEC-Lab, Facoltà di Scienze della Comunicazione,
Università della Svizzera Italiana
Via G. Buffi 13 TI - 6900 Lugano (Switzerland)
[2] HOC – Hypermedia Open Center
Department of Electronics and Information, Politecnico di Milano
Via Ponzio 34/5, 20133 Milano (Italy)
davide.bolchini@lu.unisi.ch, franca.garzotto@polimi.it

Abstract. What are the quality factors that define a "good" usability evaluation method and contribute to its acceptability and adoption in a real business context? How can we measure such factors? This paper investigates these issues and proposes to decompose the broad, general concept of "methodological quality" into more measurable, lower level attributes such as *performance*, *efficiency*, *cost effectiveness,* and *learnability.* We exemplify how to measure such attributes, reporting an empirical evaluation study of a usability inspection method for web applications called MiLE+.

Keywords: web usability, quality, empirical study, inspection, heuristics.

1 Introduction

In spite of the large variety of existing usability evaluation methods, both for interactive systems in general [4, 6, 12, 13], and for web applications in particular [2, 10, 11, 14], the factors that define their *quality* are seldom discussed in the literature, and relatively few empirical studies exist that attempt to measure them [5, 9, 15]. Consider for example heuristic evaluation, one of the most popular methods to inspect the usability of web sites [11, 12]. It is claimed to be "simple" and "cheap", implicitly assuming that these attributes are quality factors. Still, little empirical data supports these claims, which are mainly founded on informal arguments (e.g., "few simple heuristics", "no user involvement", "no need of special equipment").

Understanding the quality factors for usability evaluation methods, defining proper measurement procedures, and developing sound comparative studies, not only represent a challenging research arena, but may also pave the ground towards the *industrial acceptability* of these methodological "products": the empirical evidence of quality is a key force to promote a method and to have it accepted and adopted in a real business context.

This paper investigates the concepts of quality and quality measurement for web usability evaluation methods, aim at raising a critical reflection on these issues. We propose to decompose the general concept of methodological quality into lower level,

M. Weske, M.-S. Hacid, C. Godart (Eds.): WISE 2007 Workshops, LNCS 4832, pp. 481–492, 2007.

more measurable attributes such as *performance*, *efficiency*, *cost effectiveness,* and *learnability*. We also discuss an empirical study in which we measured the above factors for a specific web usability inspection method called MiLE+.

2 MiLE+ at a Glance

MiLE+ (Milano Lugano Evaluation Method – version 2) is the evolution of two previous inspection techniques for the usability of hypermedia and web applications - SUE [10] and MiLE [1, 14] - developed by the authors' research teams. It also borrows some concepts from various "general" usability evaluation methods (heuristic evaluation, scenario driven evaluation, cognitive walkthrough, task based testing).

The main purpose of MiLE+ is to be *more systematic* and *structured* than its "inspirators", and to be particularly suited 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.

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

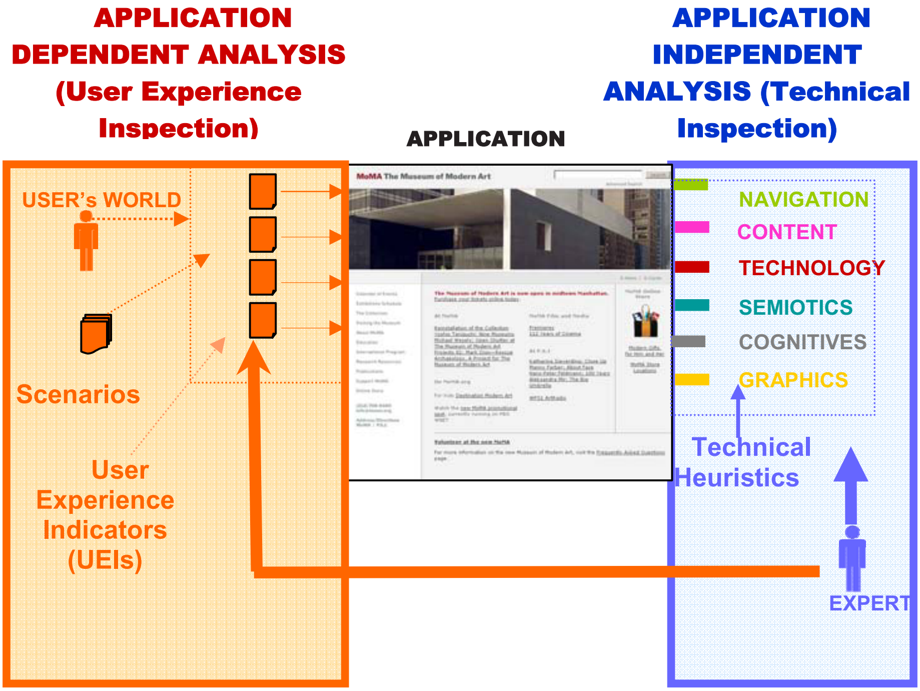

APPLICATION DEPENDENT ANALYSIS (User Experience Inspection)
APPLICATION
APPLICATION INDEPENDENT ANALYSIS (Technical Inspection)
USER's WORLD
Scenarios
User Experience Indicators (UEIs)
MoMA The Museum of Modern Art
NAVIGATION
CONTENT
TECHNOLOGY
SEMIOTICS
COGNITIVES
GRAPHICS
Technical Heuristics
EXPERT

Fig. 1. MILE+ at a glance

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* (see examples in Table 1)*:*

- *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).

Table 1. Classification of MiLE+ Technical Heuristics

Dimension		Examples of Heuristics
Navigation		Consistency of navigation patterns
		Index Backward Navigation
Content		Text accuracy
		Multimedia consistency
Technology/Performance		System reaction to user errors
		Operations management
Interface design		
	Cognitive	Information overload
		Scannability
	Graphics	Background contrast
		Text layout
	Semiotics	Ambiguity of link labels
		Conventionality of interaction images

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.

Table 2. Examples of MiLE+ User Experience Indicators

Categories	Examples of User Experience Indicators
Content Experience UEIs	Completeness
	Multilinguism
Navigation & Cognitive Experience UEIs	Predictability
	Memorability
Operational Flow Experience UEIs	Naturalness
	Recall

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* [3,4] to guide User Experience Evaluation. In general terms, scenarios are "stories of use" [3]. 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 (see Table 3). 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.

Table 3. A MiLE+ scenario for a museum website

Scenario description	A well-educated American tourist knows he 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

In principle, scenarios should be extracted from the documentation built during user requirements management or design (the application development phases in which scenarios are frequently used). In practice, in most cases such documentation is missing and scenarios are defined by the evaluators in cooperation with the different stakeholders (the client, domain experts, end-users, …).

3 Quality Attributes for a Usability Evaluation Method

Quality is a very broad and subjective concept, oftentimes defined in terms of "fitness to requirements" [7], 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 [8]. 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.

Definition 1: Performance

Performance 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 (P_i) in given inspection conditions (e.g. time at disposal) against the total number of existing problems (P_{tot})

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

Definition 2: Efficiency

Efficiency 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:

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

where P_i is the number of problems detected by the *i-th* inspector in a time period $t_{i.}$

Definition 3: Cost-effectiveness

Cost-effectiveness 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.

Definition 4: 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.

All the above definitions use, explicitly or implicitly, the notion of *usability problem,* which deserves a precise definition for web applications. Clearly, *a* usability problem has to do with a violation of a usability principle (heuristic, user experience indicator…) in some pages of the application. We must consider that most pages might be "typed", i.e., they share content structure, lay-out properties, and navigational or operational capabilities as defined by their "type" or "class". If a usability violation occurs in one page of a given type, it may occur in other, if not all, pages of the same type, which share the same design. Thus we will count the violations of the *same* principle in *a set* of *pages of the same type* as *one* usability problem.. In contrast, when we consider untyped, or "singleton", pages that represent a "unique" topic or functionality and cannot be reduced to a class, we should count *each* violation in *each* singleton page as one problem. This approach is expressed by the following definition:

Definition 5: Usability Problem
A Usability Problem is a violation of a usability principle in a singleton page, or the equivalence class of the violations of the same usability principle in any set of pages of the same type.

4 An Empirical Study on MiLE+

The purpose of our empirical study was to measure the "quality" of MiLE+ evaluation process in terms of the factors defined in the previous section: 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 MiLE+ during the course, for approximately *5 hours* consisting of an introduction to MiLE+, discussed of evaluation case studies, and Q&A sessions. All students were provided with the same learning material, composed of: a MiLE+ overview article [1], the "MiLE+ Library of Technical Heuristics and User Experience Indicators" (including guidelines and examples), the complete professional evaluation reports in two industrial cases, course slides, an Online Usability Course developed by the University of Lugano (http://athena.virtualcampus.ch/webct/logonDisplay.dowebct).

4.2 Procedure of Study 1: 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 MiLE+ to evaluate a portion of an assigned 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 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 MiLE+.

4.3 Procedure of Study 2: Mile+ Evaluation "Project"

The purpose of Study 2 was to investigate the *perceived difficulty* of *learning* and *using MiLE+*, and the *effort* needed to perform a *professional* evaluation. We also wanted to explore the effort needed for the different 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 two months time period, from the mid to the end of semester 2. Since we wanted to investigate an as much as possible *realistic* evaluation process using 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 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 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 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.

4.4 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.

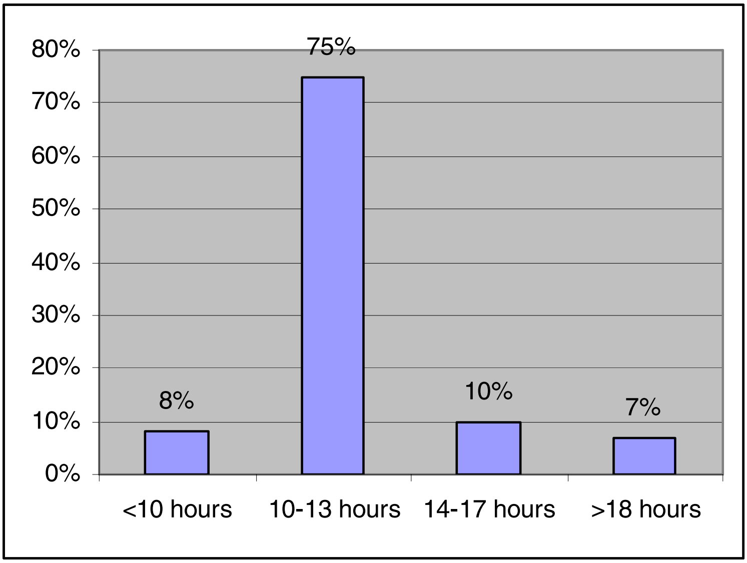


Fig. 2. MiLE+ Learning Effort

Concerning the *learning effort*, participants invested in the preliminary study of MiLE+ an average amount of time of *10-13 hours* (see Fig. 2), which is comparable with the estimated effort of Como students. Concerning *learning difficulty*, a large majority of participants (*73%*) found MiLE+ study activity *rather simple*- see Fig. 3.

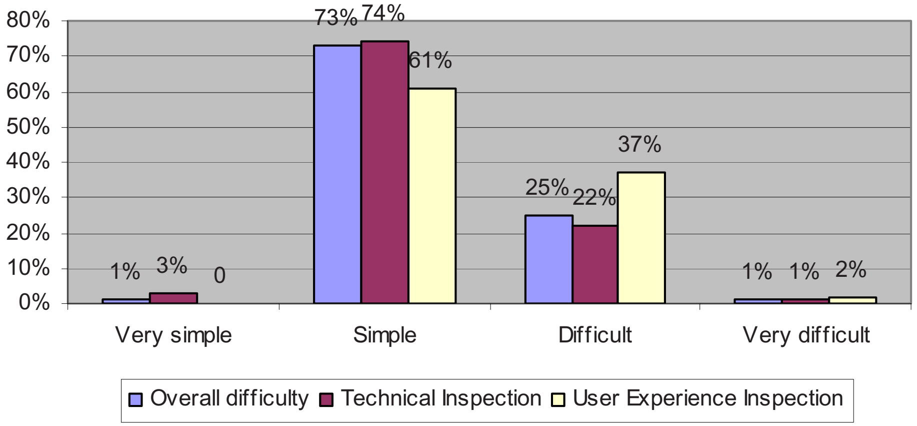


Fig. 3. Perceived Difficulty of Learning MiLE+

Fig. 4 highlights that students perceived the *use* of MiLe+ in a real project as more complex than studying it. Only *47%* of the students scored "*simple*" the use of MiLE+, while *53%* judged it *difficult* or *very difficult*.

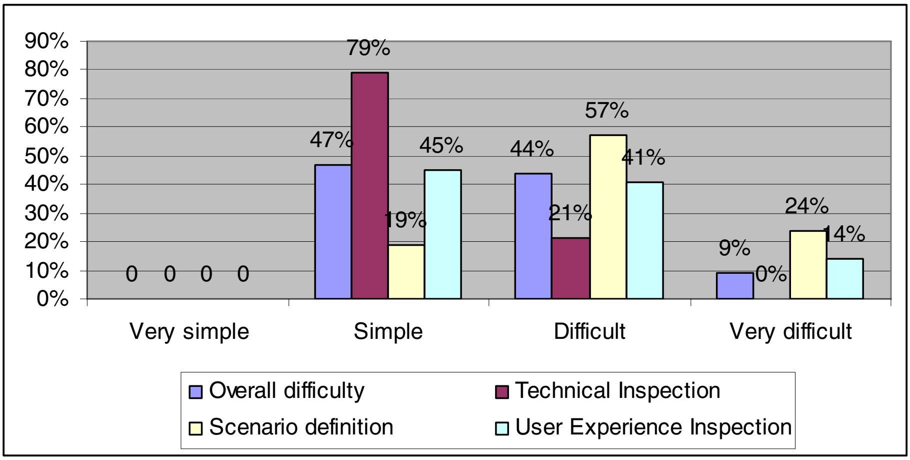


Fig. 4. Perceived Difficulty of Using MiLE+

Fig. 4 also shows that the User Experience Evaluation was perceived slightly more difficult than it was expected from the study (compare fig. 4 with fig. 3). These data may indicate a weakness of the MiLE+ method: although the number of User Experience Indicators (32) is smaller than the number of Technical Heuristics (82), 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*. Fig. 4 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.

Concerning *cost-effectiveness,* Fig. 5 & 6 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.

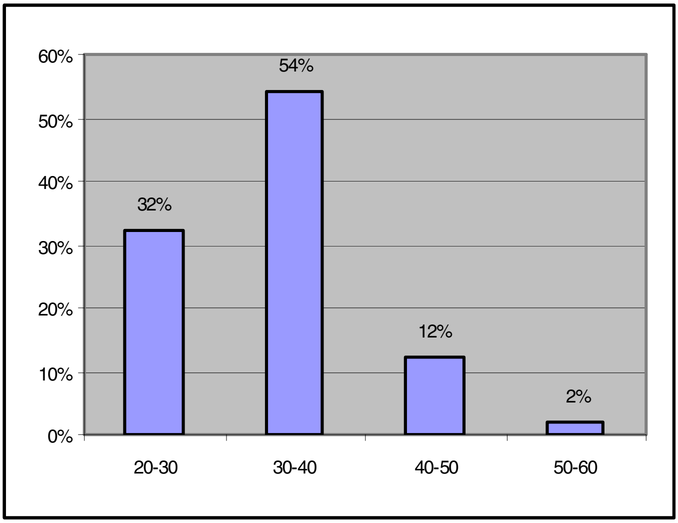
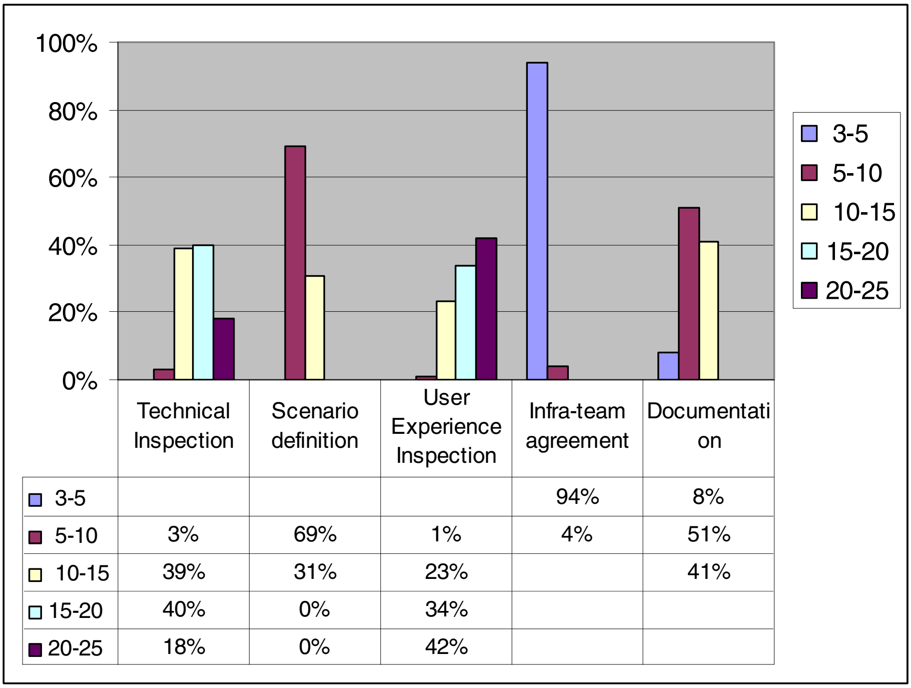


Fig. 5. Individual Effort for a Professional Evaluation Process (in person/hours)

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 business context
- consistently with the results in Fig. 4, 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 MiLE+ supports the standardization of the inspection process and the homogenization of results.



	Technical Inspection	Scenario definition	User Experience Inspection	Infra-team agreement	Documentati on
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%		

Fig. 6. Individual Effort distribution per Task in a Professional Evaluation Process

In summary, the analysis of the experimental results has proved that 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 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. In this paper, we suggest that *learnability*, *performance*, *efficiency*, and *cost effectiveness* are possible measurable attributes for methodological quality of web usability evaluation techniques, since they are critical factors for the *acceptance* and *adoption* of methodological products in the practitioners’ world. We have discussed how the above factors can be measured, presenting an empirical study that evaluated the quality of the MiLE+ usability inspection method.

Our work is only a first step towards the definition of a quality assessment framework for web usability evaluation methods, and further discussion and investigation of these concepts are needed. We plan to perform the evaluation of other methods (e.g., Nielsen's heuristic evaluation and walkthrough) using our quality criteria and metrics, both to compare these techniques with MiLE+, and to test the soundness of our quality approach.

Acknowledgments. The authors are grateful to Luca Triacca for his contribution in study 1, and to all students from our HCI classes in Milano and Como participating in our research.

References

1. Bolchini, D., Triacca, L., Speroni, M.: MiLE: a Reuse-oriented Usability Evaluation Method for the Web. In: Proc. HCI International Conference 2003, Crete, Greece (2003)
2. Brinck, T., Gergle, D., Wood, S.D.: Usability for the web. Morgan Kaufmann, San Francisco (2002)
3. Carroll, J.: Making Use – Scenario-based design of Human-Computer Interactions. MIT Press, Cambridge (2002)
4. Cato, J.: User-Centred web Design. Addison Wesley, Reading (2001)
5. De Angeli, A., Costabile, M.F., Matera, M., Garzotto, F., Paolini, P.: On the advantages of a Systematic Inspection for Evaluating Hypermedia Usability. In International Journal of Human Computer Interaction, Erlbaoum Publ. 15(3), 315–336 (2003)
6. Dix, A., Finlay, J., Abowd, G., Beale, R. (eds.): Human Computer Interaction, 2nd edn. Prentice-Hall, Englewood Cliffs (1998)
7. Fenton, N.E. (1991) Software Metrics: A Rigorous Approach, 2nd edn. Chapman & Hall, Sydney, Australia (2002)
8. Garzotto, F., Perrone, V.: Industial Acceptability of Web Design Methods: an Empirical Study. Journal of Web Engineering 6(1), 73–96 (2007)
9. Lim, K.H., Benbasat, I., 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., 3(1), 1–37 (1996)
10. 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 32(1) (2002)
11. Nielsen, J.: Designing Web Usability. New Riders, Indianapolis (1999)
12. Nielsen, J., Mack, R.: Usability Inspection Methods. Wiley, Chichester (1994)
13. Rosson, M.B., Carroll, J.: Usability Engineering. Morgan Kaufmann, San Francisco (2002)
14. Triacca, L., Bolchini, D., Botturi, L., Inversini, A.: MiLE: Systematic Usability Evaluation for E-learning Web Applications. In: ED Media 04, Lugano, Switzerland (2004)
15. Whiteside, J., Bennet, J., Holtzblatt, K.: Usability engineering: Our experience and evolution. In: Helander, M. (ed.) in Handbook of Human-Computer Interaction, pp. 791–817. North-Holland, Amsterdam (1988)