First International Symposium on Empirical Software Engineering and Measurement

Usability Evaluation Based on Web Design Perspectives

Tayana Conte[1], Jobson Massollar[1], Emilia Mendes[2], Guilherme H. Travassos[1]

[1]PESC-COPPE/UFRJ, Cx. Postal 68.511, CEP 21945-970, Rio de Janeiro, RJ, Brasil
{tayana, jobson, ght}@cos.ufrj.br
[2]Computer Science Department, The University of Auckland, Private Bag 92019, Auckland, New Zealand
emilia@cs.auckland.ac.nz

Abstract

Given the growth in the number and size of Web Applications worldwide, Web quality assurance, and more specifically Web usability have become key success factors. Therefore, this work proposes a usability evaluation technique based on the combination of Web design perspectives adapted from existing literature, and heuristics. This new technique is assessed using a controlled experiment aimed at measuring the efficiency and effectiveness of our technique, in comparison to Nielsen's heuristic evaluation. Results indicated that our technique was significantly more effective than and as efficient as Nielsen's heuristic evaluation.

1. Introduction

Web applications have quickly grown in their scope and extent of use, significantly affecting all aspects of our life [1]. Many industries nowadays (e.g. commerce, banking) all use Web-based applications to improve and increase their operations. The Web's characteristics – ubiquity, open standards, interlinking, easy access to information and services – have created new business opportunities, such as e-commerce or online auction systems [2], and the spread of Web applications into areas of communication and commerce makes it one of the leading and most important branches of the software industry [3].

As new business opportunities appear and the reliance on larger and more complex Web applications increases, so does the need to use better practices to develop applications that are delivered on time, within budget, with a high level of quality and easy to maintain [4]. To meet such requirements, a new research field has emerged: Web Engineering – the application of engineering principles to develop quality Web applications [5]. The main goal of Web Engineering is to develop correct Web applications, where structure, functionality, navigation and user interaction have to be properly represented [6].

Web development processes differ substantially from traditional software processes in many aspects [4], and to date there is no agreement as to how the core activities for a sound Web development process should be addressed [7]. In addition, there are no standardized Web development techniques or large data sets of Web projects that can be used to inform researches and practitioners [8]. Within the scope of Web development, the processes employed and the assessment of applications' quality are in general ad-hoc [9], based on common sense, intuition and developers' expertise [10]. There is a need for quality assurance activities tailored to Web development processes, and proposed using a scientific process that can support knowledge building.

The dominant Web development drivers comprise three quality criteria [3]:

- **Reliability**: Applications that work well do not crash, and do not provide incorrect data [4].
- **Usability**: A Web application with poor usability will quickly be replaced by another more usable as soon as its existence becomes known to the target audience [4]. We take usability as prescribed by the ISO 9241 standard: "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use".
- **Security**: Applications that handle customer data and other information safely, so that problems such as financial loss, legal consequences, and loss of credibility can be avoided [4].

In addition, several development methods proposed specifically for Web development have tried to demonstrate the need to capture and represent design perspectives, which are particularly relevant for this class of applications, especially those related to domain concepts, navigation and presentation issues [6]. For

0-7695-2886-4/07 $20.00 © 2007 IEEE
DOI 10.1109/ESEM.2007.30

example, according to Fraternalli and Paolini [11] the design of Web applications should consider three different design perspectives:

- Structural: defines the information organization handled by the application and its relationships.
- Navigational: represents how information can be accessed.
- Presentation: describes how information is presented to the user.

Other design perspectives, such as conceptual, presentation, and navigation, have been incorporated in several methods proposed in the literature [12]. However, it seems that there is no consensus regarding what perspectives should be used, which is not surprising given that there are several categories of Web applications (e.g. ubiquitous, transactional) [13].

Using the combination of design perspectives and heuristics, this paper proposes and validates, using a formal experiment, a technique to evaluate usability of Web applications. We propose that our technique is more efficient and effective than a conventional Heuristic Evaluation [14], where efficiency and effectiveness are defined as follows:

- **Efficiency:** the ratio between the number of detected defects and the time spent in the inspection process.
- **Effectiveness**: the ratio between the number of detected defects and the total number of existing (known) defects.

The remainder of this paper is structured as follows: Section 2 presents background information on Web usability inspections, and on Web design perspectives. Section 3 presents the Web usability inspection technique proposed and the results of a previous study used to improve the technique. In Section 4, the controlled experiment to evaluate the second version of the technique is discussed in detail, including goals and experimental design. In Section 5, the results are presented and discussed. Section 6 describes threats to the validity of our study. Finally, conclusions and comments on future work are given in Section 7.

2. Background

2.1 Web Usability Evaluation

In addition to the ISO 9241 standard, another international standard – ISO 9126 – also takes usability as one of the basic characteristics of software quality [15], and defines it as: “the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions”. Usability is then subdivided into five sub-characteristics, detailed in [15]: Understandability, Learnability, Operability, Attractiveness, and Compliance. Therefore, to evaluate the usability of a software product, one has to measure the value reached for the system on each of the relevant usability sub-characteristics.

Usability evaluation methods can be divided in two categories: (1) Usability Inspections - evaluation methods based on Experts’ Analyses; and (2) Evaluation Methods involving User Participation, such as: Laboratory studies, Think Aloud, Cooperative Evaluation, Protocol Analysis, Post-task Walkthroughs, Interviews and Questionnaires, and Physiological measurement [16].

When using Evaluation Methods involving User Participation, usability problems are discovered by means of observation and interaction with users, while they perform tasks or provide suggestions about the interface design and its usability. In Usability Inspection, problems are discovered by inspectors (experts) and their inspection techniques. Different usability inspection techniques have been developed and used. Between the main ones we have: Heuristic Evaluation [14], Cognitive Walkthrough [17], and Guidelines and Checklists [16].

The main disadvantages of usability inspections are their great subjectivity (different inspectors can produce different results) and its reliance on inspectors’ skills and experience. Conversely, its main advantage is to be less expensive than evaluation methods that involve user participation, since it does not need any special equipment or laboratory [18]. This advantage has increased the use of usability inspections, mainly in industrial environments [18].

Usability inspections represent the focus of this research work, since the usability of Web applications can greatly determine their success or failure [19].

2.2 Web Design Perspectives

To date different design perspectives have been proposed to drive Web development [12]. In addition, our own experience with industrial Web projects suggests that, for some Web applications, the following types of design perspectives should be used: conceptual, presentation, navigational and structural. The motivation to use these four types of design perspectives is as follows:

- The need to map the application’s domain to the presentation perspective, that is, to define views of the conceptual model take into account the presentation requirements of different user groups;
- The need to map the application’s domain to the navigation perspective, that is, to define responsibilities according to the requirements and rules defined in the navigation model;

- There are user requirements that cannot be mapped to or handled by the conceptual, presentation, navigational design perspectives, as they are more related to business rules than to domain concepts, having impact on the final architecture of the Web application.

Therefore, we consider the use of four Web design perspectives:

- **Conceptual**: represents the conceptual elements that make up the application domain.
- **Presentation**: represents the characteristics related to application layout and arrangement of interface elements.
- **Navigation**: represents the navigational space, defining the information access elements and their associations.
- **Structural**: represents the structural and architectural characteristics of the application, that is, how the application is structured in terms of components and their associations.

These design perspectives were evaluated by an observational study carried out in the second semester of 2005, using as subjects fifteen undergraduate students, organized in five teams with three developers each. Each team was responsible for the requirements elicitation, system analysis and design of a small Web application, in a specific domain, thus leading to 5 different applications of similar size, as follows:

a. A tool for e-Service Configuration and Allocation in Experimental Software Engineering Environments.
b. A tool for Build and Maintain Requirements Traceability Matrix in Software Projects.
c. A system to Manage Computers Configuration and Maintenance Cycle.
d. A system to Manage and Control Inventory.
e. Components and Services for e-Voting.

All teams were trained in the four design perspectives abovementioned and their respective representations using UML-based models. They could create model instances for each one of the four design perspectives, according to the demands of each application. Each team had three weeks to complete the task and, at the end of the process, had to provide the documents and models built during the development, and the follow up (feedback) questionnaire. Although all teams provided the documents and models, only ten subjects filled out the follow up questionnaire.

The evaluation of the design models allowed the identification of design perspectives relative to an application domain. We observed that, for regular Web applications, such as information systems (b-e), the conceptual, presentation and navigation perspectives were often used. However, for the non-conventional application, the e-science-based environment (a), there was a need to also use the structural perspective, besides the other three perspectives.

The analysis of the follow up questionnaires provided us with the following feedback:

- Four of ten subjects (40%) answered that the design perspectives made the development easier, while three subjects said that they neither helped nor hindered the development. Two subjects said they had difficulties in the use of the design perspectives. Finally, one subject considered dispensable the use of design perspectives for small Web applications.
- Seven of ten subjects (70%) said that all the suggested UML models could have been developed and applied to the project. One subject complained about the complexity of the UML navigation model. Two subjects argued against the use of some UML models with small Web application and one subject argued against the use of the UML presentation model to communicate between team members.

The results of the pilot study provided us useful feedback regarding the feasibility of using such design perspectives, and highlighted the need to: (1) understand what design perspectives should be used in the development of Web applications; and (2) improve aspects related to the representation (modeling) of design perspectives.

These results informed the proposal of a usability technique based upon Web Design Perspective, detailed in Section 3.

3. Web Design Perspectives-based Usability Evaluation Technique

Zhang et al. [20] suggests that it is difficult for an inspector to detect all types of problems at the same time, and propose a perspectives-based usability inspection technique called “Usability Based Reading” (UBR). This technique prescribes that each inspection session should focus on a subset of a usability issue. It defines three perspectives for usability: novice use, expert use and error handling.

The assumption behind perspective-based inspection techniques is that each inspection session can detect a greater percentage of defects in comparison to other techniques that do not use perspectives. In addition, the combination of different perspectives can detect more defects than the same number of inspection sessions using a general inspection technique [20].

Motivated by Zhang’s et al. [20] results, and assuming that design perspectives can be employed while building Web applications, we argue that the adoption of design perspectives could also improve the efficiency of Heuristic Evaluation (HEV) [14]. Basically, design perspectives are used to interpret Nielsen's heuristics in light of these perspectives. Therefore, the focus of each heuristic evaluation session evaluates a Web application's usability in relation to domain concepts, navigation, presentation and application structure. This new technique we call Web Design Perspectives-based Usability Evaluation (WDP), and its principles are as follows:

- Usability applied to Domain Concepts: relates to the clarity and concision of the problem domain's elements. Under this perspective, the usability is satisfactory if the domain terms have a representation easily understood for the different users, which does not let them make mistakes because of ambiguous, inconsistent or unknown terms.
- Usability applied to Presentation: relates to how consistent the information is presented to the user. Under this perspective, the usability is satisfactory if the arrangement of interface elements allows a user to accomplish his/her tasks effectively, efficiently and pleasantly.
- Usability applied to Navigation: relates to the different accesses to the application's functionalities. Under this perspective, the usability is satisfactory if the navigation options allow the user to accomplish his/her tasks effectively, efficiently and pleasantly.
- Usability applied to Structure: relates to how the application's structure is presented to the users. Application structure represents the associations between this application's components or modules. Under this perspective, the usability is satisfactory if the arrangement of the components or modules allows the user to accomplish his/her in an effective, efficient and pleasant way.

To evaluate the first version of the WDP technique (see Figure 1), a pilot study was carried out in the first semester of 2006, using twenty undergraduate students, attending a one semester undergraduate course in Human Computer Interaction at the Federal University of Rio de Janeiro (UFRJ). Our aim was to observe the feasibility (number of defects detected) and applicability (use of heuristics interpreted according to the design perspectives) of such inspection technique. We arranged the subjects into four teams of five inspectors each. We also measured their experience with Software Engineering using a three-point scale (Low, Medium and High). Two teams (A–Medium to High experience and B– Medium to Low experience) used the HEV technique while the other two (C–High experience and D-Low experience) used the WDP technique.

The objective of the usability inspection was to evaluate a real Web application called JEMS[1] (Journal and Event Management System). The JEMS application is used to support the process of conference creation, users' registration, submission, revision, acceptance and publishing of papers from conferences and journals.

Usability related to Navigation

Heuristic: User control and freedom

- Evaluate if the interface allows the user to navigate easily using different steps of a given task.
- Evaluate if the interface has undo or redo functions or similar functions that allow the user to exit in case of wrong choices.
- Evaluate if the interface allows the user to return to the main flow of a task after a detour or after performing a secondary task.

Heuristic: Flexibility and efficiency of use

- Evaluate if the interface provides different ways of accessing the main tasks.
- Evaluate if the interface provides accelerator keys or short-cuts when a user performs the main tasks.

Heuristic: Error prevention

- Evaluate if the interface prevents navigation errors, providing accessibility for the different types of users.
- …

Figure 1. Extract of WDP technique (version 1)

The context of the inspection comprised sixteen use cases involving three actors (Conference Chair, Reviewer and Paper Author). The teams had three weeks to complete the task and they were free to organize themselves to reach their goals, which were to inspect all scenarios related to the sixteen use cases and prepare a defect report. Four teams finished the inspections on time and provided their reports.

The evaluation of the defect reports provided useful feedback, detailed as follows: Figures 2 and 3 show the percentage of defects overlapping for each technique, where we noticed that:

- Teams A and B presented a significant difference in the number of reported defects despite their similar experience.
- Conversely, teams C and D do not present a significant difference in the number of reported

[1] https://submissoes.sbc.org.br

defects, even though their levels of experience differ largely.

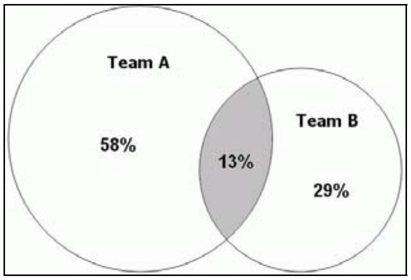

Figure 2. Overlapping defects for HEV technique

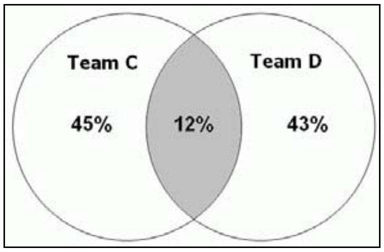

Figure 3. Overlapping defects for WDP technique

These results may suggest that the WDP technique is less sensitive to inspector's experience, when compared to HEV.

From this result and feedback obtained using the follow up questionnaires, it was possible to review the WDP technique, as follows:

- We reviewed the relationship between the heuristics and the design perspectives.
- We improved the heuristics' interpretations regarding each design perspective, that is, to make it clear what kind of defects that relationship relates to.
- We improved the training material.

Based on these results, a new version of the WDP technique was proposed and assessed using a controlled experiment (see Section 4).

Table 1 shows all the associations between heuristics [14] and the design perspectives in the WDP's version 2. Figure 4 shows an example of a heuristic and its corresponding interpretation according to the design perspectives.

Table 1. Existing relationships between heuristics and design perspectives in WDP version 2

Heuristic	Concep.	Present.	Navig.	Struct.
Visibility of system status	✓	✓		✓
Matching between system and real world	✓	✓		✓
User control and freedom			✓	
Consistency and standards	✓	✓		
Error prevention		✓	✓	
Recognition rather than recall		✓		✓
Flexibility and efficiency of use		✓	✓	
Aesthetic and minimalist project		✓		
Help users recognize, diagnose and recover from errors	✓	✓	✓	
Help and documentation	✓	✓	✓	

Consistency and Standards

Definition: The interface must adhere to conventions of the development platform and normally accepted interface standards. Users must not have to guess if different words, situations or actions all mean the same.

Conceptual perspective:

- Evaluate if the terms of the problem domain are presented to the user in a consistent way.

Presentation perspective:

- Evaluate if the terms, graphs and symbols of the interface are consistent.
- Evaluate if the interface adhere to the adopted standards for layout and controls.
- Evaluate if the interface is consistent for equivalent tasks.
- …

Figure 4. Extract of WDP technique (version 2)

4. The Experiment

4.1 Goal

The goal of this experiment, presented using the GQM paradigm [21], can be formalized as:

Analyze	Web Design Perspective-based Usability Evaluation Technique (WDP)
For the purpose of	understanding
With respect to	its effectiveness and efficiency, compared to the Heuristic Evaluation Technique (HEV)
From the point of view	of Web project inspectors

In the context of	the evaluation of the usability of a real Web application by undergraduate and postgraduate students.

4.2 Context Selection

Our experimental object was the same application used in the study described in Section 3 (JEMS). However, due to the large number of use cases, we only selected the use cases associated with the role 'Reviewer'. All the remaining tasks that needed to simulate a real conference (e.g conference creation, deadlines setup, papers submission) were carried out by the authors. This way, the following use cases comprised the inspection's context:

- Confirm participation in the conference
- User authentication
- Inform topics of interest
- Inform topics of conflict
- Retrieve paper for revision
- Submit revision

4.3 Hypotheses

The experiment was used to test the following hypotheses (null and corresponding alternative hypotheses are given):

- **H01**: There is no difference between the efficiency of techniques WDP and HEV.
- **HA1**: The efficiency of the WDP technique is greater than the efficiency of the HEV technique.
- **H02**: There is no difference between the effectiveness of techniques WDP and HEV.
- **HA2**: The effectiveness of the WDP technique is greater than the effectiveness of the HEV technique.

4.4 Variables Selection

The independent variable was the usability evaluation technique and the dependent variables were the efficiency and effectiveness of a technique. Efficiency and effectiveness were calculated for each subject as the ratio between the number of defects detected and the time spent in the inspection process, and the ratio between the number of detected defects and the total number of existing (known) defects, respectively.

4.5 Subjects

Subjects were chosen by convenience from three different courses (2nd Semester/2006) offered at UFRJ: the first course contributed three MSc students, the second three undergraduate students and the third eight MSc and PhD students. All fourteen subjects signed a consent form and filled out a characterization form that measured their expertise with software inspections, software usability and software engineering in general. The characterization data was used to organize subjects into three categories (low, medium or high experience).

The three MSc students attending the first course had high expertise in software inspections and were familiar with our research project. Thus, to minimize bias, these three subjects comprised the control group (see Section 4.6). The remaining eleven subjects had their experience measured using an ordinal scale, from 1 (low experience) to 3 (high experience). Three subjects (27.27%) had high experience in software engineering (more than five projects in industry); another three (27.27%) had medium experience (at least one software project in industry) and five (45.45%) had low experience (no industry experience in software engineering). Table 4 shows each subject's characterization.

4.6 Experimental Design

The experiment was planned as a blocked subject-object study [22], and the fourteen subjects were organized into three groups, as follows:

- **Group 1**: represents the control group and includes three subjects (having high expertise in software inspections) who used an *ad-hoc* inspection technique.
- **Group 2**: six subjects who used the HEV technique.
- **Group 3**: five subjects who used the WDP technique.

Group 1 (control group) was created due to the following factors:

- The experimental object was a real Web application, so it was not possible to anticipate all the possible defects one could find (defects oracle), and also if the defects were real or false positives. This way, the list of defects encountered by the control group was used to check the defects encountered by Groups 2 and 3.
- Due to the lack of a defects oracle, Group 1 took part in an inspection meeting where the lists of defects from Groups 2 and 3 were analyzed and the false-positives were identified. The authors attended this meeting as moderators, i.e., only the control group decided whether a defect was genuine or a false positive. Group 1 was also responsible to identify replicated defects.

Groups 2 and 3 were organized according to subjects' experience levels, where experience levels were balanced to minimize the effect of confounding factors (see Table 4). Each subject carried out the usability evaluation individually.

4.7 Instrumentation

Several artifacts were defined to support the experiment: characterization and consent forms, specification of HEV and WDP techniques including examples of their use, JEMS specification describing the use cases that would be executed during inspection, worksheets to support defects discrimination and inspection meeting, follow up questionnaire and setup of the JEMS system (conference creation, user registration, paper submission, paper allocation to reviewers). All artifacts were validated by the authors.

4.8 Preparation

The training package was organized into four modules:

- **JEMS**: training about the JEMS application (30 minutes). It details the use cases relevant to the role Reviewer.
- **UI**: training about the general goals of Usability Inspections, without any details about any specific technique (45 minutes).
- **HEV**: training about Heuristic Evaluation technique (90 minutes).
- **WDP**: training about Web Design Perspectives Evaluation technique (90 minutes).

Table 2 describes which training was given to which group.

Table 2. Groups and trainings

	JEMS	UI	HEV	WDP
Group 1	✓	✓		
Group 2	✓	✓	✓	
Group 3	✓	✓		✓

4.9 Execution

All subjects received the instruments at the same time and the JEMS application was made available on the Web as well. Subjects had 4 days to perform the evaluation. During the inspection, they had to fill out a worksheet with the defects encountered. At the end of the evaluation process, all subjects returned the defects worksheet with the total time spent in the inspection and the follow up questionnaire filled out.

4.10 Data Validation

All defects' worksheets were checked for incorrect and missing information, and one worksheet (Group 2) was discarded because the subject did not accomplish the work on time. No outlier was eliminated.

5. Results

At the end of the experiment, a meeting attended by the control group and two authors took place. A list of all discrepancies each group found (summarized in Table 3), without any information about which technique was used, was analyzed resulting in one single aggregated list of real defects.

Table 3. Discrepancies and detection time per subject group

Group	#Discrepancies	Total Time (min)
1 (ad-hoc)	43	187
2 (HEV)	53	696
3 (WDP)	86	1080

The results of this meeting were as follows:

- A list classifying discrepancies, real defects and false positives.
- A list of replicated defects, that is, equivalent defects detected by different inspectors.
- A list with the known defects, totalizing 90 defects, and 48 false positives. The control group found 36 defects.

5.1 Quantitative Analysis

The statistical analysis presented in this Section was carried out using the statistical tool SPSS V 12.1.0, and $\alpha = 0.10$. This choice of statistical significance was motivated by the small sample size used in this experiment [23].

Table 4 shows the overall results of the usability evaluation. Column #Defects represents the defects found by each inspector (also includes replicates).

Table 4. Summary of inspection results per subject

Group	Subj.	Exp.	Time (min)	Discrep.	# False Positive	# Def.
2	1	H	116	22	1	21
	2	L	40	6	2	4
	3	M	190	9	1	8
	4	H	260	23	5	18
	5	L	90	13	2	11
3	6	M	110	24	4	20
	7	L	170	22	4	18
	8	H	295	22	3	19
	9	M	300	41	16	25
	10	L	205	24	3	21

Table 4 suggests that the number of defects found per inspector using the WDP technique is very similar, independently of subject's experience. This trend is supported by the small standard deviation of Defects per Subject (see Table 5). The same behavior was not observed for the HEV technique, also supported by its larger standard deviation. These results suggest that the WDP technique enabled inspectors with diverse inspecting skills to find a similar number of defects, thus reducing reliance upon inspectors' skills, one of the problems of usability inspections pointed out by Garzotto et al. [18] (see Section 2).

Table 5. Techniques Characteristics

		Technique	
		HEV	WDP
Time	Average	139.2	216.0
	Std. Dev.	86.56	81.81
	% Std. Dev.	62.2%	37.9%
Defects per Subject	Average.	12.40	20.60
	Std. Dev.	7.02	2.70
	% Std Dev.	56.6%	13.1%
Defects/Hour per Subject	Average	6.18	6.45
	Std. Dev.	3.19	2.68
	% Std Dev.	51.7%	41.6%

We used the non-parametric Chi-Square test to check if experience levels differed between Groups 2 and 3, and results showed that they did not (p = 0.819).

H1: Efficiency of Techniques WDP and HEV

The boxplots with the distribution of efficiency per subject, per technique (see Figure 5) show that Group 3 (EfficiencyWDP) was relatively more efficient than Group 2 (EfficiencyHEV) to inspect the usability of the JEMS application: Group 3's median is slightly higher than Group 2's. However, when we compared the two samples using the Mann-Whitney test, a non-parametric test, we found no significant differences between the two groups (p = 0.754). These results suggest that despite the trends see using boxplots, both techniques provided similar efficiency when used to inspect the JEMS application.

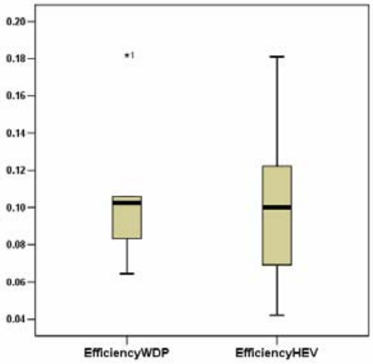

Figure 5. Boxplots for efficiency per subject per technique

We also used the One-Way ANOVA test to check if there was a significant relationship between technique and efficiency, however none was found (p = 0.885).

These results support the null hypothesis H01 that there are no differences in efficiency between techniques WDP and HEV, and in consequence, reject HA1.

H2: Effectiveness of Techniques WDP and HEV

The boxplots with the distribution of effectiveness per subject per technique (see Figure 6) show that Group 3 (EffectivenessWDP) was much more effective than Group 2 (EffectivenessHEV) when inspecting the usability of the JEMS application: Group 3's median is much higher than Group 2's, and all of Group 3's boxplot is above Group 2's third quartile. The Mann-Whitney test confirmed that WDP's effectiveness was significantly higher than HEV's effectiveness (p = 0.095), thus supporting the alternative hypothesis HA2, and, conversely, rejecting the null hypothesis H02. These results suggest that the technique WDP was more effective than HEV when used to inspect the JEMS application.

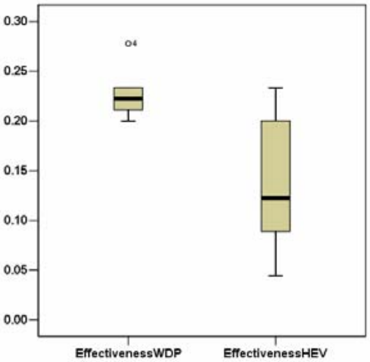

Figure 6. Boxplots for number of defects found per subject per technique

Herein we also used the One-Way ANOVA test to check if there was a significant relationship between technique and effectiveness, and this test confirmed that the relationship between technique and effectiveness was statistically significant (p = 0.041).

We also computed two indicators – an effectiveness indicator (Defects ratio), computed as the average number of defects found per group (Avg. Defects) divided by the total number of known defects (90); and an efficiency indicator (Avg. Defects/Hour) , computed as the average number of defects found per group divided by the amount of time spend per group (see Table 6). Results show that both techniques were not very effective at finding the total number of defects; however the WDP technique was nearly twice as effective as the HEV technique; however their efficiency was similar.

Table 6. Effectiveness and efficiency per technique

Tech.	Avg. Time (min)	Avg. Defects/Hour	Avg. Defects	Defects ratio %
HEV	139.2	5.34	12.4	13.78
WDP	216	5.72	20.6	22.89

5.2 Qualitative Analysis

The follow-up questionnaires from Group 3 provided useful feedback about the use of the WDP technique. The follow-up questionnaire contained six open-ended questions about:

- WDP's understandability and operability.
- Heuristics description.
- Design perspectives description.
- Help in finding discrepancies.
- Need for additional comments about how to apply a particular pair of heuristic/design perspective; and
- Suggestions for improvement.

All subjects, except one, considered the WDP technique adequate for use. One subject complained about the difficulty of use, because of the large number of heuristic/design perspective pairs.

None asked for a more detailed heuristic description. One argued for a better description of the conceptual perspectives and four complained about the description of the structural perspective. They also reported difficulty to apply the structural perspective during the usability inspection.

All subjects said that the WDP technique helped find discrepancies. Two subjects added that the WDP technique made the classification of discrepancies easier to do. And one said that the table with the relationships between heuristics and design perspectives (Table 1) made the inspection more agile.

Related to the need for additional comments on how to apply a particular pair of heuristic/design perspective, one asked for additional comments in all pairs, including the structural perspective, and another said that the pair "Error Prevention/ Navigation" needed a more detailed comment.

Finally, there were many suggestions to improve WDP's use: to create a "quick reference" and replace the comments for questions in each heuristic/ design perspective pair.

6. Threats to Validity

As in all studies, there were various threats that could affect the validity of our results. In this Section, we discuss those threats; break them down into the following categories: internal, external, conclusion and construct.

Internal validity: in our experiment, we considered three main threats that represent a risk for an improper interpretation of the results: training effects, experience classification and time measurement. There might be a training effect if the training about HEV technique had worst quality than the training about WDP. We controlled training effect by preparing equivalent training courses with the same examples of discrepancies detection applying technique HEV or WDP. Related to experience classification, it was based on self classification. We tried to mitigate this risk comparing subjects' self-selected classification with our own observation during the courses. Finally, considering time measurement, we asked subjects to be as precise as possible, but there is no warranty that the time reported was accurate.

External validity: even though experiments using students as subjects and running on an academic environment are valuable as a way to characterize or understand aspects of new proposals, two issues must be taken into consideration: (1) probably students are not good substitutes for real world inspectors; and (2) usually, academic environments are far from simulating the existing conditions in industrial development environments. However, JEMS represents a real application and students may to some extent present similar inspection skills to less experienced practitioners, thus our results may still be suitable to industry and scale up. In addition, Carver et al. [24] point out a series of benefits that researchers can gain from empirical studies using students.

Conclusion validity: the small number of data points is not ideal from the statistical point of view, however small sample sizes are a known problem difficult to overcome.

Construct validity: We measured efficiency and effectiveness using the same approach proposed in [25]. In addition, these two measures are often used in studies that investigate defect detection techniques [25], which is also our case.

7. Conclusions

In this paper we motivate, propose and validate a new usability evaluation technique based on Web design perspectives (WDP) and heuristics. We used a formal experiment to compare the efficiency and effectiveness between the WDP and HEV techniques.

Results showed that the WDP technique was significantly more effective than the HEV technique, with similar efficiency.

An important lesson learned regarding experiment design was the creation of the control group and the introduction of an inspection meeting after the defects detection phase. This arrangement allowed provided greater reliability on the list of known defects.

Further research regarding usability evaluation should include: (1) a replication of the formal experiment reusing the same statistical models presented on Section 5.1, and using a greater number of subjects,; (2) improvement of the technique based on a detailed analysis of the influence of each pair heuristic/perspective in the final list of detected defects; (3) to investigate new arrangements for inspector/perspective relationship, for example, to investigate if efficiency and effectiveness can be improved if each inspector focuses on just one perspective, and (4) to replicate the experiment in an industrial development environment.

We also plan to conduct further research to investigate the use of design perspectives in the inspection of others design-related artifacts.

8. Acknowledgements

We thank Marcos Kalinowski for his help with the first study using the WDP technique, and the undergraduate and postgraduate students for their participation in the experiment and pilot studies. This work was supported by CNPq (Experimental Software Engineering Project – 472135/2004-0) and FAPEAM. It also represents a result from the ESEWEB Project (http://lens.cos.ufrj.br:8080/ESEWEB)

9. References

[1] Ginige, A., Murugesan, S., "Web Engineering: an Introduction", IEEE Multimedia, 8(1), Jan – Mar. 2001.

[2] Gómez, J., Cachero, C., Pastor, O., "Conceptual Modeling of Device-Independent Web Applications", IEEE Multimedia, 8(2), April – June 2001.

[3] Offutt, J., "Quality attributes of Web software applications", IEEE Software, 19(2), March/April 2002.

[4] Mendes, E., Mosley, N., Counsell, S., "The Need for Web Engineering: An Introduction", In "Web Engineering" (Eds: Emilia Mendes and Nile Mosley), Springer, 2006.

[5] Pressman, R., "What a tangled Web we have", IEEE Software, Jan – Feb. 2000.

[6] Pastor, O. (2004), "Fitting the Pieces of the Web Engineering Puzzle" - Invited Talk, XVIII Simpósio Brasileiro de Engenharia de Software (SBES), Brasília, October- 2004.

[7] Gómez, J., Cachero, C., "OO-H Method: Extending UML to Model Web Interfaces", "Information Modeling for Internet Applications". Idea Group Publishing. 2003.

[8] Mendes, E., Mosley, N., Counsell, S. (2004), "Investigating Web size metrics for early Web cost estimation", Journal of Systems and Software, 77(2), August 2005.

[9] Standing, C., "Methodologies for developing Web applications", Information and Software Technology, 44(3), 2002.

[10] Abrahão, S., Condori-Fernández, N., Olsina, L., Pastor, O., "Defining and Validating Metrics for Navigational Models", Proceeding of the Ninth International Software Metrics Symposium (METRICS'03), 2003.

[11] Fraternalli, P., Paolini, P., "A Conceptual Model and a Tool Environment for Developing More Scalable, Dynamic, and Customizable Web Applications", EDBT 98, 1998.

[12] Conte T., Mendes, E., Travassos, G., H., " Development Processes for Web Applications: A Systematic Review", in Proceedings of XI Simpósio Brasileiro de Multimídia e Web - WebMedia, Poços de Caldas, Brasil, 2005. (In Portuguese)

[13] Kappel, G., Michlmayr, E., Pröll, B., Reich, S., Retschitzegger, W., "Web Engineering - Old wine in new bottles?", Web Engineering, Springer LNCS 3140, 2004.

[14] Nielsen, J., "Heuristic evaluation". In "Usability Inspection Methods" (Eds. Nielsen, J., and Mack, R.L.), John Wiley & Sons, New York, NY, 1994.

[15] ISO, "ISO/IEC FDIS 9126-1: Software Engineering - Product quality – Parts 1-4", 2000.

[16] Dix, A., Finlay, J., Abowd, G., Beale, R., "Human-Computer Interaction", Pearson/ Prentice Hall, Third Edition, 2004.

[17] Polson, P., Lewis, C., Rieman, J., Wharton, C., "Cognitive Walkthroughs: a method for theory-based evaluation of users interfaces", International Journal of Man-Machine Studies, 36, 1992.

[18] Garzotto F., Matera M., Paolini P., "Abstract Tasks: a tool for the inspection of Web sites and off-line hypermedia" Proceedings of the tenth ACM Conference on Hypertext and hypermedia, 1999.

[19] Matera, M., Rizzo, F., Carughi, G., "Web Usability: Principles and Evaluation Methods", In "Web Engineering" (Eds: Emilia Mendes and Nile Mosley), Springer, 2006.

[20] Zhang, Z., Basili, V. and Shneiderman, B. "Perspective-based Usability Inspection: An Empirical Validation of Efficacy", Empirical Software Engineering: An International Journal, vol. 4(1), March 1999.

[21] Basili, V., Rombach, H., "The TAME Project: Towards Improvement-Oriented Software Environments", IEEE Transactions on Software Engineering, 14, 1988.

[23] Dyba, T.; Kampenes, V.; Sjoberg, D. A Systematic Review of Statistical Power in Software Engineering Experiments. Information and Software Technology. Elsevier, 2005.

[22] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., and Wesslén, A., "Experimentation in Software Engineering - An Introduction", Kluwer Academic Press, 2000.

[24] Carver, J., Jaccheri, L., Morasca, S., and Shull, F., "Issues in using students in empirical studies in software engineering education", Proceedings Ninth International Software Metrics Symposium, 3-5 Sept. 2003.

[25] Denger, C., Kolb, R., "Testing and Inspecting Reusable Product Line Components: First Empirical Results", Proceedings 5th International Software Metrics Symposium, 2006.