A CONTROLLED EXPERIMENT FOR MEASURING THE USABILITY OF WEBAPPS USING PATTERNS[i]

F. Javier García, María Lozano, Francisco Montero, Jose Antonio Gallud, Pascual González
Computer Science Department,University of Castilla-La Mancha, Albacete, Spain

Carlota Lorenzo
Marketing Department, Universitity of Castilla-La Mancha, Albacete, Spain

Keywords: Internet services, Dial-up networking

Abstract: Usability has become a critical quality factor of software systems in general, and especially important regarding Web-based applications. Measuring quality is the key to developing high-quality software, and it is widely recognised that quality assurance of software products must be assessed focusing on the early stages of the development process. This paper describes a controlled experiment carried out in order to corroborate whether the patterns associated to a quality model are closely related to the final Web application quality. The experiment is based on the definition of a quality model and the patterns associated to its quality criteria to prove that applications developed using these patterns improve its usability in comparison with other ones developed without using them. The results of this experiment demonstrate that the use of these patterns really improves the quality of the final Web application in a high degree. The experiment is formally based on the recommendations of the ISO 9126-4.

1 INTRODUCTION

Usability has become a critical quality factor of software systems in general, but with the increase use of internet for everyday activities, it is especially important in web-based applications. Usability is a key factor for users to decide whether or not a web application (WebApp) is satisfying.

Thus, it is important to design WebApps with a basic level of quality in general and usability in particular, and also developing methods which allow evaluating this quality. In this sense, different Web metrics have been defined (Ivory, 2001), (ISO, 2001), which can help us to measure the final usability or quality in use of a web application.

Lots of efforts have been made in order to improve the quality of software systems, such as definition of quality metrics (Olsina, 1999), usability metrics (Ivory, 2001), usability evaluation methods (Nielsen, 93), (Constantine et al., 2000), etc. All these mechanisms allow us to check the usability of a software system but they have to be used on a final and running application. Nowadays, it is widely recognised that quality assurance of software products must be assessed focusing on the early stages of the development process. It is necessary to incorporate new mechanisms at the very beginning of the software process to produce high-quality software applications. In this sense, the use of design patterns in general has proved to be good for these purposes, especially interaction patterns regarding usability.

Some interaction patterns have been proposed in the literature, but the novelty of our approach is to establish a clear association between a quality factor defined in a quality model and one or more concrete patterns in such a way that the use of that pattern in the construction of a WebApp makes it to assess the corresponding quality criteria.

To validate these associations and the goodness of using interaction patterns to satisfy the corresponding quality factors, a controlled experiment has been carried out.

This paper is organized as follows: firstly, we present a general idea about web quality, usability and patterns, to have a global vision of the frame work where this study is included and the related work.

Javier García F., Lozano M., Montero F., Antonio Gallud J., González P. and Lorenzo C. (2005).
A CONTROLLED EXPERIMENT FOR MEASURING THE USABILITY OF WEBAPPS USING PATTERNS.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems - DISI*, pages 67-74

Then we describe the experiment that we have carried out to prove the importance of using patterns and quality models to design quality WebApps, as well as the use of different metrics as a mechanism for evaluating usability.

Finally, we finish the paper with interesting conclusions based on the experiment results and future research aspects are proposed.

2 RELATED WORK

Before describing the experiment carried out for this study is necessary to talk about web quality and web usability to establish the aim and the context of the experiment.

Brajnik states (Brajnik, 2002) that the quality in this context is a property of a website defined in terms of a group of attributes, like consistency of background colours or average download time. ISO 9126 defines web quality dividing it into more abstract terms, as effectiveness, efficiency and satisfaction. (ISO, 2001).

Because of the variety and quantity of quality attributes proposed in the literature, it is necessary to develop a quality model that helps to know which attributes are important for the analysis, which one is more important than others, and which measurement methods have to be used to assess the attributes values.

Usability is widely recognized as one of the most important attributes regarding Web quality, so we focus on usability to define a complete quality model (see figure 1) and define an accurate association between the final and more concrete criteria with one or more interaction pattern. The experiment aims to prove the goodness of the quality model and the interaction patterns associated.

The model is centred specifically on usability criteria as regarding web environments this factor is more meaningful than others because of the new interactive features of internet.

This quality model (Montero et al. 2003) centred on usability is based on usability features defined by ISO 9126, some ergonomic criteria and its sub-criteria. These ergonomic criteria are implemented by means of patterns, as a way to materialize the abstract concept of a quality criterion with something implementable as it is a pattern.

Pattern concept in computer science in this context is defined as a tuple of three elements: a problem, a context and a solution. Many patterns have been proposed for web development (Percel et al.,1999), (Tidwell, 2002), (Welie, 2003), (Van Duyne et al., 2002), (Rossi et al., 12), or (Montero et al., 2002b) can be cited.Once we have seen the general terms and web quality model we base the experiment on, we continue this paper describing the case study about assessing the usability of an e-commerce WebApp by using patterns and the complete description of the experiment.

3 DESCRIPTION OF THE EXPERIMENT

The description of the experiment follows the structure and recommendations of ISO 9126-4.

3.1 Goals

The website that we use for the experiment is a fictitious on-line store named e-fashion in order to eliminate the effects of prior experience. This site is an e-commerce WebApp where you can buy men and women clothes.

The content of an online shop is based on a homepage (Nielsen, 2001) which includes the same links and websites as other online apparel stores.

We developed six different versions of this site according to different quality patterns we wanted to validate.

The main goal of the experiment is to prove that the association established between interaction patterns and the criteria defined on the proposed quality model (Montero et al. 2003) is correct (see figure 1), in such a way that the use of the patterns in the construction of a WebApp makes it to assess the corresponding quality criteria, and for this reason the WebApp versions that includes the recommended patterns for solving usability problems have more quality that the ones that do not include them. In this case we carried out the experiment choosing only one of the quality model criteria: “Guide”. The same method could be used with the rest of quality factors and patterns related.

We had to evaluate the websites designed for the experiment using usability metrics to validate the hypotheses.

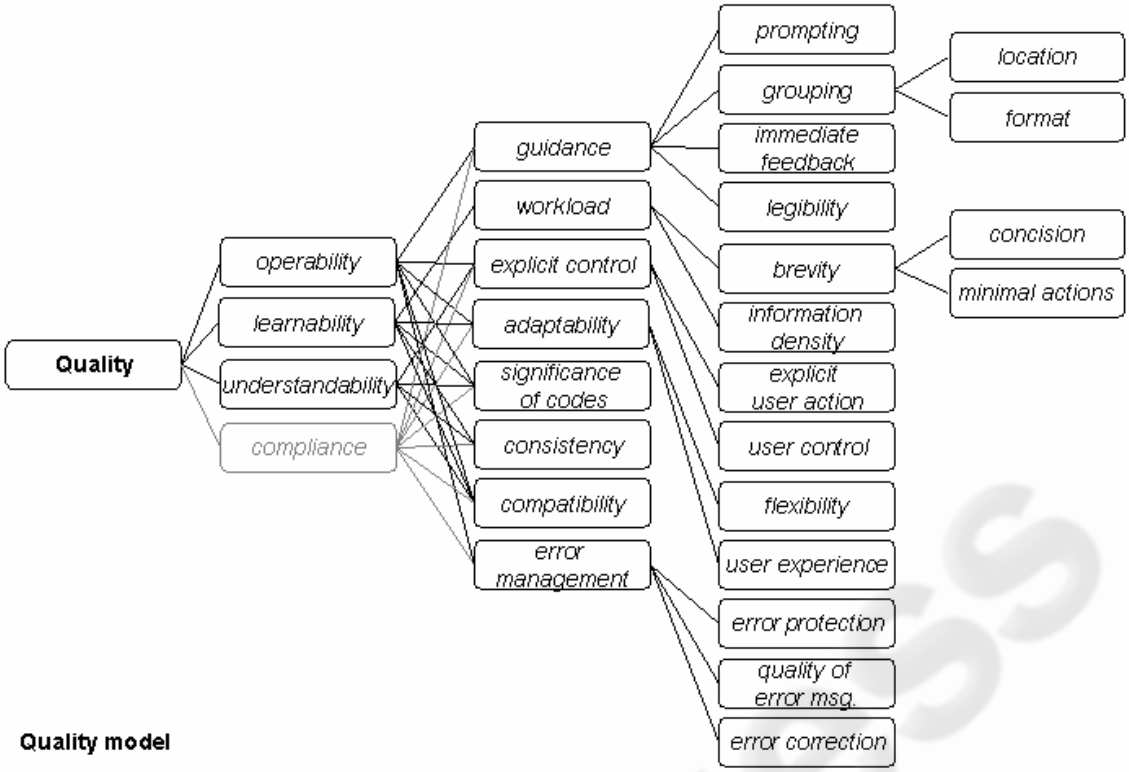

Figure 1: Web Quality Model

3.2 Method

To do the experiment for each feature of the criteria "Guide" represented in the quality model we based on, we had to design six versions of e-fashion. Thus, we had to choose six different groups of people, with the same experience on the use of internet and buying in electronic shops.

Each group had to do four tasks, and then we had to measure them with usability metrics to reach the goals of the experiment.

Each version of e-fashion was made using the patterns that are recommended according to the features of the quality criteria "Guide" of our quality model (Montero et al. 2003). We wanted to measure if the association established between the patterns and this features is correct and their use improves the usability of the WebApp generated.

The sub-criteria defined for the criteria "Guide" are the following:

-Prompting
-Grouping
-Immediate feedback
-Legibility

The patterns, defined by D. Van Duyne (Van Duyne et al. 2002), associated to each sub-criteria are the following:

- *Prompting*:
  - **D3***: Headlines and blurbs*
  - **D9***: Distinctive HTML titles*
  - **G1***: Featured products*
  - **G2***: Cross-selling and up-selling*
  - **H6***: Pop-up windows*
  - **H8***: Context-sensitive help*
  - **K6***: Location bread crumbs*

- *Grouping*:
  - **B3***: Hierarchical organization*
  - **B4***: Task-based organization*
  - **B5***: Alphabetical organization*
  - **B6***: Chronological organization*
  - **B7***: Popularity-based organization*
  - **D1***: Page templates*
  - **D7***: Inverse-pyramid writing style*
  - **F1***: Quick-flow checkout*
  - **G1***: Featured products*
  - **H7***: Frequently asked questions*
  - **J3***: Organized search results*
  - **K1***: Navigation bar*

- *Immediate feedback*:
  - **C2***: Up-front value proposition*
  - **D9***: Distinctive HTML titles*
  - **F3***: Shopping cart*
  - **F7***: Order summary*

F8: *Order confirmation and thank-you*
H6: *Pop-up windows*
I3: *Clear first reads*
K5: *High-visibility action buttons*
K10: *Obvious links*
K14: *Page not found*

- *Legibility*:
D7: *Inverse-pyramid writing style*
D9: *Distinctive HTML titles*
I2: *Above the fold*
I3: *Clear first reads*
I4: *Expanding width screen size*
I5: *Fixed-width screen size*

Taking into account these associations, we implemented the six different websites of e-fashion: one version using all the patterns (e-fashion 3), another one using any of them (e-fashion 4), another one using only the patterns defined for the sub-criteria *Prompting* (e-fashion 5), another one using only the patterns defined for the sub-criteria *Grouping* (e-fashion 6), another one using only the patterns defined for the sub-criteria *Immediate Feedback* (e-fashion 7), and the last one using only the patterns defined for the sub-criteria *Legibility* (e-fashion 8).

3.2.1 Participants

The selection of participants was not easy as they should not be experts on using internet and buying through websites.

Finally the participants selected to carry out the experiment were high school students between 16 and 17 years. All of them had the same experience on using internet and no experience on buying in internet. There were only 2 users that had previously bought something in internet but just once.

3.2.2 Context of Product Use in the Experiment

3.2.2.1 Tasks

The proposed tasks that each user had to do on the experiment were designed according to the factor "Guide" that we wanted to evaluate and considering that the WebApps designed for the experiment were e-commerce sites.

The tasks each user had to do were the following:

- **Task 1**: To buy two grey men jerseys.
- **Task 2**: To write the price of a white woman shirt with black stripes
- **Task 3**: To add to the shopping chart four pairs of uncovered woman shoes and a white man belt.
- **Task 4**: To buy a brown woman skirt.

After the execution of these tasks the users had to do the satisfaction test proposed in the experiment.

With the results obtained using metrics we got conclusions about the use of patterns to create usable websites, as described afterwards.

3.2.2.2 Context Used for the Experiment

The evaluation was made on the Albacete high school called CEDES, in Spain, on June the 15th and 16th of 2004.

The participants were constantly observed by the person in charge of the experiment during the whole time.

3.2.2.3 Participant's Computing Environment

All the participants used the same computer machines and the same internet connection. The computers used were Pentium MMX with 32 MB of RAM, with 15" monitors and a screen resolution of 800x600. The operating system was Windows 2000.

The internet connection was 150 kb/s ADSL, but shared by all the machines. This was determining in the development of the experiment, because the loading of the pages was very slow, because of the high quantity of images shown in the web.

3.2.3 Experiment Design

Six groups of participants were established, with a media of 12 users per group. Thus, the total amount of participants was 74 peoplel. Each group of users did the proposed tasks in one of the versions of e-fashion. Each user had to do the 4 proposed tasks and to fill the satisfaction test.

3.2.3.1 Procedure

When the participants arrived at the laboratories where the experiment was carried out, all of them were informed about the goals of the experiment at the same time. We told them that the experiment was made to measure the usability of the website e-fashion where they had to do the proposed tasks to find out whether it met the needs of users as them. They were told to read each task (described on a link on the home page of e-fashion) and to make these tasks one by one. Finally they had to fill the satisfaction test available by a link on the home page.

The experiment was about 50 minutes long for each one of the six groups. In this time the users had to be able to execute the four proposed tasks and the satisfaction test.

The participants were given basic instructions describing the environment. The evaluator reset the state of the computers before the coming of each new users group, and gave them the appropriate and

convenient instructions. The participants also were informed that they could not be helped by the evaluator, because the web was enough to help the users to make the tasks properly.

Any incident occurred during the experiment was solved by the evaluator in the most properly way.

The evaluator finally asked the participants about the difficulties they had encountered to have a best vision about the results of the experiment.

3.2.4 Metrics

The metrics we used on the experiment were the following:

- ***Efficiency:*** We use as efficiency metric the "*Task Time"*, which is the time that each user spends completing each task. We did the mean time for all users on each task.

- ***Effectiveness:*** The effectiveness was measured using the metric *"Task Completion"* and *"Error frequency"*. With these metrics we can obtain the number of tasks that users did not completed and the number of errors the users had made on each task.

In this case we considered as completed task the one where the user had done what we asked to do, independently of if he did it properly or not.

As well as the number of completed and not completed tasks, we showed the results as a percentage to allow a simple analysis.

One task is not completed for example if we asked to buy something and the user has added the products to the shopping chart but he has not paid for them.

The results of the metric *"Error Frequency"* allow us to evaluate the errors that each user has made. If one of the task was to buy a brown skirt, but the user have bought another issue, it will be considered that the task is completed but with errors.

If the task consisted on taking note of the price of a certain shirt and this price was written from other shirt, then the task will be considered completed but with errors.

Another possible way to have contemplated this metrics would have been to consider that the task with errors it is not completed.

- ***Satisfaction*****:** The satisfaction metric was measured using a satisfaction test, that was created based on some questions of the evaluation test SUSS, developed by Constantine (Constantine et al., 1999) to measure key elements in the interfaces design as could be personal tastes, aesthetic, organization, understandability, and learning.

Other questions of the test are based on SUMI test, accepted by expert evaluators and international organizations as ISO.

Our test is about 20 questions that allow us to evaluate in a simple way the satisfaction of the users who have carried out the four tasks in the website.

Each user had to answer each question of the test choosing an option between 1 and 5, like a Likert scale, except in the two first questions and the three final questions that were questions with a scale of three points.

- ***Comprensibility and learning:*** Finally the metric learning was measured. This kind of metric measures how the user have learned to use the website. To measure this metric we use a special question of the satisfaction test and we compared also the *"Task Time"* of task 1 and task 4, because these two tasks consisted on buying something. Thus, it was supposed that if the user had been able to learn, the Task Time for Task 4 would be inferior to Task Time for Task 1, as indeed occurred.

3.2.5 Results

Before comparing the data obtained by the metrics in the different versions of e-fashion we can say that the association established between the interaction patterns and the feature *Guide* in general is good and improves notably the quality of websites.

Special case was e-fashion 5 because the server felt down for some minutes and some tasks could not be completed, and some Task Times were higher than expected, so the results of e-fashion 5 were taken carefully taking into account this unexpected situation.

3.2.5.1 Comparative of Medium Task Time

Figure 2 shows a graphic where the medium times of task are represented for each version of e-fashion.

We highlight the fact that it occurs something curious on the versions of e-fashion that use all the patterns of the criteria "Guide", or some of them. These sites had more information to load, for example more quantity of images, than e-fashion 4, the "worst" version of all. For this reason the loading time of these pages was slower than the loading time on e-fashion 4.

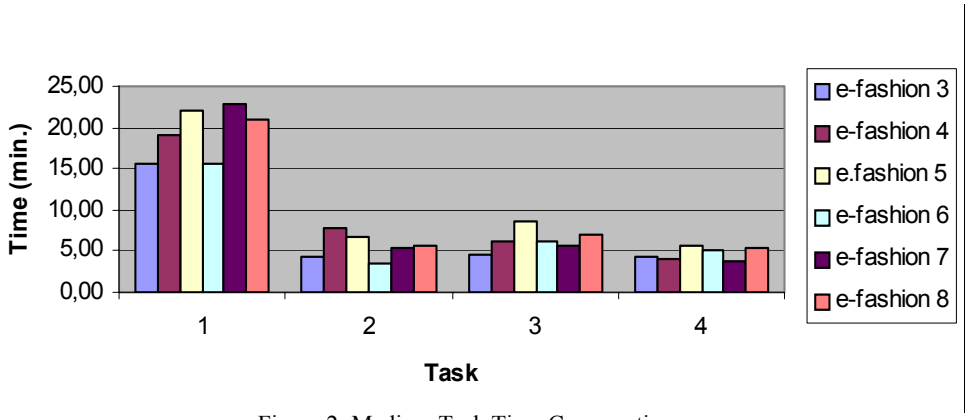

Figure 2: Medium Task Time Comparative

This workload that made the loading slower was due to products images, for example on the home page that loads the new products of the month, and this could affect to the final time task.

Despite this situation, on the Figure 2 we can see that in general, the users take more time to execute the tasks on e-fashion 4 than on e-fashion 3, as we expected. If we add the handicap that has been talked about, we have to take into account the times in function of the load of the page. In this case there is no doubt that e-fashion 4 is the worst of all versions of the sites, and e-fashion 3 is the best of the six versions because its task times are slower on 3 of the 4 tasks.

These results allow us to conclude that the association between the patterns with the criteria "Guide" is good in general and helps to improve the quality of websites.

Each colour in Figure 2 represents each version of e-fashion that was used on the experiment. The X-axis shows the number of task and the Y-axis shows the time expressed in minutes.

3.2.5.2 Comparative of Task Completion

Figure 3 shows a graphic where the completion of tasks is compared on the six versions of e-fashion. We want to remind that when the experiment was made with e-fashion 5 the server felt down, and some tasks were not completed. So, this 23% of users that had not completed the tasks on e-fashion 5 do not reflect the real usability of the web.

If we analyse the rest of groups, we can see that on e-fashion 4 there was a 90% of uncompleted tasks, what shows that this web was the worst of the six versions, because in e-fashion 7 there were a 2 % of users that did not complete all the tasks, but it is a better result than e-fashion 4 and in the rest of the versions there was a 100% of task completed.

3.2.5.3 Errors Frequency comparative

Figure 4 shows which version of e-fashion had more error frequency on its tasks, and as we expected e-fashion 4 was the worst site of the six.

This result reinforces the hypothesis that it is

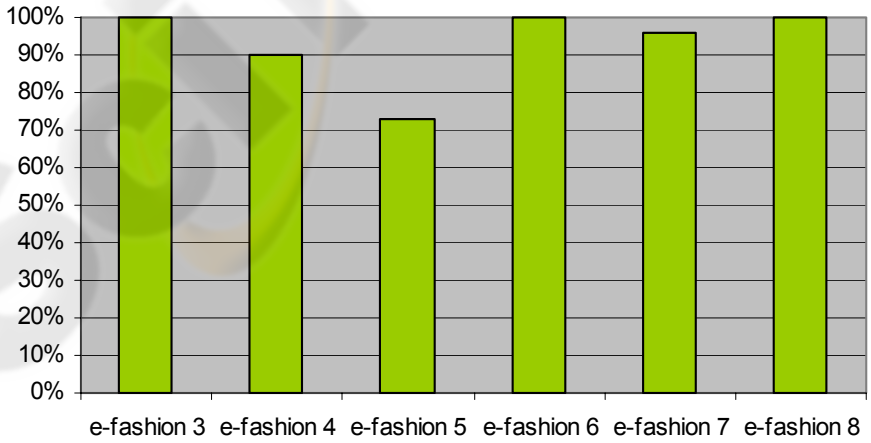

Figure 3: Task Completion Comparative

much better to design websites according to a quality model and using the ergonomic patterns identified.

3.2.5.4 Satisfaction comparative

Now we show some of the results obtained from the satisfaction test bringing face to face the different versions of e-fashion showing some comparative graphics.

We start with the affirmation *"Working with this site is satisfactory"*, because this question shows the general satisfaction of the users that have used the web application. In general, the satisfaction of the users was higher on e-fashion 3, because the most chosen option was number 5, which is the best of the different options. This shows that in general the satisfaction of the users that executed the tasks of the experiment was good.

However, on e-fashion 4 the most chosen option was number 2, which is near to the worst option (Totally disagree), which indicates that in general, on e-fashion 4 the satisfaction was the worst.

Another interesting question of the test was number 4: *"The website is very attractive for me",* because it shows the capacity of the website to attract users. In this case the results were also favourable to e-fashion 3.

In general the results of e-fashion 3 were positives, if we have into account the percentages of the different versions. E-fashion 4 again was the worst of the six versions.

The conclusions that we obtained about the rest of questions in the test were favourable to e-fashion 3 too. This indicates again that using patterns for designing websites is good to improve their final quality.

3.2.5.5 Learning comparative

In this case as we could see in the answers of question number 21 of the satisfaction test, in the notes taken by the evaluator while the experiment was carried out, and comparing the times obtained on the tasks where the users had to buy something (number 1 and number 4), the users proved to have understood the functionality of the website and have learned the basic use of it. Again this result was clearer on e-fashion 3, the version that was implemented using all the patterns defined for the criteria "Guide".

4 CONCLUSIONS AND FUTURE WORKS

Based on the interesting results obtained from the experiment, we can conclude that the use of ergonomic patterns to characterize quality criteria defined in a quality model produces better web applications with much more quality than others implemented without using the patterns.

Moreover, the realization of the experiment has been useful to evaluate the metrics used on it. The metrics are a useful mechanism for evaluating the usability of websites.

With the data obtained from the experiment we can conclude that the satisfaction metric is useful because it shows a good view about what users think about when they use the website on true conditions. However, we can also state that due to the subjective nature of this metric, and because we cannot assure that all the users say the truth when they fill a test, is important not to use only this metric to evaluate the quality of a website. In this sense, it is better to compare its results with the results obtained by using other metrics, as for instance, Task Time, Error Frequency and so on.

The validity of the experiment is clear in this case as the global results of the metrics used are

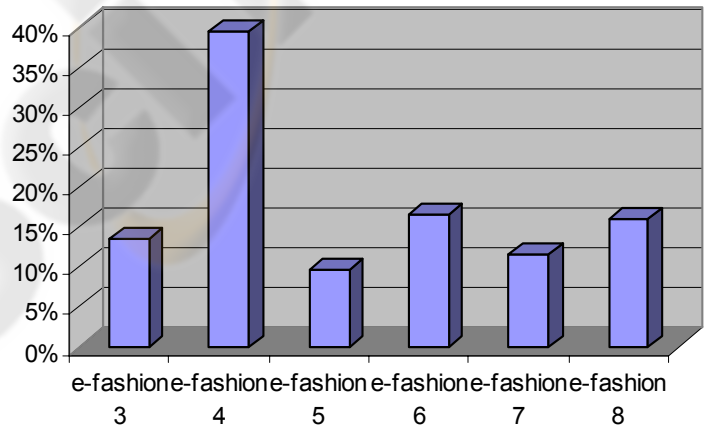

Figure 4: Errors Frequency Comparative.

consistent one with the others,

Another interesting conclusion is that *"Task Time"* and *"Error Frequency"* metrics are useful when we need to evaluate the quality of a website, because they show in an objective way the usability of the system when users execute tasks on it.

However, in the experiment we can see that it is very important when we want to evaluate the results of the metric *"Task Time"* the loading time of the page, because the loading time can give a false view about the time users spend on doing the tasks. Thus, it is very important that all the users can carry out the tasks on computers with the same technical specifications. About the "*learning*" metric we must say that it is also a good metric because it gives a general idea about the capacity of the website to be learned and used by the user.

So, applications easier to learn are more satisfactory for the user as we can conclude from the experiment results. The most important conclusion is that the use patterns to characterize the criteria defined in a quality model allow us to construct web applications with a higher degree of quality than others designed without any reference model nor patterns.

As a final remark, we can state that due to the complexity and difficulty of carrying out an experiment with real people it is much better to develop web quality models which allow developing websites with a basic level of usability. The design websites using a quality model and interaction patterns associated to the different criteria avoids making usability experiments because the final quality will be guaranteed.

REFERENCES

Brajnik, G (2002). Quality models based on automatic webtesting. Automatically evaluating usability of Web Sites, Minneapolis, CHI April 2002.

Constantine, L.L., Lockwood, L.A.D. (1999), Software for use, A practical guide to the models and methods of usage-centered design. ACM Press

ISO/IEC 9126-1. (2001) Software Engineering – Product Quality- Part 1: Quality Model, International Organization for Standardization, Geneva, 2001.

ISO/IEC 9126-4. (2001) Software Engineering – Software Product Quality- Part 4: Quality in use metrics, 2001.

Ivory, M.Y. (2001) An Empirical Foundation for Automated Web Interface Evaluation. Ph D. Thesis, Berkeley University, California.

Montero, F., Lozano, M., González, P., Ramos, I. (2002) Designing Websites by Using Patterns. Second Latin American conference on Pattern Languages of Programming. SugarLoafPLoP02. Itaipava. Rio de Janeiro. Brasil. ISBN: 85-87837-07-9. pp. 209-224.

Montero, F.; Lopez-Jaquero, V., Lozano, M.; González, P. (2003) A Quality Model For Testing the Usability of Web Sites. HCII´03.

Nielsen, J. (1993). Usability Engineering. Morgan Kaufmann,

Nielsen, J. (2001). Homepage Usability: 55 websites deconstructed. New Riders, pp. 315

Olsina, L. (1999) Metodología Cuantitativa para la Evaluación y Comparación de la Calidad de Sitios Web. PhD. Thesis, Universidad Nacional de La Plata, Argentina.

Perzel, K., Kane, D. (1999). Usability Patterns for Applications on the World Wide Web. PloP´99.

Shneiderman, B (1998). Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley Publishers.

Tidwell, J. (1999) Commond Ground: A pattern language for HCI design. http://www.mit.edu./~jtidwell/interaction_patterns.html

Van Duyne, D.K.; Landay, J. A.; Hong, J. I.; (2002) The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience, Publisher: Addison Wesley, ISBN: 0-201-72149-X

Welie, M. (2003) Interaction Design Patterns. http://www.welie.com/patterns.

[i] This work is partially supported by the Spanish CICYT TIN2004-08000-C03-01 and PBC-03-003 grants.