Towards to the Validation of a Usability Evaluation Method for Model-Driven Web Development

Adrian Fernandez
ISSI Research Group
Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain.
+34 96 387 73 50
afernandez@dsic.upv.es

Silvia Abrahão
ISSI Research Group
Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain.
+34 96 387 73 50
sabrahao@dsic.upv.es

Emilio Insfran
ISSI Research Group
Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain.
+34 96 387 73 50
einsfran@dsic.upv.es

ABSTRACT
The challenge of developing more usable Web applications has promoted the emergence of several usability evaluation methods. However, there is a lack of empirically validated methods that can properly be integrated during the early stages of Web development processes. This has motivated us to propose a Web Usability Evaluation Method (WUEP) which can be integrated into model-driven Web development processes. This paper presents the first steps in the empirical validation of WUEP through a controlled experiment. This experiment was designed in order to evaluate the effectiveness, efficiency, perceived ease of use, and satisfaction with WUEP in comparison to a widely-used inspection method: Heuristic evaluation (HE). Results show that WUEP is more effective and efficient than HE in the detection of usability problems in artifacts obtained from a model-driven Web development process. The evaluators were also satisfied when applying WUEP, and found it easier to use than HE.

Categories and Subject Descriptors
D.2.4 [Software Engineering]: Software/Program Verification - Validation; D.2.9 [Software Engineering]: Management - Software quality assurance.

General Terms

Keywords
Empirical Software engineering, Web usability evaluation, Model-driven Web development.

1. MOTIVATION
Despite the large number of usability evaluation methods with which to address Web usability issues, most of these approaches only consider usability evaluations during the final stages of the Web development process. Works such as that of Juristo et al. [10] claim that the usability evaluation should also be performed during the early stages of the Web development process in order to improve user experience and decrease maintenance costs.

This is in line with the results of a systematic review which was performed to investigate which usability evaluation methods have been employed to evaluate Web artifacts and how they were employed [8]. This study revealed findings such as a lack of usability evaluation methods for the Web that have been empirically validated and can be properly integrated into the early stages of the Web development processes. In an earlier work [5], these results motivated our proposal for the Web Usability Evaluation Process (WUEP), which can be instantiated and integrated into Model-Driven Web Development processes (MDWD). In MDWD processes, intermediate artifacts (models), which represent different views of a Web application, are used in all the steps of the development process, and the final source code is automatically generated. In this context, evaluations of these models can provide early usability evaluation reports to identify problems that can be corrected prior to the generation of the source code.

Several empirical studies related to the assessment of usability evaluation methods for the Web exist (e.g., [4]). However, they are focused on traditional Web development processes. There are few empirical studies based on the MDWD processes (e.g., [1][13]). In our view, empirical studies are needed to evaluate and improve any evaluation method proposed. In addition, these studies will provide useful information when the method is compared to others. The aim of this paper is in line with this statement and presents the first results of an empirical validation of WUEP through a controlled experiment.

This paper is structured as follows. Section 2 shows the methods evaluated. Section 3 describes the controlled experiment. Section 4 shows the analysis of the results obtained. Section 5 discusses threats to the validity of the experiment, and Section 6 presents our conclusions and further work.

2. METHODS EVALUATED
The methods to be evaluated were two inspection methods: our proposal (WUEP) and the Heuristic Evaluation (HE) [12] proposed by Nielsen. Inspection methods are performed by evaluators to evaluate artifacts (normally user interfaces) with regard to certain principles in order to detect usability problems. These methods are commonly employed since they can be applied in several stages of a Web development process and not only when the Web application has been completed.

The Web Usability Evaluation Process (WUEP) extends and adapts the quality evaluation process proposed in the ISO 25000 (SQuaRE) [9] with the purpose of integrating usability evaluations.
into MDWD processes. WUEP employs a Web usability model that decomposes the usability concept into sub-characteristic and measurable attributes. Metrics with a generic definition are associated with these attributes in order for them to be operationalized at different abstraction levels (Platform-Independent Models, Platform-Specific Models and final User Interfaces) in any model-driven Web development process. The aim of applying metrics was to reduce the subjectivity inherent to existing inspection methods. There are three roles involved in WUEP: evaluation designer, evaluation executor, and Web developer. The evaluation designer performs the establishment of evaluation requirements (scope, Web application selection, usability attributes selection, Web artifacts selection, etc.), the specification of the evaluation (operationalization of metrics, rating levels for metrics, etc.), and the design of the evaluation (number of evaluators, evaluation plan, etc.). The evaluation executor applies the evaluation plan designed in the execution stage (calculation of metrics, usability problem reports), and finally, the Web developer performs the analysis of changes in order to correct the usability problems.

The Heuristic Evaluation (HE) requires a group of evaluators to examine the user interface in compliance with recognized usability principles called heuristics. HE proposes ten heuristics that are intended to cover the best practices in the design of any user interface. (e.g., minimize the user workload, error prevention, etc.). There are two roles involved in HE: evaluation designer and evaluation executor. The evaluation designer determines the scope of the evaluation and makes the evaluation plan. The evaluation executor applies the heuristics to Web artifacts in order to identify any usability problems. HE was selected because it is widely-used in industrial domains since it can also be applied in intermediate artifacts produced during the early stages of the Web development process (e.g., mock-ups). It is important to note that there is no other method based on the MDWD process with which to compare WUEP.

3. CONTROLLED EXPERIMENT

The controlled experiment was designed by considering the guidelines proposed by Wohlin et al. [14]. According to the Goal-Question-Metric (GQM) [2], the goal of the experiment is: to analyze WUEP for the purpose of evaluating it with regard to its effectiveness, efficiency, perceived ease of use, and the evaluators’ perceived satisfaction of it in comparison to HE from the viewpoint of a set of usability evaluators.

The context of the experiment is the usability evaluation of a real Web application performed by research associates. This context is determined by the Web application to be evaluated, the usability evaluation methods to be applied and the subject selection. The Web application selected is an intranet employed for the management of Web projects in a company. It was developed by a Web development company through an MDWD method called the Object-Oriented Hypermedia method (OO-H) [6]. Two different functionalities, which are similar in complexity, were selected for the composition of the experimental objects (O1 and O2). Each experimental object contains three artifacts related to one functionality: a navigational model, an abstract user interface model, and a final user interface. The methods to be evaluated were WUEP and HE, and only their execution stages were considered. The Subjects were chosen for convenience from a group of research associates with no previous experience in usability evaluation, from the Department of Information Systems and Computation of the Universidad Politécnica de Valencia. This selection was made up of 12 PhD students all of whom have a degree in Computer Science Engineering and whose research activities are performed in the Software Engineering field. This number of evaluators was selected by considering a recent study that claims that 10±2 evaluators are needed to perform a usability evaluation to find around 80% of the usability problems [7].

The method has been applied to two independent variables: the evaluation method (WUEP and HE) and the experimental objects (O1 and O2). There are two objective dependent variables: effectiveness, which is calculated as the ratio between the number of usability problems detected and the total number of existing (known) usability problems; and efficiency, which is calculated as the ratio between the number of usability problems detected and the total time spent on the inspection process. There are also two subjective dependent variables: perceived ease of use and evaluators’ perceived satisfaction. Both are calculated by closed-questions from a five-point Likert-scale questionnaire which also includes open-questions to obtain feedback from the evaluators.

The hypotheses of the experiment are:

- **H1**: There is no difference between the effectiveness of WUEP and HE /H1/: The effectiveness of WUEP is different from the effectiveness of HE.
- **H2**: There is no difference between the efficiency of WUEP and HE /H2/: The efficiency of WUEP is different from the efficiency of HE.
- **H3**: There is no difference between the perceived ease of use of WUEP and HE /H3/: The perceived ease of use of WUEP is different from the perceived ease of use of HE.
- **H4**: There is no difference between the evaluators’ perceived satisfaction of employing WUEP and HE /H4/: The evaluators’ perceived satisfaction of employing WUEP is different from the evaluators’ perceived satisfaction of employing HE.

The experiment is planned as a within subjects design, signifying that the same subjects use both methods in a different order and with different experimental objects (the subjects’ assignment to the tasks was random). Table 1 shows the schedule of the experiment in more detail. In addition, before the controlled experiment, a control group was created in order to provide an initial list of usability problems by applying an ad-hoc inspection method, and to determine whether the usability problems reported by the subjects were real or false positives. This group was formed of two independent evaluators who are experts in usability evaluations, and one of the authors of this paper.

<table>
<thead>
<tr>
<th>Table 1. Schedule of the controlled experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Day</td>
</tr>
<tr>
<td>1D Sub. (15+20 m)</td>
</tr>
<tr>
<td>Training</td>
</tr>
<tr>
<td>1st Session (90 min)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Break (180 min)</td>
</tr>
<tr>
<td>Training (20 min)</td>
</tr>
<tr>
<td>2nd Session (90 min)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Several documents were designed as instrumentation for the experiment: slides for training session, appendices with an explanation of the method, documents for gathering data (4 docs: combinations of WUEP/HE with O1/O2), and two questionnaires.

4. ANALYSIS OF RESULTS

After the execution of the experiment, the control group analyzed all the usability problems detected by the subjects. If a usability problem was not in the initial list, this group determined whether it could be considered as a real usability problem or a false positive. Replicated problems were considered only once. Discrepancies in this analysis were solved by consensus. The control group determined a total of 13 and 14 usability problems in the experimental objects O1 and O2, respectively.

The quantitative analysis was performed by using the SPSS v16 statistical tool and $\alpha=0.05$. Table 2 summarizes the overall results of the usability evaluations.

**Table 2. Overall results**

<table>
<thead>
<tr>
<th>Problems/Subject</th>
<th>False positives/Subject</th>
<th>Replicated Prob./Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{x}$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>HE</td>
<td>4.25</td>
<td>1.4</td>
</tr>
<tr>
<td>WUEP</td>
<td>7</td>
<td>2.21</td>
</tr>
</tbody>
</table>

The results obtained have allowed us to interpret that WUEP did not provide false positives and replicated problems. The lack of false positives can be explained by the fact that WUEP tends to minimize the subjectivity of the evaluation. The lack of replicated problems can be explained by the fact that WUEP provides operationalized metrics that are classified to be applied in one type of artifact.

The SPSS guide recommends the Shapiro-Wilk test to verify whether the data is normally distributed when the sample size is smaller than 50. Our aim was to select which tests are needed in order to verify our hypotheses. Table 3 shows the results of the normality test, in which $^*$ signifies that this variable is not normally distributed in this evaluation method.

**Table 3. Shapiro-Wilk Normality test results**

<table>
<thead>
<tr>
<th></th>
<th>Effec.</th>
<th>Effic.</th>
<th>PEU</th>
<th>PSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>0.761</td>
<td>0.045 $^*$ (&lt; 0.05)</td>
<td>0.425</td>
<td>0.118</td>
</tr>
<tr>
<td>WUEP</td>
<td>0.029 $^*$ (&lt; 0.05)</td>
<td>0.805</td>
<td>0.573</td>
<td>0.158</td>
</tr>
</tbody>
</table>

The boxplots with the distribution of each dependent variable per subject per method (see Figure 1) show that WUEP was more effective and efficient than HE, and WUEP was also perceived by the evaluators as being easier use and more satisfactory than HE. In order to determine whether or not these results were significant, we applied: the Mann-Whitney non-parametric test to verify H1 and H2 (since WUEP_Effec. and HE_Effic. are not normally distributed), and the One-Way ANOVA parametric test to verify H3 and H4.

![Figure 1. Boxplots for each dependent variable](image)

The $p$-values obtained from the Mann-Whitney test for the Effec. and Effic. variables were 0.003 and 0.000 respectively. The $p$-values obtained from the One-Way ANOVA test for the PEU and PSU variables were 0.006 and 0.000, respectively. These results therefore support the rejection of all the null-hypotheses and the acceptance of their respective alternative-hypotheses ($p < 0.05$).

In order to strengthen our analysis, we used the method suggested in [3] to test the effect of the order of methods and the order of experimental objects (both independent variables). We used the Diff function: $\text{Diff} = \text{observation}(A) - \text{observation}(B)$, where $x$ denotes a particular subject, and $A_B$ are the two possible values of one independent variable. We created Diff variables from each dependent variable (e.g., Effec Diff(WUEP) represents the difference in effectiveness of the subjects who used WUEP first and HE second. On the other hand, Effec Diff(HE) represents the difference in effectiveness of the subjects who used HE first and WUEP second). The aim was to verify that there were no significant differences between Diff functions since that would signify that there was no influence in the order of the independent variables. The Shapiro-Wilk test showed that all the Diff functions were normally distributed. We therefore applied the parametric $t$-test in order to verify whether the effects were significant. Table 4 shows that all the $p$-values obtained were $> 0.05$. We can therefore conclude that there was no effect with regard to the order of methods and experimental objects for any dependent variable.

**Table 4. $t$-test results for Diff functions**

<table>
<thead>
<tr>
<th>Order of Experimental Objects</th>
<th>Effec.</th>
<th>Effic.</th>
<th>PEU</th>
<th>PSU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>0.161</td>
<td>0.846</td>
<td>0.871</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td>0.394</td>
<td>0.910</td>
<td>0.908</td>
<td>0.514</td>
</tr>
</tbody>
</table>

Finally, a qualitative analysis was performed by analyzing the open-questions that were included in the questionnaire. This analysis revealed several important issues which can be considered to improve WUEP. For example, the evaluators suggested that WUEP might be more useful if the evaluation process were automated or computer-aided (particularly the calculation of certain metrics). They also detected that providing more examples of how to apply metrics might improve the application of the method.

5. THREATS TO VALIDITY

The main threats to the internal validity of the experiment are: learning effect, evaluation design, subject experience, information exchange among evaluators, and understandability of the documents. The learning effect was alleviated by ensuring that each subject applied each method to different experimental objects, and all the possible order combinations were considered. The evaluation design might have affected the results owing to the...
The main threats to the external validity of the experiment are: representativeness of the results, and duration of the experiment. Despite the fact that the experiment was performed in an academic context, the results could be representative with regard to evaluators with no experience in usability evaluations. However, the previous selection of attributes with their operationalized metrics and the selection of the Web application might have affected the representativeness. In order to alleviate these issues, we intend to evaluate more Web applications, and to carry out surveys to determine guidance for the evaluator designers in order to select attributes by considering the Web application family. Since the duration of the experiment was limited to 90 minutes, only three representative artifacts were selected from the different types of artifacts available.

The main threats to the construct validity of the experiment are: measures that are applied in the quantitative analysis and the reliability of the questionnaire. Measures that are commonly employed in this kind of experiment were applied in the quantitative analysis [4]. The reliability of the questionnaire was tested by applying the Cronbach test. Questions related to perceived ease of use obtained a Cronbach’s alpha of 0.91, and questions related to evaluator satisfaction of use obtained a Cronbach’s alpha of 0.80. These values are higher than the acceptable minimum (0.70) [11].

The main threat to the conclusion validity of the experiment is the validity of the statistical tests applied. This was alleviated by applying the most common tests that are employed in the empirical software engineering field [11]. However, more replications are needed in order to confirm these results.

6. CONCLUSIONS AND FUTURE WORK
This paper presents a controlled experiment that represents the first steps in the empirical validation of our proposal: the Web Usability Evaluation Process. This experiment was designed in order to compare the effectiveness, efficiency, perceived ease of use and satisfaction of our proposal with a widely-used inspection method: Heuristic Evaluation (HE). The results show that WUEP was more effective and efficient than HE in the detection of usability problems in artifacts obtained from a model-driven web development process. In addition, the evaluators were satisfied when they applied WUEP, and found it easier to use than HE.

Although the experimental results provided good results as regards the usefulness of WUEP as a usability inspection method for Web applications, we are aware that more experimentation is needed to confirm these results. These results need to be interpreted with caution since they are only valid within the context established in this experiment. However, we have obtained valuable feedback from this empirical study with which to improve our proposal.

As future work, we plan to replicate this experiment by considering subjects with different levels of experience in usability evaluations (by including practitioners) and by considering other kinds of Web applications.