Further Analysis on the Validation of a Usability Inspection Method for Model-Driven Web Development

Conference Paper · September 2012

DOI: 10.1145/2372251.2372278

4 authors:

Adrian Fernandez
Softonic International
22 PUBLICATIONS **235** CITATIONS
SEE PROFILE

Silvia Abrahão
Universitat Politècnica de València
135 PUBLICATIONS **1,036** CITATIONS
SEE PROFILE

Emilio Insfran
Universitat Politècnica de València
123 PUBLICATIONS **1,149** CITATIONS
SEE PROFILE

Maristella Matera
Politecnico di Milano
158 PUBLICATIONS **2,349** CITATIONS
SEE PROFILE

Further Analysis on the Validation of a Usability Inspection Method for Model-Driven Web Development

Adrian Fernandez[1], Silvia Abrahão[1], Emilio Insfran[1], Maristella Matera[2]

[1]Universitat Politècnica de València
Camino de Vera, s/n, 46022, Valencia, Spain
+34 96 38773 50
{afernandez,sabrahao,einsfran}@dsic.upv.es

[2]Politécnico di Milano
via Ponzio, 34/5 - 20133, Milan, Italy
+39 02 23993408
matera@elet.polimi.it

ABSTRACT

Currently, there is a lack of empirically validated usability evaluation methods that can properly be integrated during the early stages of Web development processes. This has motivated us to propose a usability inspection method called WUEP that can be integrated into different model-driven Web development processes. In previous work, we presented the operationalization and validation of WUEP in a specific process based on the Object-Oriented Hypermedia (OO-H) method. In this paper, we present further analysis of the empirical validation of the operationalization of WUEP into WebML, which is one of the most well-known industrial model-driven Web development process. The effectiveness, efficiency, perceived ease of use, and satisfaction of WUEP was evaluated in comparison to Heuristic Evaluation. The results show that WUEP is more effective and efficient than heuristic evaluation in the detection of usability problems. The inspectors were also satisfied when applying WUEP, and found it easier to use than heuristic evaluation.

Categories and Subject Descriptors

D.2.4 [**Software Engineering**]: Software/Program Verification - *Validation*; D.2.9 [**Software Engineering**]: Management - *Software quality assurance*.

General Terms

Measurement, Design, Experimentation, Human Factors.

Keywords

Usability Inspection, Model-driven Web development, Controlled Experiment.

1. INTRODUCTION

Usability is considered to be one of the most important quality factors for Web applications, along with others such as reliability and security. The challenge of developing more usable Web applications has promoted the emergence of a large number of usability evaluation methods. However, most of these approaches only consider usability evaluations during the final stages of the Web development process. Works such as that of Matera *et al.* and [10] and Juristo *et al.* [9] claim that usability evaluations should also be performed during the early stages of the development process in order to improve the user experience and decrease maintenance costs.

To address these issues, we have proposed a usability inspection method (i.e., Web Usability Evaluation Process – WUEP [7]), which can be instantiated and integrated into different model-driven Web development processes. In this type of processes, intermediate artifacts (i.e., 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 from these models. In this context, inspections of these models can provide early usability evaluation reports to identify usability problems that can be corrected prior to the generation of the source code.

Besides the need of evaluation method we also envision the need of empirical studies to evaluate and improve any new proposed evaluation method. These studies can indeed provide useful information when a method is compared to others. Several empirical studies for validating Web usability evaluation methods exist (e.g., [5]). However, they focus on traditional Web development processes. There are few empirical studies based on the model-driven Web development processes (e.g., [10][1][6][13]). Among these studies, we presented in [6] an operationalization and validation of WUEP in a specific process followed by the Object-Oriented Hypermedia (OO-H) method. In this work, WUEP was compared against Heuristic Evaluation (HE) and the results showed that WUEP is more effective and efficient than HE in the detection of usability problems.

However, in other to verify the generalization of WUEP into another process this inspection method has been operationalized for use with the Web Modeling Language (WebML) [4], which is one of the most well-known industrial model-driven Web development process. This operationalization consisted in adapting the generic measures taken from the Web Usability Model [7] (that drives the inspection process followed by WUEP) to apply them to WebML artifacts as a way to predict the usability of Web applications early on the process. In this work, we present the results of a controlled experiment aimed at providing further analysis about the effectiveness, efficiency, perceived ease of use, and satisfaction of WUEP in detecting usability problems when integrated for use with WebML. WUEP was evaluated in comparison to Heuristic Evaluation (HE), which is a widely-used inspection method in industry.

This paper is structured as follows. Section 2 shows the evaluated usability inspection methods. 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.

ESEM'12, September 19–120 2012, Lund, Sweden.

2. EVALUATED INSPECTION METHODS

The evaluated methods are two usability inspection methods: our proposal (WUEP), and the Heuristic Evaluation (HE) proposed by Nielsen [12]. Inspection methods are used by evaluators to evaluate artifacts (normally User Interfaces – UIs) 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 development process and not only when the software application has been completed and deployed.

The Web Usability Evaluation Process (WUEP) extends and adapts the quality evaluation process proposed in the ISO 25000 (SQuaRE) [8] with the purpose of integrating usability evaluations into model-driven Web development processes. WUEP employs a Web Usability Model that decomposes usability into sub-characteristics and measurable attributes. Measures with a generic definition are associated to these attributes in order for them to be operationalized at different abstraction levels (e.g., abstract UI) in any model-driven Web development process. The aim of applying measures 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 (e.g., scope, Web application selection, attributes selection, Web artifacts selection), the specification of the evaluation (e.g., operationalization of measures, rating levels for measures), and the design of the evaluation (e.g., number of evaluators, evaluation plan). The *evaluation executor* applies the evaluation plan designed in the execution stage (measures calculation, usability problem reports), and finally, the *Web developer* performs the analysis of changes in order to correct the usability problems detected.

The Heuristic Evaluation (HE) requires a group of evaluators to examine the UI in compliance with recognized usability principles called *heuristics*. HE proposes 10 heuristics that are intended to cover the best practices in the design of any UI (e.g., minimize the user workload, error prevention). There are two roles involved in HE: evaluation designer and evaluation executor. The *evaluation designer* determines the scope of the evaluation and defines the evaluation plan. The *evaluation executor* applies the heuristics to Web artifacts to identify and report the usability problems. HE was selected because i) it is widely-used in industrial settings and ii) it can be applied to intermediate artifacts (e.g., mock-ups) produced during early stages of Web development. It is worth to mention that there is no other inspection method for model-driven Web development processes 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** the WUEP operationalization for the WebML development process, **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 novice usability evaluators.

The context of the experiment is the usability evaluation of two Web applications performed by novice inspectors. This context is determined by the *Web applications* to be evaluated, the *usability evaluation methods* to be applied and the *subject selection*. The *Web applications* selected are a Web Calendar for meeting appointment management, and an e-commerce application for a Book Store. They were developed by a Web development company using the WebML model-driven development process.

Two different functionalities of the Web Calendar application (Appointment management and User comments support) were selected for defining the experimental object O1, whereas two different functionalities of the Book Store application (Book search and Book shopping) were selected for defining the experimental object O2. Each experimental object contains two Web artifacts: a Hypertext model (HM), specifying the structure of the applications through the WebML design notation, and a Final User Interface (FUI). We selected these four functionalities since they are relevant to the end-users and similar in size and complexity. The *usability inspection methods* to be evaluated were WUEP and HE, and only their execution stages were considered. Thirty *subjects* were chosen from a group of fifth-year Computer Science students from the Universitat Politècnica de València, who were enrolled on an Advanced Software Engineering course from September 2011 to January 2012.

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:

- **H1-0:** There is no significant difference between the effectiveness of WUEP and HE / **H1-a:** WUEP is significantly more effective than HE.
- **H2-0:** There is no significant difference between the efficiency of WUEP and HE / **H2-a:** WUEP is significantly more efficient than HE.
- **H3-0:** There is no significant difference between the perceived ease of use of WUEP and HE / **H3-a:** WUEP is perceived to be significantly easier to use than HE.
- **H4-0:** There is no significant difference between the evaluators' perceived satisfaction of applying WUEP and HE / **H4-a:** WUEP is perceived to be significantly more satisfactory to use than HE.

The experiment was planned as a balanced within-subject design with a confounding effect, signifying that the same subjects use both methods in a different order and with different experimental objects (the subjects' assignation 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. Several documents were designed as instrumentation for the experiment: slides for training session, an explanation of the methods, gathering data forms, and two questionnaires.

Table 1. Schedule of the controlled experiment

	1st Day		2nd Day	
Training (15+20 m)	WebML introduction			
	Inspection using HE		Inspection using WUEP	
1st Session (90 min)	HE in O1	HE in O2	WUEP in O1	WUEP in O2
	HE Questionnaire		WUEP Questionnaire	
Break (180 min)				
Training (20 min)	Inspection using WUEP		Inspection using HE	
2nd Session (90 min)	WUEP in O2	WUEP in O1	HE in O2	HE in O1
	WUEP Questionnaire		HE Questionnaire	

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 9 and 11 usability problems in the experimental objects O1 and O2, respectively.

The quantitative analysis was performed by using the SPSS v16 statistical tool and α=0.05. Table 2 summarizes the overall results of the usability evaluations. Mean and standard deviation were used as descriptive statistics also for the PEU and PSU subjective variables, being the five-point Likert scale adopted for their measurement an interval scale

Table 2. Overall results

	# Problems / Subject		**False positives / Subject**		**Replicated Prob. / Subject**	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
HE	3.29	1.08	1.38	1.24	0.88	0.80
WUEP	**6.50**	1.14	**0.54**	0.66	**0.00**	0.00

	Duration (min)		**Effectiveness (Effec) (%)**		**Efficiency (Effic) (prob / min)**	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
HE	**70.13**	13.52	33.04	10.85	0.05	0.02
WUEP	80.88	18.46	**65.32**	11.54	**0.08**	0.02

	Perceived Ease of use (PEU)		**Perceived Satisfaction of use (PSU)**	
	$\bar{x}$	σ	$\bar{x}$	σ
HE	3.38	0.73	3.63	0.67
WUEP	**3.80**	0.72	**3.92**	0.75

The overall results obtained have allowed us to interpret that WUEP has achieved the subjects' best performance in about all the analyzed statistics (see cells in bold), The only exception is the duration of the evaluation session, which however was longer for WUEP due to the longer time required to read the material containing the WUEP description As indicated by the results, WUEP tends to provide a low degree of 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 measures that are classified to be applied in one type of Web artifact.

Since the sample size is smaller than 50, we applied the Shapiro-Wilk test to verify whether the data was normally distributed. 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 usability inspection method.

Table 3. Shapiro-Wilk Normality test results

	Effec.	**Effic.**	**PEU**	**PSU**
HE	0.219	0.722	0.414	0.281
WUEP	0.021 * (< 0.05)	0.296	0.072	0.053

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 to 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 (since *WUEP_Effec* is not normally distributed), and the 1-tailed *t*-test for independent samples to verify H2, H3 and H4.

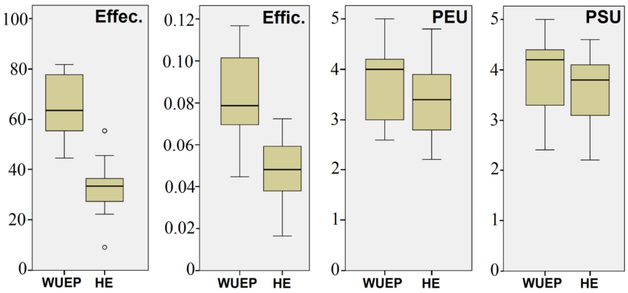

Figure 1. Boxplots for each dependent variable

The *p*-values obtained from the Mann-Whitney test for the *Effec*. variable was 0.000. The *p*-values obtained from the 1-tailed *t*-test test for the *Effic*., *PEU* and *PSU* variables were 000, 0.026 and 0.086, respectively. These results therefore support the rejection of all the null-hypotheses and the acceptance of their respective alternative-hypotheses except from the H4 (0.086 > 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: Diff_x = $\text{observation}_x(A)$ - $\text{observation}_x(B)$, where x denotes a particular subject, and A, B are the two possible values of one independent variable. We created Diffs 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, with the exception of Effic_Diff (HE) that was not. We therefore applied the parametric 2-tailed *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 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

Order of	**Effec.**	**Effic.**	**PEU**	**PSU**
Methods	0.095	0.291	0.173	0.560
Experimental Objects	0.989	0.932	0.709	0.560

Finally, a qualitative analysis was performed by analyzing the open-questions that were included in the questionnaire. This analysis revealed some important issues which can be considered to improve WUEP (e.g., the evaluators suggested that WUEP might be more useful if its evaluation process were automated by a tool (particularly the calculation of certain metrics).

5. THREATS TO VALIDITY

The main threats to the **internal validity** of the experiment are: learning effect, evaluation design, subject experience, and information exchange among evaluators. 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 selection of attributes to be evaluated during the design stage of WUEP. We attempted to alleviate this threat by considering relevant usability attributes, although empirical studies that involve experts in the Web domain are needed to provide the evaluator designer with the most relevant usability attributes for each Web application family. Subject experience was alleviated due to the fact that none of the subjects had any experience in usability evaluations. Information exchange might have affected the results since the experiment took place over two days, and it is difficult to be certain whether the subjects exchanged any information with each other.

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 novice evaluators with no experience in usability evaluations. However, the previous selection of usability attributes with their operationalized measures and the selection of the Web application might have affected the representativeness. To alleviate these issues, we intend to carry out a survey with Web designers to determine the relative importance of the usability attributes for different categories of Web applications. Since the duration of the experiment was limited to 90 min, only 3 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 used in the quantitative analysis [5]. The reliability of the questionnaire was tested by applying the Cronbach test. Questions related to PEOU and PU obtained a Cronbach's alpha of 0.80 and 0.78, respectively. 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 presented a controlled experiment for validating a usability inspection method (WUEP) when integrated into the WebML model-driven development process. The effectiveness, efficiency, perceived ease of use and satisfaction of WUEP were compared against 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 WebML artifacts. The evaluators found it easier to use than HE. Although they were also more satisfied when applying WUEP, this last variable resulted not statistically significant. These results confirmed our previous findings [6] when an operationalization of WUEP into the OO-H method was compared against 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 with subjects with different level of experience in usability evaluations (including practitioners) and by considering other kinds of Web applications such as mashups.

7. ACKNOWLEDGMENTS

This research work is funded by the MULTIPLE project (TIN2009-13838), the FPU program (AP2007-03731).

8. REFERENCES

[1] Abrahão, S., Iborra, E.,Vanderdonckt, J. 2007. Usability Evaluation of User Interfaces Generated with a Model-Driven Architecture Tool. Maturing Usability: Quality in Software, Interaction and Value, Springer, pp. 3-32.

[2] Basili, V., Rombach, H. 1988. The TAME Project: Towards Improvement-Oriented Software Environments, IEEE Transactions on Software Engineering 14(6), pp. 758-773.

[3] Briand, L., Labiche, Y., Di Penta, M., Yan-Bondoc, H. 2005. "An experimental investigation of formality in UML-based development", IEEE TSE, 31(10), pp. 833–849.

[4] Ceri, S., Fraternali, P., Bongio, A. 2000. Web modeling language (WebML): a modeling language for designing Web sites. 9th World Wide Web Conference, pp. 137–157.

[5] Conte, T., Massollar, J., Mendes, E., Travassos, G. H. 2007. Usability Evaluation Based on Web Design Perspectives. In Proc. of ESEM'07, Spain, pp. 146-155.

[6] Fernandez, A., Abrahão S., Insfran E. 2010. Towards to the validation of a usability evaluation method for model-driven web development, In Proc. of ESEM'10, Bolzano, Italy.

[7] Fernandez, A., Insfran, E., Abrahão, S. 2009. Integrating a Usability Model into a Model-Driven Web Development Process. In Proc. WISE'09, pp. 497-510, Springer.

[8] ISO/IEC. 2005. ISO/IEC 25000 series, Software Product Quality Requirements and Evaluation (SQuaRE).

[9] Juristo, N., Moreno, A., Sánchez-Segura, M.I. 2007. Guidelines for eliciting usability functionalities. IEEE Transactions on Software Engineering 33 (11), pp. 744-758.

[10] Matera, M., Costabile, M. F., Garzotto, F., Paolini, P. 2002. SUE inspection: an effective method for systematic usability evaluation of hypermedia. IEEE Transactions on Systems, Man, and Cybernetics, Part A 32(1): 93-103

[11] Maxwell, K. 2002. Applied Statistics for Software Managers. Software Quality Institute Series, Prentice Hall.

[12] Nielsen, J. 1994. Heuristic evaluation. Usability Inspection Methods, John Wiley & Sons, NY.

[13] Panach, I., Condori, N., Valverde, F., Aquino, N., Pastor, O. 2008. Understandability measurement in an early usability evaluation for MDD. In Proc. of ESEM'08, pp. 354-356.

[14] Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., Weslen, A. 2000. Experimentation in Software Engineering - An Introduction, Kluwer.