Empirical Validation of a Usability Inspection Method for Model-Driven Web Development

Adrian Fernandez[a,1], Silvia Abrahão[a], Emilio Insfran[a]

[a] *ISSI Research Group, Department of Information Systems and Computation, Universitat Politècnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain.*

Abstract

Web applications should be usable in order to be accepted by users and to improve the success probability. Despite the fact that this requirement has promoted the emergence of several usability evaluation methods, there is a need for empirically validated methods that provide evidence about their effectiveness and that can be properly integrated into early stages of Web development processes. Model-driven Web development processes have grown in popularity over the last few years, and offer a suitable context in which to perform early usability evaluations due to their intrinsic traceability mechanisms. These issues have motivated us to propose a Web Usability Evaluation Process (WUEP) which can be integrated into model-driven Web development processes. This paper presents a family of experiments that we have carried out to empirically validate WUEP. The family of experiments was carried out by 64 participants, including PhD and Master's computer science students. The objective of the experiments was to evaluate the participants' effectiveness, efficiency, perceived ease of use and perceived satisfaction when using WUEP in comparison to an industrial widely-used inspection method: Heuristic Evaluation (HE). The statistical analysis and meta-analysis of the data obtained separately from each experiment indicated that WUEP is more effective and efficient than HE in the detection of usability problems. The evaluators were also more satisfied when applying WUEP, and found it easier to use than HE. Although further experiments must be carried out to strengthen these results, WUEP has proved to be a promising usability inspection method for Web applications which have been developed by using model-driven development processes.

Keywords: Usability inspection, Web applications, Model-driven development, Family of experiments

[1] Corresponding author: Tel: +34 96 387 73 50 (Ext: 83525); Fax: +34 96 387 73 59
Email addresses: afernandez@dsic.upv.es (Adrian Fernandez), sabrahao@dsic.upv.es (Silvia Abrahão), einsfran@dsic.upv.es (Emilio Insfran),

1. INTRODUCTION

Web applications play an important role in business activities, information exchange, and social networks. The acceptability of Web applications relies on the ease or difficulty that users experience with this kind of systems [Matera et al. 2006]. Usability is therefore considered to be one of the most important quality factors for Web applications [Offut 2002].

The challenge of developing more usable Web applications has led to the emergence of usability evaluation methods with which to address Web usability. These methods can be principally classified into two different types: *empirical methods* and *inspection methods*. Empirical methods are based on observing, capturing and analyzing usage data from real end-users, whereas inspection methods are performed by expert evaluators or Web designers, and are based on reviewing usability principles in Web artifacts (e.g., mockups, conceptual models, user interfaces) with regard to their conformance with a set of guidelines.

The employment of usability evaluation methods to evaluate Web artifacts was investigated through a systematic mapping study in a previous work [Fernandez et al. 2011a]. This study revealed various findings such as:

a) There is a lack of usability evaluation methods that can be properly integrated into the early stages of Web development processes.
b) There is a shortage of usability evaluation methods that have been empirically validated.

These results, and particularly finding 'a)', motivated us to propose the Web Usability Evaluation Process (WUEP) in previous research [Fernandez et al. 2011b]. WUEP is an inspection method which can be instantiated and integrated into different model-driven Web development (MDWD) processes. Most MDWD processes break the Web application design up into three dimensions: content, navigation and presentation. These dimensions allow proper levels of abstraction to be established [Casteleyn et al. 2009]. An MDWD process basically transforms models that are independent of technological implementation details (i.e., Platform-Independent Models - PIMs), such as structural models, navigational models or abstract user interface (UI) models, into other models that contain specific aspects from a specific technological platform (i.e., Platform-Specific Models - PSMs), such as specific user interface models or database schemas. This is done automatically by applying transformation rules. PSMs can be automatically compiled to generate the source code of the final Web application. In this respect, evaluations of these models can provide early usability evaluation reports with which to identify problems that will appear at the final Web application and consequently, can be corrected prior to the generation of the source code.

With regard to finding 'b)', we performed a controlled experiment as an initial step in the empirical validation of the Web Usability Evaluation Process (WUEP) [Fernandez et al. 2010]. However, replications are necessary if the results are to have greater validity. The concept of replication is extended to the "family of experiments" reported by Basili et al. [1999]. A family is composed of multiple similar experiments that pursue the same goal to build the knowledge needed to extract significant conclusions.

In this paper, we present the results of a family of three controlled experiments carried out to investigate the *Effectiveness*, *Efficiency*, *Perceived Ease of Use*, and *Perceived Satisfaction of Use* of WUEP when compared with a representative inspection method that is commonly applied in industry: Heuristic Evaluation (HE) [Nielsen 1994]. These controlled experiments were conducted with Computer Science Master's degree students from the Universitat Politècnica de València (UPV) in Spain.

The paper is structured as follows. Section 2 discusses related work in the empirical validation of inspection methods with which to evaluate Web usability. Section 3 briefly introduces the inspection methods (WUEP and HE) which were used in the family of experiments. Section 4 presents the family of experiments. Section 5 provides details of the individual designs of each experiment. Section 6 reports and analyzes the results obtained from each experiment. Section 7 summarizes the results of the family of experiments, and a meta-analysis to provide a global analysis of the individual experiments. This section also discusses possible threats to validity. Finally, Section 8 presents our conclusions and final remarks.

2. RELATED WORK

Since the late 1980s, usability inspection methods have emerged as a cost-effective alternative to empirical methods for identifying usability problems [Cockton et al. 2003]. In this context, several inspection methods (e.g., Heuristic Evaluation, Cognitive Walkthrough) were proposed by usability experts from the Human-Computer Interaction (HCI) field. Since the term "Web Engineering" was first published in 1997 [Gellersen et al. 1997], these existing HCI methods have been adapted and improved in order to be applied to Web applications, and other new usability evaluation methods specifically crafted for the Web domain have also appeared. In this section, we discuss related works that report on empirical validations and comparisons of usability inspection methods for Web applications.

2.1 Empirical Studies Involving Usability Inspection Methods for Traditional Web Development

Several empirical studies with which to validate the performance of usability inspection methods have been reported. These studies can be classified in two types according to their aim: a) empirical studies that were intended to perform comparative studies involving well-known usability inspection methods in order to guide researchers and practitioners, and b) empirical studies that were intended to empirically validate a specific usability inspection method which had been specifically proposed for the Web domain.

The following representative examples of comparative studies involving well-known usability inspection methods should be highlighted:

- Hvannberg et al. [2007] reported an experiment in which two usability inspection methods were compared: Heuristic Evaluation and Gerhardt-Powals Principles. A within-subjects experimental design was applied to evaluate the usability of a Web portal. The study found that there were no significant differences between both methods as regards their effectiveness and efficiency in the specified context.
- Koutsabasis et al. [2007] reported a case study in which the effectiveness of four usability evaluation methods was compared. Participants were divided into 9 groups, of which 3 and 2 groups of participants applied the Heuristic Evaluation and Cognitive Walkthrough inspection methods, respectively, and 3 and 1 groups of participants applied two empirical methods: Think-aloud protocol and Co-discovery Learning, respectively. The Co-discovery Learning method was found to be slightly more effective than the others.
- Ssemugabi and De Villiers [2007] reported a case study whose aim was to investigate the extent to which Heuristic Evaluation identifies usability problems in a Web-based learning application by comparing the results with those of Survey Evaluations among end-users. The Heuristic Evaluation performed by four expert evaluators proved to be an appropriate and effective usability evaluation method for e-learning applications.

- Tan and Bishu [2009] reported an experiment in which Heuristic Evaluation was compared to User Testing. Although Heuristic Evaluation was able to identify more usability problems, there were no significant conclusions regarding the effectiveness and efficiency of both methods since they aimed to evaluate different aspects of the Web application.

Most of the aforementioned empirical studies presented comparisons between usability inspection methods and empirical methods. It is important to highlight that these kinds of comparisons are useful for practitioners in that they provide guidance in the selection of proper usability evaluation methods in a specific context. However, we argue that usability inspection methods should be compared to other usability inspection methods since empirical methods tend to evaluate usability aspects discovered during user interaction rather than usability aspects discovered in Web artifacts.

The following representative examples of empirical validations of a specific usability inspection method which had been specifically proposed for the Web domain should be highlighted:

- Costabile and Matera [2001] presented the empirical validation of the Systematic Usability Evaluation (SUE) method which employed operational guidelines called Abstract Tasks. Two experiments involving 26 and 20 novice evaluators, respectively, were conducted. The first experiment confirmed that the SUE method enhanced the effectiveness and efficiency of the usability evaluation, along with the evaluators' satisfaction. The second experiment aimed to predict the number of evaluators needed to achieve a certain percentage of usability problems detected.
- Chattratichart and Brodie [2004] presented the empirical validation of the Heuristic Evaluation Plus method (HE-Plus), which is an extended version of the Heuristic Evaluation (HE) [Nielsen 1994]. The experiment consisted of two groups containing five participants each, which were randomly assigned to the two methods. The results showed that HE-Plus was more effective than HE.
- Hornbæk and Frøkjær [2004] presented the empirical validation of the Metaphor of Human-Thinking method (MOT). The experiment compared the proposed method with the Cognitive Walkthrough method. Evaluators applied both methods in a different order. The results showed that the participants were more effective in the detection of usability problems when using MOT. In addition, it achieved a broader coverage in the type of usability problems detected.
- Blackmon et al. [2005] presented the empirical validation of the Cognitive Walkthrough for the Web method (CWW). The experiment showed that CWW was more effective than the Cognitive Walkthrough method on which it is based, and it also considered CWW to be an effective inspection method with which to repair usability problems related to unfamiliar and confusable links.
- Conte et al. [2009] presented the empirical validation of the Web Design Perspectives method (WDP), which defines a set of heuristics by considering four different perspectives of a Web application: conceptual, structural, navigation and presentation. Two experiments that pursued different goals were performed in order to refine the approach. The results of the first experiment showed that WDP was a feasible method with which to detect usability problems, whereas the second experiment showed that WPD was more effective when it was compared to the Nielsen's Heuristic Evaluation.
- Malak and Sahraoui [2010] presented the definition and empirical validation of a probabilistic approach for building Web quality models in order to manage uncertainty and subjectivity, which are inherent to quality evaluation. This

approach was instantiated to evaluate the navigability of Web applications, which is considered to be a relevant sub-characteristic of usability [Leavit and Shneiderman 2006]. The results of an experiment conducted showed that the scores given by the proposed model are strongly correlated with navigability as perceived by the user.

Although the aforementioned empirical studies present the empirical validation of inspection methods, the majority of them tend to be isolated and are not replicated by using similar experimental settings in order to support a meta-analysis to aggregate empirical evidences from individual studies. In addition, most of these empirical studies consider only the objective dependent variables of the usability inspection method (mainly their effectiveness) when it is compared to another method. Although objective dependent variables such as effectiveness and efficiency are relevant, subjective dependent variables related to the evaluator's perceptions should also be considered since they likewise contribute to the acceptance of the usability inspection method in practice.

2.2 Empirical Studies Involving Usability Inspection Methods for Model-driven Web Development

Studies such as that of Juristo et al. [2007] claim that usability evaluations should also be performed during the early stages of the Web development process in order to improve user experience and decrease maintenance costs. We argue that model-driven Web development (MDWD) processes provide an appropriate context in which to conduct early usability evaluations, since models which are applied at all stages can be evaluated throughout the entire Web development process. Despite the fact that several MDWD processes have been proposed since the late 2000s, and are still evolving [Valderas and Pelechano 2011], few works address usability evaluations in model-driven Web development (e.g., Abrahão and Insfran [2006], Sottet et al. [2007], and Molina and Toval [2009]). There are consequently few studies that present empirical studies in this context. Some examples are Abrahão et al. [2007] and Panach et al. [2008].

Abrahão et al. [2007] present an empirical study which evaluates the user interfaces that were generated automatically by a model-driven development tool. This study applies two usability evaluation methods: an inspection method (i.e., Action Analysis [Olson and Olson 1990]) and an empirical method (i.e., User Testing) with the aim of comparing what types of usability problems are detected in the user interfaces and what their implications are for transformations rules and platform-independent models. However, the usability evaluation methods employed were not adapted to be applied in Web artifacts and no dependent variables were defined in order to compare the performance of both methods.

Panach et al. [2008] extended the usability model proposed in Abrahão and Insfran [2006], which decomposes usability into measurable attributes that are applied to software products obtained as result of a model-driven development process. The aim was to provide metrics with which to evaluate the understandability of Web applications (i.e., a usability sub-characteristic) and to aggregate the values obtained in order to provide attribute indexes. These indexes were compared to the perception of these same attributes by end users. However, the empirical validation was based on correlations between metric calculation and attribute perception. Moreover, it did not consider any performance measure of method use.

2.3 Discussion

The analysis of the aforementioned studies has allowed us to detect some limitations in the empirical validation of usability inspection methods such as: 1) the low number of empirical studies, particularly in the context of model-driven Web development; 2) the lack of frameworks and standard criteria for the comparison of usability evaluation methods; and 3) the fact that the majority of empirical validations tend to be isolated and not replicated.

The first limitation is in line with the results of our systematic mapping study, which revealed that only 44% of Web usability studies have reported empirical validations of the proposed and/or employed usability evaluation methods [Fernandez et al. 2011a]. This study showed that experiments were one of the most frequently employed types of empirical methods used for validation purposes since they provide a high level of control and are useful for comparing usability evaluation methods in a more rigorous manner. However, the majority of these experiments involved usability inspection methods that are oriented towards traditional Web development processes, and usability evaluations therefore principally took place in the later stages of the Web development process.

The second limitation is in line with studies such as that of Gray and Salzman [1998] in which it is claimed that most of the experiments based on comparisons of usability evaluation methods do not clearly identify which aspects of these methods are being compared. This issue was also detected by Hartson [2003], in which several studies were analyzed in order to determine which measures had been used in the validation of usability evaluation methods. The majority of these studies evaluated the effectiveness of usability evaluation methods using the thoroughness metric (i.e., the ratio between the number of real usability problems found and the number of total real usability problems). This study also claimed that the majority of these comparative studies did not provide the descriptive statistics needed to perform a meta-analysis of the empirical findings extracted from different sources.

The third limitation is in line with studies that have been performed in the Software Engineering field, such as that of Sjøberg et al. [2005]. This work claims that only 20 out of 113 controlled experiments are replications, and of these, 15 are differentiated replications (i.e., replications that introduce variations in essential aspects of the experimental conditions, such as executions of replications with different kinds of participants). Dealing with experimental replications has been addressed by the concept of the family of experiments. Although many empirical studies of this type have been applied in the Software Engineering field (e.g., Cruz-Lemus et al. [2011]; Abrahão et al. [2011]), few families of experiment have been reported in the Web Engineering field (e.g., Abrahão and Poels [2009]). Another issue also appears which is specific to the Web Engineering field: the majority of empirical studies cannot be considered to be methodologically rigorous. A systematic review presented by Mendes [2005] was performed to determine the rigor of claims of Web Engineering research. This review demonstrated that only 5% should be considered as rigorous. It also found that numerous Web Engineering papers used incorrect terminology (e.g., they used the term *experiment* rather than *experience report* or the term *case study* rather than *proof of concept*).

3. METHODS EVALUATED

The methods evaluated through the family of experiments were two inspection methods: our proposal (WUEP) and the Heuristic Evaluation (HE) proposed by Nielsen [1994]. An overview of both methods is presented in the following sub-sections. The rationale for selecting HE as the method used to compare our proposal is based on the following statements:

- WUEP should be compared with other inspection method since these methods allow us to evaluate Web artifacts that are produced during the early stages of the Web development process. Empirical methods which involve the participation of real users and are often used after development to assess a design are therefore discarded (e.g., User Testing or End-user Questionnaires). In this work, we are thus interested in comparing WUEP against other method that can be applied to obtain formative evaluations (i.e., evaluations carried out during development to improve a design).

- HE is one of the best-known inspection methods. This allows us to gather more accuracy information about its employment [Hollingsed and Novick 2007]..
- He is one of the most widely-used evaluation methods in industry. For instance, half of the ten Web intranets that won a 2005 competition used this method [Nielsen 2005].
- HE covers a broader range of usability aspects than other inspection methods such as, for instance, Cognitive Walkthroughs, whose usability definition is more focused on ease of navigation.
- HE has provided useful results when used to conduct Web usability evaluations [Sutcliffe 2002; Allen et al. 2006; Ssemugabi and De Villiers 2007].
- HE has often been used for comparison with other inspection methods [Costabile and Matera 2001; Chattratichart and Brodie 2004; Conte et al. 2009].
- No usability evaluation method has been previously defined for application in model-driven Web development processes. Since there is currently no standard inspection method for conducting Web usability evaluations, we cannot evaluate WUEP against a control method.

3.1 Web Usability Evaluation Process (WUEP)

The Web Usability Evaluation Process (WUEP) [Fernandez et al. 2011b] extends and adapts the quality evaluation process proposed in the ISO 25000 standard (SQuaRE) [2005] with the aim of integrating usability evaluations into model-driven Web development processes. WUEP employs a Web Usability Model [Fernandez et al. 2009] that decomposes the usability concept into sub-characteristics 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 Web Usability Model (including all the sub-characteristics attributes and their associated generic metrics) is available at http://www.dsic.upv.es/~afernandez/ WebUsabilityModel.

The aim of applying metrics was to reduce the subjectivity inherent to existing inspection methods. It is important to note that by applying metrics, the evaluators inspect these artifacts in order to detect problems related to the usability for end-users but not related to the usability of model-driven artifacts themselves. Therefore, inspection of these models (by considering the traceability among them) allows the source of the usability problem to be discovered and facilitates the provision of recommendations to correct these problems during the earlier stages of the Web development process. In other words, we are referring to a Web application that can be usable by construction [Abrahão et al. 2007].

Figure 1 shows an overview of the main stages of WUEP in which three roles are involved: evaluation designer, evaluator, and Web developer. The evaluation designer performs the first three stages: 1) Establishing the requirements of the evaluation; 2) Specification of the evaluation; and 3) Design of the evaluation. The evaluator performs the fourth stage: 4) Execution of the evaluation, and the Web developer performs the last stage: 5) Analysis of changes. A brief description of each stage is provided as follows:

1. In the *establishment of the evaluation requirements* stage, the scope of the evaluation is defined by a) establishing the purpose of the evaluation; b) specifying the evaluation profiles (type of Web application, Web development method employed, context of use); c) selecting the Web artifacts to be evaluated; and d) selecting the usability attributes from the Web usability model which are going to be evaluated.
2. In the *specification of the evaluation* stage, the metrics associated with the selected attributes are operationalized in order for them to be applied to the Web artifact to be evaluated. This operationalization consists of establishing a mapping between the generic description of the metric and the concepts that are represented in the Web artifacts (modeling primitives in models or UI elements in the final Web application). In addition, rating levels are established for ranges of values obtained for each metric by considering their scale type and the guidelines related to each metric whenever possible. These rating levels provide a classification of usability problems based on their severity: low, medium, or critical. It is important to note that the operationalization needs to be performed once by a concrete Web development method, and can be reused in further evaluations that involve Web applications from the same Web development method.
3. In the *design of the evaluation* stage, the template for usability reports is defined and the evaluation plan is elaborated (e.g., number of evaluators, evaluation restrictions).
4. In the *execution of the evaluation* stage, the evaluator applies the operationalized metrics to the selected artifacts in order to detect usability problems by considering the rating levels of each metric. An example of this metric application can be found in Appendix A.1.
5. In the *analysis of changes* stage, all the usability problems detected are analyzed in order to propose changes with which to correct the affected artifacts from a specific stage of the Web development process. The changes are applicable to the previous intermediate Web artifacts (i.e., platform-independent models, platform-specific models and model transformations if the evaluation is performed on the final Web user interface).

<< Insert Figure 1 approximately here >>

3.2 Heuristic Evaluation (HE)

The Heuristic Evaluation (HE) method requires a group of evaluators to examine Web artifacts (commonly user interfaces) in compliance with commonly-accepted 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, recognition rather than recall).

In order to facilitate both the method application and the method comparison, we have structured the method in the same main stages provided by WUEP. Figure 2 shows an overview of these stages in which three roles are also involved: evaluation designer, evaluation executor and Web developer. The evaluation designer performs the first three stages: 1) Establishing the requirements of the evaluation; 2) Specification of the evaluation; and 3) Design of the evaluation. The evaluator performs the fourth stage: 4) Execution of the evaluation, and the Web developer performs the last stage: 5) Analysis of changes. A brief description of each stage is provided as follows:

1. In the *establishment of the evaluation requirements* stage, the scope of the evaluation is defined by: a) establishing the purpose of the evaluation; b) specifying the evaluation profiles (type of Web application, Web development

method employed, context of use); and c) selecting the Web artifacts to be evaluated.

2. In the *specification of the evaluation* stage, the ten heuristics are described in detail by providing guidelines about which elements from the selected artifacts can be affected by each heuristic.
3. In the *design of the evaluation* stage, the template for usability reports is defined (e.g., structured reports or verbalized finding), and the evaluation plan is elaborated (e.g., number of evaluators, mechanisms to aggregate results, evaluation restrictions).
4. In the *execution of the evaluation* stage, the evaluator applies the heuristics to the selected artifacts (when its expressiveness allows the heuristic to be applicable) in order to detect usability problems. An example of this heuristic application can be found in Appendix A.2.
5. In the *analysis of changes* stage, all the usability problems detected are analyzed in order to propose changes with which to correct the affected artifacts.

<< Insert Figure 2 approximately here >>

4. THE FAMILY OF EXPERIMENTS

An increasing understanding exists that empirical studies are needed to create, improve, or assess processes, methods, and tools for software development [Basili et al. 1986; Basili 1996; Fenton 1993], maintenance [Colosimo et al. 2009; Dzidek et al. 2008], and quality evaluation [Bolchini and Garzotto 2007]. An empirical study is generally an act or operation by which to discover something that is unknown, or to test hypotheses [Basili 1993]. Research strategies include controlled experiments, qualitative studies, surveys, and archival analyses [Juristo and Moreno 2001; Wohlin et al. 2000]. However, replications of these studies are necessary if their results are to achieve greater validity [Shull et al. 2008; Kitchenham 2008]. In this respect, the “family of experiments” as an empirical research methodology has arisen with the aim of extracting significant conclusions from multiple similar experiments that pursue the same goal.

In this section, we present the family of experiments that we performed to empirically validate WUEP. This empirical study is also intended to contribute to Software Engineering research through proposing a well-defined framework that can be reused by other researchers in the empirical validation of their usability evaluation methods. The research methodology adopted is an extension of the five-steps proposed by Ciolkowski et al. [2002], in which the fifth step, “Family data analysis”, has been replaced with “Family data analysis and meta-analysis”, and it was guided by the experimental process of Wohlin et al. [2000].

4.1 Step 1. Experiment Preparation

The experiment was prepared by carrying out the following steps: 1) the establishment of the goal of the family of experiments; 2) the selection of variables; 3) the formulation of hypotheses; and 4) the experimental design, which all the individual experiments have in common. These issues are described in the following subsections.

4.1.1. Goal of the family of experiments. According to the Goal-Question-Metric (GQM) paradigm [Basili and Rombach 1988], the goal of our family of experiments is to analyze the Web Usability Evaluation Process (WUEP) in order to evaluate it with regard to its effectiveness, efficiency, perceived ease of use, and perceived satisfaction in comparison to the Heuristic Evaluation (HE) from the viewpoint of a set of usability inspectors. This experimental goal will also allow us to show the feasibility of our approach when it is applied to Web artifacts from a model-driven Web development process, in addition to detecting issues that can be improved in future versions of WUEP.

4.1.2. Independent and Dependent Variables. There are two independent variables in the family of experiments:

- The evaluation method, with nominal values: WUEP and HE.
- The experimental objects (collection of Web artifacts) to which both methods are applied, with nominal values: O1 and O2. A detailed description of these experimental objects is provided in Section 4.2.1.

There are two objective dependent variables, which were selected by considering works such as Hartson et al. [2000] and Gray and Salzman [1998]:

- *Effectiveness*, which is calculated as the ratio between the number of usability problems detected and the total number of existing (known) usability problems. We consider one usability problem as one defect that can be found in different artifacts independently of its severity level and its total number of occurrences.
- *Efficiency*, which is calculated as the ratio between the number of usability problems detected and the total time spent on the inspection process.

The measurement of these variables involves several issues. Since the experimental objects have been extracted from a real Web application, it is not possible to anticipate all the existing problems in the artifacts to be evaluated. For this reason, a control group (formed of two independent evaluators who are experts in usability evaluations and one of the authors of this paper) was created in order to provide a baseline of usability problems by applying an Expert Evaluation as *ad-hoc* inspection method based on their own expertise. In addition, this control group was also responsible to determine whether the usability problems reported by the participants in each experiment were false positives (no real usability problems), problems that have been reported more than once (replicated problems), or new problems that need to be added to the baseline (increasing the total number of existing usability problems). Disagreements among control group members were resolved by consensus.

There are also two subjective dependent variables, which were based on constructs from the Technology Acceptance Model (TAM) [Davis 1989] since TAM is one of the most widely applied theoretical model to study user acceptance and usage behavior of emerging information technologies, and it has received extensive empirical support through validations and replications [Venkatesh 2000]:

- *Perceived Ease of Use*, which refers to the degree to which evaluators believe that learning and using a particular evaluation method will be effort-free.
- *Perceived Satisfaction of Use*, which refers to the degree to which evaluators believe that the employment of a particular evaluation method can help them to achieve specific abilities and professional goals.

Both variables are measured using a set of 8 closed-questions: 5 questions with which to measure Perceived Ease of Use (PEU), and 3 questions with which to measure Perceived Satisfaction of Use (PSU). The closed-questions were formulated by using a 5-point Likert scale, using the opposing statement question format. In other words, each question contains two opposite statements which represent the maximum and minimum possible values (5 and 1), in which the value 3 is considered to be a neutral perception. Each subjective dependent variable was quantified by calculating the arithmetical mean of its closed-question values. Table 1 presents the questions associated with each subjective dependent variable.

<< Insert Table 1 approximately here >>

It is important to note that both objective and subjective variables are related to the employment of Web usability evaluation methods, not the usability evaluation of a Web application by involving end users.

4.1.3. Hypotheses. We formulated the following null hypotheses, which are one-sided since we expected WUEP to be superior to HE for each dependent variable. Each null hypothesis and its alternative hypothesis are presented as follows:

- $H1_0$: 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 perceived satisfaction of employing WUEP and HE.
- $H4_a$: WUEP is *perceived to be significantly more satisfactory to use* than HE.

4.1.4. Experimental Design. The experiment was planned as a balanced within-subject design with a confounding effect, signifying that the same number of participants used both methods in a different order and with different experimental objects. Table 2 shows the schema of the experimental design which has been used in all the individual experiments. Although this experimental design was intended to minimize the impact of learning effects on the results, since none of the participants repeated any of the methods in the same experimental object, other factors were also present that needed to be controlled since they may have influenced the results. These factors were:

- Complexity of experimental objects, since the comprehension of the modeling primitives from Web artifacts may have affected the application of both inspection methods. We attempted to alleviate the influence of this factor by selecting representative Web artifacts that were considered suitable, in both size and complexity, for application in the time available for the execution of the experiment, and also by providing a complete description of the Web artifacts to be evaluated (graphical and textual).
- Order of experimental objects and methods, since this may have caused learning effects, thus biasing results. We attempted to check the influence of this factor by applying proper statistical tests.

<< Insert Table 2 approximately here >>

4.2 Step 2. Context Definition

The context was determined by a) the Web application to be evaluated; b) the usability evaluation methods to be applied; and c) the subject selection. These are described in the following subsections.

4.2.1. Web Application Evaluated. We contacted a Web development company located in Alicante (Spain) in order to obtain Web artifacts from a real Web application. This Web application.was developed through the use of a model-driven Web development method called the Object-Oriented Hypermedia (OO-H) [Gomez et al. 2000] which is supported by the VisualWade tool[2].

OO-H provides the semantics and notation needed to develop Web applications. The platform-independent models (PIMs) that represent the different concerns of a Web application are: a class model, a navigational model, and a presentation model. The Class Model is UML-based and specifies the content requirements; the navigational model is composed of a set of Navigational Access Diagrams (NADs) that specify the functional requirements in terms of navigational needs and users' actions; and the presentation model is composed of a set of Abstract Presentation Diagrams (APDs), whose initial version is obtained by merging the Class Model and NADs, which are then refined in order to represent the visual properties of the final UI. The platform-specific models (PSMs) are embedded in a model compiler, which automatically obtains the source code (CM) from the Web application by taking all the previously mentioned platform-independent models as input.

The type of the provided Web application was an intranet for task management to be used in the context of a software development company. Two different functional features (Task management and Report management) were selected for the composition of the experimental objects (O1 and O2), as Table 3 shows in detail. We selected these functional features because they are relevant to the application users. These functional features are also similar in complexity, and their related Web artifacts are also similar in size. Each experimental object contains three Web artifacts: a Navigational Access Diagram (NAD), an Abstract Presentation Diagram (APD) model, and a Final User Interface (FUI).

<< Insert Table 3 approximately here >>

4.2.2. Inspection Methods Evaluated. Since the context of our family of experiments was from the viewpoint of a set of usability inspectors, we evaluated the execution stages of both methods (WUEP and HE), or in other words, the evaluators' application of both methods. Two of the authors therefore performed the evaluation designer role in both methods in order to design an evaluation plan. In critical activities such as the selection of usability attributes in WUEP, we required the help of two external Web usability experts. The outcomes of the stages performed by the evaluation designers are described as follows.

With regard to the establishment of the evaluation requirements stage, the first three activities (i.e., purpose of the evaluation, evaluation profiles, and selection of Web artifacts) were the same for both methods. In the case of the HE, all 10 heuristics were selected. In the case of the WUEP, a set of 20 usability attributes were selected as candidates from the Web Usability Model through the consensus reached by the two evaluator designers and the two Web usability experts. The attributes were selected by considering the evaluation profiles (i.e., which of them would be more relevant to the type of Web application and the context in which it is going to be used). Only 12 out of 20 attributes were randomly selected in order to maintain a balance in the number of metrics and heuristics to be applied.

With regard to the specification of the evaluation stage, the 10 heuristics from the HE were described in detail by providing guidelines concerning which elements can be considered in the Web artifacts to be evaluated. Examples of these heuristics can be found in Appendix B.1.2. In the case of the WUEP, 13 metrics associated with the 12 selected attributes were obtained from the Web Usability Model, and then associated with the artifact in which they could be applied. Since metrics can be applied at different abstraction levels, the highest level of application was selected.

[2] www.visualwade.com

Once the metrics had been associated with the artifacts, these metrics were operationalized in order to provide a calculation formula for artifacts from the OO-H method and to establish rating levels for them. Examples of these operationalized metrics can be found in Appendix B.1.1.

With regard to the design of the evaluation stage, the same evaluation plan (i.e., the experiment design), along with the same template with which to report usability problems, were defined for both methods. The templates employed for both inspection methods can be found in Appendix B.4.

4.2.3. Subject selection. Although expert evaluators are able to detect more usability problems than novice evaluators [Hertzum and Jacobsen 2001], we focus on this latter evaluator profile since the intention is to provide a Web usability evaluation method which enables inexperienced evaluators to perform their own usability evaluations. Therefore, the following groups of subjects were identified in order to facilitate the generalization of results:

– Master's students, all of whom had previously obtained a degree in Computer Science. At the moment of each experiment, they were attending a "Quality of Web Information Systems" course on the Masters in Software Engineering course at the Universitat Politècnica de València. It has been shown that, under certain conditions, there is no great difference between this type of students and professionals [Basili et al. 1999; Höst et al. 2000], and they could therefore be considered as the next generation of professionals [Kitchenham et al. 2002]. We therefore believe that their ability to understand Web artifacts obtained with model-driven Web development processes, and to apply usability evaluation methods to them, can be comparable to that of typical novice practitioners. With regard to their participation, all the Master's students were given one point in their final grades, regardless of their performances.
– PhD students, all of whom had previously obtained a degree in Computer Science and whose research activities are performed in the Software Engineering field. At the moment of each experiment, they were participants in the PhD Doctorate Program in Computer Science at the Universitat Politècnica de València. The participation of these PhD students in the experiments was voluntary.

We did not establish a classification of participants, since neither the Master's nor the PhD students had any previous experience in conducting usability evaluation studies. The assignation of the participants to the experimental groups was therefore random. With regard to the total number of participants, we tried to enroll the maximum possible participants in each individual experiment. Despite recent studies such as Hwang and Salvendy [2010] claims that 10±2 evaluators are needed to perform a usability evaluation to find around 80% of usability problems, these studies are related to the usability evaluation of a Web application, but not the evaluation of the performance and perceptions of Web usability evaluation methods. However, we ensured that at least 12 participants were involved as a sample size for each experiment in order to detect a representative number of usability problems.

4.3 Step 3. Experimental Tasks and Materials

The material was composed of the documents needed to support the experimental tasks and the training material. The documents used to support the experimental tasks were:

– Four kinds of data gathering documents in order to cover the four possible combinations (WUEP-O1, WUEP-O2, HE-O1, and HE-O2). Each document contained: the set of Web artifacts from the experimental object with a description of their modeling primitives (an example of the Web artifact evaluated can be found in Appendix B.2); and the description of the tasks to be performed in these artifacts (an example of these tasks for both usability inspection methods can be found in Appendix B.3). Although only three artifacts were evaluated (NAD, APD, and FUI), we also included a Class Diagram in order to provide a better understanding of the Web application's structure and content.

- Two appendixes containing a detailed explanation of each evaluation method (WUEP and HE) appear at the end of this paper.
- Two questionnaires (one for each method), which contained the closed-questions presented in Section 4.1.2 with which to evaluate the two subjective dependent variables (i.e., Perceived ease of use and Perceived satisfaction). Various questions belonging to the same dependent variable (i.e., construct group) were randomized to prevent systemic response bias. In addition, in order to ensure the balance of items in the questionnaire, half of the questions on the left-hand side were written as negative sentences to avoid monotonous responses [Hu and Chau 1999]. We also added two open-questions in order to obtain feedback on how to improve the ease of use and the employment of both methods. These open-questions were formulated as follows:
    - Q1: What suggestions would you make in order to improve the method's ease of use?
    - Q2: What suggestions would you make in order to make the metrics/heuristics more useful in the context of Web usability evaluations?

The training materials included: i) a set of slides containing an introduction to the Object Oriented Hypermedia method in order to present the modeling primitives of Web artifacts; (ii) a set of slides describing the WUEP method, with examples of metric application and the procedure to be followed in the experiments; and (iii) a set of slides describing the HE method with examples of heuristic application and the procedure to be followed in the experiments.

All the documents were created in Spanish, since this was the participants' native language. All the material (including the experimental tasks and the training slides) is available for download at www.dsic.upv.es/~afernandez/JSS/familyexp.html.

4.4 Step 4. Individual Experiments

Figure 3 summarizes the family of experiments by representing each individual experiment as a rectangle. This figure shows the order in which the experiments were executed (e.g., 1st experiment), the kind of participants involved and their number, the name associated with each experiment (e.g., EXP), and the kind of replication (e.g., internal replication). It is important to note that the number of participants is according to the final accepted samples, since we discarded incomplete samples, in addition to random samples when it was necessary to maintain the balanced within-subject design (i.e., the same number of participants per group).

The second and third experiments (REP1 and REP2) were differentiated replications of the original experiment (i.e., EXP) in different settings with different participants. In particular, REP2 is a strict replication of REP1 since the only variation in the execution of the experiment was the number of participants. With regard to the classification provided in Shull et al. [2008], the replications are exact replications since the procedure followed is as close as possible to the original experiment.

<< Insert Figure 3 approximately here >>

4.5 Step 5. Family Data Analysis and Meta-Analysis

The results of each individual experiment and the family of experiments were collected and analyzed.

With regard to the analysis of each individual experiment, we used boxplots and statistical tests to analyze the data collected. In particular, we tested the normality of the data distribution by applying the Shapiro-Wilk test. The results of the normality test allowed us to select the proper significance test in order to test our hypotheses. When data was assumed to be normally distributed (p-value ≥ 0.05), we applied the parametric one-tailed t-test for independent samples [Juristo and Moreno 2001]. However, when data could not be assumed to be normally distributed (p-value < 0.05), we applied the non-parametric Mann-Whitney test [Conover 1998].

In order to test the influence of Order of Method and Order of Experimental Objects (both independent variables), we used a method similar to that proposed by Briand et al. [2005]. We used the Diff function:

$$Diff_x = observation_x(A) - observation_x(B) \qquad (1)$$

where x denotes a particular subject, and A,B are the two possible nominal values of an 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. We also applied the Shapiro-Wilk test to prove the normality of the Diff functions. Table 4 presents the hypotheses related to the Diff functions, which are two-sided since we did not make any assumption about whether one specific order would be more influential than another. We verified these hypotheses by applying the parametric two-tailed t-test for independent samples or the non-parametric Mann-Whitney test depending on the results of the normality test.

<< Insert Table 4 approximately here >>

These statistical tests have been chosen because they are very robust and sensitive, and have been used in experiments similar to ours in the past, e.g., [Ricca et al. 2010; Briand et al. 2005; Conte et al. 2005]. As usual, in all the tests we decided to accept a probability of 5% of committing a Type-I-Error [Wohlin et al. 2000], i.e., of rejecting the null hypothesis when it is actually true.

We also performed a meta-analysis in order to aggregate the results, since the experimental conditions were very similar for each experiment. This analysis, which is detailed in Section 7.2, enabled us to extract more general conclusions with regard to each individual experiment.

5. DESIGN OF INDIVIDUAL EXPERIMENTS

In this section, we describe the main characteristics of each of the three individual experiments that constitute our family of experiments. In order to avoid useless redundancies, we discuss some clarifications of the original experiment related to the information presented in the previous section, and we only discuss the differences in the replications with regard to the original experiment.

5.1 The Original Experiment (EXP)

5.1.1. Planning. This section details the experimental plan by describing the context, the variables, hypotheses, experiment design, and instrumentation.

The context of the experiment: we used both of the experimental objects described in Section 4.2.1 (O1 and O2), we evaluated the execution stages by providing an evaluation design as described in Section 4.2.2 (10 heuristics to be applied with the HE method and 13 metrics to be applied with the WUEP method), and we selected 12 PhD students as participants whose profile is described in Section 4.2.3.

The variables: we selected all the independent and dependent variables described in Section 4.1.2.

The hypotheses: we tested all the hypotheses related to each dependent variable (Section 4.1.3) and all the hypotheses related to the influence of the order of methods and order of experimental objects (Section 4.5).

The experimental design: we used the balanced within-subject design with a confounding effect, presented in Section 4.1.4. Three participants were randomly assigned to each of the four groups, since there was no difference in their experience in Web usability evaluations.

The instrumentation: we used the documents presented in Section 4.3 to support the experimental tasks (4 data gathering documents, 2 appendices and 2 questionnaires) and the training material (3 slide sets).

5.1.2. Operation. This section details the experimental operation by describing the preparation, the execution, the data recording, and the data validation.

With regard to the preparation of the experiment, the experiment was planned to be conducted in two days owing to the participants' availability and the optimization of resources. Table 5 shows the planning for both days. The subjects were given a training session before each of the inspection methods was applied, in which they were also informed about the procedure to follow in the execution of the experiment. We established a time slot of 90 minutes as an approximation for each method application. However, we allowed the participants to continue the experiment even though these 90 minutes had passed in order to avoid a possible ceiling effect [Sjøberg et al. 2003].

<< Insert Table 5 approximately here >>

With regard to the execution of the experiment, the experiment took place in a single room and no interaction between participants was allowed. We logged all the interventions that were necessary to clarify questions concerning the completion of the experimental tasks, along with possible improvements that could be made to the experiment material. Finally, with regard to the data validation, we ensured that all the participants had completed all the requested data, and it was not therefore necessary to discard any samples.

5.2 The Second Experiment (REP1)

This second experiment (first replication) was different in three respects as regards the original experiment. These differences are described as follows:

- Subject selection. The participants were initially 38 Master's students. The profile of these subjects is described in Section 4.2.3, and all of them attended the "Quality of Web Information Systems" course which took place from April 2010 to July 2010. This course was selected because the necessary preparation and training, and the experimental task itself, fitted the scope of this course well. We took a "convenience sample" (i.e., all the students available in the class). We created two groups of 10 participants, and two groups of 9 participants, despite the fact that it would later be necessary to discard samples in order to maintain a balanced design.
- Metrics selection. Since only 12 out of 20 usability attributes were randomly selected from the Web Usability Model in the original experiment, we made minimal variations in order to enable new attributes to be evaluated as long as the evaluation design was not altered. In particular, we replaced one usability attribute with another, and we also replaced a metric from an existing attribute with another metric. We therefore maintained the same number of metrics to be applied, which was 13.

– Questionnaire. Table 6 presents the two new closed-questions that were added in order to evaluate the Perceived Satisfaction of Use. The questionnaire therefore contained a total of 10 closed-questions.

<< Insert Table 6 approximately here >>

With regard to the experiment preparation, the experiment was planned to be conducted over three days owing to the course timetable and the optimization of resources. Table 7 shows the planning for these days. On the first day, the participants were given the complete training and they were also informed of the procedure to follow in the execution of the experiment. They were told that their answers would be treated anonymously, and were also informed that their grade for the course would not be affected by their performance in the experiment. On the second and third days, the participants were given an overview of the complete training before applying the evaluation method, since all the groups were located in the same session. As in the previous experiment, we established a time slot of 90 minutes without a time limit for each method application.

<< Insert Table 7 approximately here >>

As in the original experiment, the experiment also took place in a single room and no interaction between participants was allowed. With regard to the data validation, we checked that all the participants had completed all the requested data. However, a total of 6 samples were discarded: 4 owing to incomplete data, and 2 of which were randomly discarded to maintain the same number of samples per group. The experiment eventually considered the results of only 32 evaluators (8 samples per group).

5.3 The Third Experiment (REP2)

This third experiment (second replication) was a strict replication of REP1. The difference with regard to REP1 was the subject selection. The participants were initially 35 Master's students (Section 4.2.3), all of whom attended the "Quality of Web Information Systems" course which took place from April 2011 to July 2011. We created three groups of 9 participants, and one group of 8 participants, despite the fact that it would later be necessary to discard samples in order to maintain a balanced design.

With regard to experiment preparation and execution, there were no differences with regard to REP1 since the same three day planning was followed. With regard to the data validation, we checked that all the participants had completed all the requested data. However, a total of 15 samples were discarded: 9 owing to incomplete data, and 6 of which were randomly discarded to maintain the same number of samples per group. The experiment eventually considered the results of only 20 evaluators (5 samples per group).

6. RESULTS AND DISCUSSION

After the execution of each 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.

In this section, we discuss the results of each individual experiment by quantitatively analyzing the results for each dependent variable and testing all the formulated hypotheses. We also analyze the influence of the order of methods and experimental objects. All the results were obtained by using the SPSS v16 statistical tool[3] with a statistical significance level of $\alpha = 0.05$. A qualitative analysis based on the feedback obtained from the open-questions in the questionnaire will also be provided.

6.1 Quantitative analysis

Table 8 summarizes the overall results of the usability evaluations performed in each experiment. The cells in bold type indicate the subjects' best performance in each statistic. The overall results obtained have allowed us to interpret that WUEP has achieved the subjects' best performance in all the statistics that were analyzed. As observed in these results, WUEP tends to provide a low degree of false positives and replicated problems. The low degree of false positives can be explained by the fact that WUEP aims to minimize the subjectivity of the evaluation by providing a more systematic procedure (metrics) to detect usability problems rather than interpreting whether the usability principles have been supported or not (heuristics). The low degree of replicated problems can be explained by the fact that WUEP provides operationalized metrics that have been previously classified to be applied in one type of artifact.

<< Insert Table 8 approximately here >>

The analysis of each dependent variable (Effectiveness, Efficiency, Perceived Ease of Use, and Perceived Satisfaction of Use) and the hypotheses testing is detailed in the following subsections.

6.1.1. Effectiveness. Figure 4 presents the boxplots containing the distribution of the Effectiveness variable per subject and per method for each of the individual experiments. These box plots show that WUEP was relatively more effective than HE when inspecting the usability of the experimental objects. Although we found the WUEP scores to be more scattered than those of HE (specifically in EXP and REP1), the median value for WUEP (between 50% and 60% of usability problems detected) was much higher than that for HE (between 20% and 40%). This may represent some variability in the participants' performance when detecting usability problems. However, the middle 50 percent of WUEP scores is above the third quartile of HE in all the individual experiments.

<< Insert Figure 4 approximately here >>

In order to determine whether or not these results were significant, we applied the Mann-Whitney non-parametric test to verify H1 in EXP, since Effectiveness(WUEP) for EXP was not normally distributed (*p*-value = 0.029), and the one-tailed *t*-test for independent samples to verify this in REP1 and REP2, since both Effectiveness(WUEP) and Effectiveness(HE) were normally distributed. The *p*-values obtained for these tests were: 0.001 for EXP, 0.000 for REP1, and 0.000 for REP2. These results therefore support the rejection of the null hypothesis $H1_0$ for each individual experiment (*p*-value < 0.05), and the acceptance of its alternative

[3] SPSS version 12.1.0 forWindows. SPSS Inc., Chicago, 2004.

hypothesis, meaning that the effectiveness of WUEP is significantly greater than the effectiveness of HE.

6.1.2. Efficiency. Figure 5 presents the boxplots containing the distribution of the Efficiency variable per subject and per method for each individual experiment. These box plots show that WUEP was relatively more efficient than HE when considering the usability of the experimental objects. As in the effectiveness results, the median value for WUEP (around 0.12 usability problems detected per minute) was much higher than that for HE (between 0.05 and 0.07). In fact, the middle 50 percent of the WUEP scores is also above the third quartile in all the individual experiments. However, we found the WUEP scores to be more scattered than those of HE in all the individual experiments. This might have been caused by differences in the duration of the evaluation in each method employment, since HE achieved a more constant and higher value than WUEP.

<< Insert Figure 5 approximately here >>

In order to determine whether or not these results were significant, we applied the Mann-Whitney non-parametric test to verify H2 in EXP, since Efficiency(HE) for EXP was not normally distributed (*p*-value = 0.045), and the one-tailed *t*-test for independent samples to verify this in REP1 and REP2, since both Efficiency(WUEP) and Efficiency(HE) were normally distributed. The *p*-values obtained for these tests were: 0.000 for EXP, 0.000 for REP1, and 0.000 for REP2. These results therefore support the rejection of the null hypothesis $H2_0$ for each individual experiment (p-value < 0.05), and the acceptance of its alternative hypothesis, meaning that the efficiency of WUEP is significantly greater than the efficiency of HE.

6.1.3. Perceived Ease of Use. Figure 6 presents the boxplots showing the distribution of the Perceived Ease of Use (PEU) variable per subject and per method for each individual experiment. These boxplots show that the participants perceived WUEP to be relatively easier to use than HE. The median value for WUEP (between3.8 and 4.4 points in the 5-point Likert scale) was slightly higher than that for HE (between 3 and 3.2 points). However, we found the HE scores to be more scattered than those of WUEP in all the individual experiments. This may represent controversial perceptions among participants.

<< Insert Figure 6 approximately here >>

In order to determine whether or not these results were significant, we applied the one-tailed *t*-test for independent samples to verify H3 in each individual experiment, since both PEU(WUEP) and PEU(HE) were normally distributed. The *p*-values obtained for these tests were: 0.003 for EXP, 0.000 for REP1, and 0.002 for REP2. These results therefore support the rejection of the null hypothesis $H3_0$ for each individual experiment (p-value < 0.05), and the acceptance of its alternative hypothesis, meaning that WUEP is perceived as easier to use than HE.

6.1.4. Perceived Satisfaction of Use. Figure 7 presents the boxplots showing the distribution of the Perceived Satisfaction of Use (PSU) variable per subject and method for each individual experiment. These boxplots show that the participants were more satisfied with WUEP than HE. The median value for WUEP (between 3.8 and 4.4 points in the 5-point Likert scale) was slightly higher than that for HE (around 3.5 points). However, we also found that the HE scores were more scattered than those for WUEP in all the individual experiments, particularly in EXP.

<< Insert Figure 7 approximately here >>

In order to determine whether or not these results were significant, we applied the one-tailed *t*-test for independent samples to verify H4 in EXP and REP1, since both PSU(WUEP) and PSU(HE) were normally distributed, and the Mann-Whitney non-parametric test to verify this in REP2, since PSU (HE) for REP2 was not normally distributed (*p*-value = 0.012). The *p*-values obtained for these tests were: 0.000 for EXP, 0.000 for REP1, and 0.025for REP2. These results therefore support the rejection of the null hypothesis $H4_0$ in each individual experiment (*p*-value < 0.05), and the acceptance of its alternative hypothesis, meaning that the subjects were more satisfied with the use of WUEP as compared to HE.

6.2 Influence of Order of Experimental Objects and Methods

We then applied the Shapiro-Wilk test to the Diff functions (Section 4.5), and this allowed us to determine that most of these functions were normally distributed (*p*-value ≥ 0.05). We also applied the two-tailed t-test for independent samples and the Mann-Whitney test (depending of the data distribution) in order to verify all the hypotheses related to the influence of order of method application (i.e., HM_1, HM_2, HM_3, and HM_4), and the influence of order of experimental object employment (i.e., HO_1, HO_2, HO_3, and HO_4). Table 9 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 method application and experimental object employment for any dependent variable.

<< Insert Table 9 approximately here >>

6.3 Qualitative Analysis

This analysis revealed several important issues which should be considered if WUEP is to be improved. With regard to the first open-question *"What suggestions would you make in order to improve the method's ease of use?"* (Section 4.3), the participants suggested that WUEP might be more useful if the evaluation process were automated or computer-aided (particularly the calculation of certain metrics). With regard to the second open-question: *"What suggestions would you make in order to make the metrics more useful in the context of Web usability evaluations?"*, the participants detected that providing more examples of how to apply the metrics might improve the application of the method. In addition, they suggested that a more detailed description of the operationalized metric might be useful since it was not always easy to identify elements of the Web artifacts involved in the metric calculation.

In the case of HE, and with regard to the first open-question, the participants recommended a previous classification of heuristics in order to determine which ones might be applicable to each kind of Web artifact obtained from a Model-driven Web development process, since this method has been commonly applied to the inspection of final user interfaces. With regard to the second open-question, the participants agreed that the heuristics need to be redefined to be more useful since their descriptions are too generic, thus leading inexperienced evaluators to obtain different interpretations.

7. FAMILY DATA ANALYSIS

This section provides a summary of the results obtained. We first present an analysis of the results in the context of the family of experiments, followed by the results of a meta-analysis that aggregates the empirical findings obtained in the individual experiments.

7.1 Summary of Results

We performed a global analysis of the results to determine whether the general goal of our family of experiments had been achieved. We also studied all the results to search for possible differences. A summary of the experiments and their results is provided in Table 10.

Three experiments were performed, in which data gathered from 64 subjects was used to test the formulated hypotheses (see Section 4.1.3). The main result of the family of experiments indicates that all the alternative hypotheses ($H1_a$, $H2_a$, $H3_a$, and $H4_a$) were supported in all the experiments. This outcome shows that WUEP was more effective and efficient than HE in the detection of usability problems in artifacts obtained using a specific model-driven Web development process (OO-H). In addition, the evaluators were more satisfied when they applied WUEP, and found it easier to use than HE.

<< Insert Table 10 approximately here >>

With regard to the Effectiveness variable (see Table VIII and Figure 4), we detected that WUEP was able to detect at least 50% of the total existing usability problems in each experiment, whereas HE accounted for at least 30% of the defects. It is important to note that only one set of metrics was selected in the evaluation design stage of WUEP, whereas in HE all ten heuristics were considered. This may represent promising results as regards the range of usability aspects that are considered in WUEP owing to the employment of its Web usability model. However, these results show that the ratio of usability problems detected are low for both methods, and could be improved by considering more usability attributes in WUEP and by refining the heuristic descriptions in HE.

With regard to the Efficiency variable (see Table VIII and Figure 5), we detected that those participants who used WUEP were able to detect one usability problem approximately every 7 minutes (between 0.14 and 0.17 usability problems per minute), whereas those participants who used HE detected one usability problem approximately every 14 minutes (between 0.05 and 0.07 usability problems per minute). This could have been owing to the fact that HE evaluators are required to spend more time on the interpretation of each heuristic in each Web artifact.

With regard to the Perceived Ease of Use variable (see Table VIII and Figure 6), we detected that WUEP achieved a mean score of 4.25, 4.16 and 3.73 points in the 5-point Likert scale, whereas HE achieved a mean score of 3.23, 3.44 and 3.03 points. This may indicate that metrics are perceived as easier to apply than heuristics. However, it is important to highlight that both scores are good results for both methods since all of them were above the neutral value established at 3 points.

With regard to the Perceived Satisfaction of Use variable (see Table VIII and Figure 7), we found that WUEP achieved a mean score of 4.32, 4.18 and 3.82 points in the 5-point Likert scale, whereas HE achieved a mean score of 3.36, 3.56 and 3.32 points. This may represent that metrics are perceived as a useful procedure by which to evaluate Web artifacts. These scores are also good results for both methods since all of them were above the neutral value established at 3 points. We also detected slight differences between both types of participants, since the PhD students achieved better results than the Master's students. This could have been owing to the former's level of experience in model-driven engineering.

With regard to the influence of other factors, statistical tests allowed us to conclude that there was no influence with regard to the order of method application and experimental object employment for any dependent variable. This strengthens the validity of our experimental design and also minimizes the possible learning effect when both methods are employed.

In summary, the results support the hypothesis that WUEP would achieve better results than HE in the specified context. According to the previously discussed results, we can conclude that WUEP can be considered as a promising approach with which to perform usability evaluations of Web artifacts obtained from a model-driven Web development process. However, WUEP was operationalized in the context of a specific process (OO-H). We plan to apply WUEP to evaluate the usability of Web artifacts obtained with other model-driven development processes (e.g., WebML [Ceri et al. 2000], UWE [Koch and Kraus 2003]). Feedback on how to improve the approach was also obtained. Running a family of experiments (including replications) rather than a single experiment provided us with more evidence of the external validity, and thus the generalization of the study results. Each replication provided further evidence of the confirmation of the hypothesis. We can thus conclude that the general goal of the empirical validation has been achieved.

7.2 Meta-Analysis

Although there are several statistical methods with which to aggregate and to interpret the results obtained from interrelated experiments [Glass et al. 1981; Hedges and Olkin 1985; Rosenthal 1986; Sutton et al. 2001], we used meta-analysis because it allowed us to extract more general conclusions.

Meta-analysis is a set of statistical techniques for combining the different effect sizes of the experiments to obtain a global effect of a factor. In particular, the estimation of effect sizes can be used after comparing studies to evaluate the average impact across studies of an independent variable on the dependent variable. Since measures may come from different settings and may be non-homogeneous, a standardized measure must be obtained for each experiment: these measures must be combined to estimate the global effect size of a factor. In our study, we considered that the usability inspection method was the main factor in the family of the experiments.

The meta-analysis was conducted by using the Meta-Analysis v2 tool [Biostat 2006]. We employed the mean value obtained using the WUEP method minus the mean value achieved when using the HE method to calculate the effect sizes for all the dependent variables (i.e., Effectiveness, Efficiency, Perceived Ease of Use, Perceived Satisfaction of Use) for each of the individual experiments, and these values were then used to obtain the Hedges' g metric [Hedges and Olkin 1985; Kampenes et al. 2007], which was used as a standardized measure. This measure expresses the magnitude of the effect of the method employed.

In order to obtain the overall conclusion, we calculated the Z-score based on the mean and standard deviation of the Hedges' g statistics of the experiments. More specifically, we used correlation coefficients, which provided the effect sizes that had a normal distribution (z_i) once they had been transformed by the Fisher transformation [Fisher 1915]. The global effect size was obtained by using the Hedges' g metric, whose weights were proportional to the experiment's size:

$$\bar{Z} = \frac{\sum_i w_i z_i}{\sum_i w_i} \qquad (2)$$

Where $w_i = 1/(n_i-3)$ and n_i is the sample size of the *i*-th experiment. The higher the value of Hedges' g, the higher the corresponding correlation coefficient is.

Table 11 summarizes the results of the meta-analysis: for each experiment, it reports the effect size, the values of the Hedges' g metric, and its significance. For studies in Software Engineering, the effect size is rated as small (0 to 0.37), medium (0.38 to 1), or large (above 1) [Kampenes et al. 2007] depending on the standardized difference between the two means m1 and m2. For example, an effect size of 0.5 indicates that m1 = m2 + (0.5 * d), where d is the standard deviation (i.e., a positive value signifies that WUEP achieved better results than HE in the dependent variable defined).

<< Insert Table 11 approximately here >>

For the reader's convenience, we show the meta-analysis results in diagram form by using a forest plot (or blobbogram). Figure 8 shows the four diagrams as provided by the tool used. On the left-hand side, the experiments are reported in chronological order from the top downwards. On the right-hand side, the effect of the Hedges' g metric is plotted for each experiment by a square whose dimensions are proportional to the weight of the experiment in the meta-analysis. The estimations for studies with a large sample size are more accurate, signifying that they make a greater contribution to the overall effect. The square size is proportional to the number of participants and the experiment effect size, and the square position with regard to the x axis indicates the Hedges' g value. The confidence intervals of each experiment are represented by the horizontal lines. Here we have considered a confidence interval of 95% for each experiment. The confidence interval [-1, 0] indicates a negative correlation, whereas the confidence interval [0, 1] indicates a positive correlation. The overall conclusion is represented by a diamond in the last row of the figure. In particular, the summary measure is the center line of the diamond, while the associated confidence interval is the lateral tips of the diamond.

<< Insert Figure 8 approximately here >>

The effect size obtained was large for the objective dependent variables (i.e., Effectiveness and Efficiency) and medium for the subjective dependent variables (i.e., Perceived Ease of Use and Perceived Satisfaction of Use). This was probably a result of the number of experiments used in the data meta-analysis. Despite the fact that the first experiment contributed to the overall results of the meta-analysis to a lesser extent, these results present a significant positive effect, and we can thus reject the null hypotheses which were formulated for each dependent variable (i.e., "there are no significant differences between WUEP and HE"). The meta-analysis therefore strengthens all the alternative hypotheses, providing promising results as regards WUEP's performance.

7.3 Threats to Validity

We must consider certain issues which may have threatened the validity of the family of experiments:

7.3.1 Internal Validity. The threats to internal validity are relevant in those studies that attempt to establish a causal relationship. In our case, the main threats to the internal validity of the family of experiments were: learning effect, subject experience, information exchange among participants, and understandability of the documents.

The learning effect was alleviated by ensuring that each participant applied each method to different experimental objects, and all the possible order combinations were considered. We also assessed the effect of order of method and order of experimental object by using statistical tests.

Subject experience was alleviated owing to the fact that none of the participants had any experience in usability evaluations. We confirmed this fact by asking the participants about their experience with usability evaluation methods. However, the training session may have affected the performance of the experiments, since the participants received the complete training immediately before the experimental tasks in the original experiment (EXP), whereas in the replications (REP1 and REP2) the participants received the complete training on the previous day. In order to alleviate this issue, we included a training slot before the experimental tasks in REP1 and REP2 in order to remind the participants of the employment of both inspection methods.

In order to minimize information exchange among participants, they were monitored by the experiment conductors to avoid communication biases while performing the tasks. However, this might have affected the results since the experiment took place over more than one day, and it is difficult to be certain whether the participants exchanged any information with each other. In order to alleviate this situation, at least to some extent, the participants were asked to return all the material at the end of each task. Moreover, since the participants in REP1 and REP2 were from the same Master's course but from different academic years, we ensured that no participants who were enrolled in REP1 were also enrolled in REP2.

Finally, understandability of the material was alleviated by clearing up all the misunderstandings that appeared in each experimental session.

7.3.2 External validity. This refers to the approximate truth of conclusions involving generalizations within different contexts. In our case, the main threats to the external validity of the family of experiments were: representativeness of the results and the size and complexity of the tasks.

The representativeness of the results might be affected by the evaluation design and the participant context selected. The evaluation design might have made an impact on the results owing to the selection of Web artifacts (experimental objects) and usability attributes to be evaluated during the design stage of WUEP. With regard to the selection of Web artifacts, we attempted to alleviate this by considering a set of artifacts with the same size and complexity, and which also contained representative artifacts of a Model-driven Web development process (i.e., navigational model, presentation model and final user interface). With regard to the selection of usability attributes, we attempted to alleviate this threat by considering a set of relevant usability attributes by involving Web usability experts in this decision. In order to alleviate these issues, we intend to evaluate more Web applications, and to carry out surveys to provide a predefined set of usability attributes to be evaluated in different Web application families (e.g., intranets, social networks, virtual marts) which will be useful as guidance for evaluator designers.

In addition, WUEP has been operationalized to be used in the context of a specific model-driven Web development method (OO-H). Consequently, our results can only be generalized to Web applications that follow a model-driven Web development process that is based on the OO-H method. Nevertheless, it can be considered a representative method of the whole set of model-driven Web development methods [Moreno and Vallecillo 2008]. The equivalence of the primitives of this method with regard to other model-driven Web development methods (e.g., UWE, WebML) is described in [Cachero et al. 2007]. By applying these guidelines, WUEP can easily be operationalized to evaluate the usability of Web artifacts obtained using other model-driven Web development methods.

Despite the fact that all the individual experiments were performed in an academic context (PhD and Master's students), the participants' performance could be considered to be representative of single-experienced evaluators (i.e., evaluators who have experience on the domain, but not in usability evaluations) since the kinds of students involved will be soon integrated into the industry's market. As further work, we are intended to conduct more experiments involving double-experienced evaluators (i.e., evaluators who have experience on the domain and in usability evaluations) in order to assess how the experience level would impact on the obtained results. In addition, since only internal replications were conducted, more external replications need to be conducted by other experimental conductors in other settings to confirm these results. In order to address the aforementioned limitations, these external replications will involve participants from different contexts and also from different levels of experience in Web usability evaluations.

The size and complexity of the tasks might have also affected the external validity. We decided to use relatively small tasks that would be applied in few representative Web artifacts since a controlled experiment requires participants to complete the assigned tasks in a limited amount of time.

7.3.3 Construct validity. The construct validity of the family of experiments may have been influenced by the measures that were applied in the quantitative analysis and the reliability of the questionnaire. We intended to alleviate the first threat by evaluating the dependent variables that are commonly employed in experiments in which usability inspection methods are involved. In particular, we employed the Effectiveness and Efficiency measures as suggested by Hartson et al. [2000] for formative evaluations (i.e., usability evaluations during the Web development process). These measures have also been employed in similar empirical studies [Conte et al. 2009]. In addition, the subjective measures employed were Perceived Ease of Use and Perceived Satisfaction of Use, based on the Technology Acceptance Model (TAM) [Davis 1989], a well-known and thoroughly validated model for evaluating information technologies.

The reliability of the questionnaires was tested by applying the Cronbach test. Table 12 shows the Cronbach's alpha obtained for each set of closed-questions intended to measure both subjective dependent variables (Perceived Ease of Use and Perceived Satisfaction of Use). All the values obtained were higher than the acceptable minimum threshold ($\alpha \geq 0.70$) [Maxwell 2002].

<< Insert Table 12 approximately here >>

7.3.4 Conclusion validity. The main threats to the conclusion validity of the family of experiments were the data collection and the validity of the statistical tests applied. With regard to the data collection, we applied the same procedure in each individual experiment in order to extract the data, and ensured that each dependent variable was calculated by applying the same formula. With regard to the validity of the statistical tests applied, we applied the most common tests that are employed in the empirical software engineering field owing to their robustness and sensitivity [Maxwell 2002].

8. CONCLUSIONS AND FUTURE WORK

The lack of usability evaluation methods that can properly be integrated into early stages of Web development processes motivated us to propose, in a previous work, the Web Usability Evaluation Process (WUEP) [Fernandez et al. 2011b] as an inspection method which can be instantiated and integrated into different model-driven Web development processes.

In this paper, we have reported the results of a family of experiments aimed at evaluating participants' effectiveness, efficiency, perceived ease of use, and perceived satisfaction of use when using WUEP in comparison to an industrial widely-used inspection method based on heuristics: Heuristic Evaluation (HE).

The results of the quantitative analysis showed up that WUEP was more effective and efficient than HE in the detection of usability problems in artifacts obtained from a specific model-driven Web development process (i.e., OO-H). These results were supported by a meta-analysis that was performed in order to aggregate empirical findings from each individual experiment. The low ratio of false positives obtained by

WUEP suggests that the use of metrics as part of the evaluation process reduces the degree of subjectivity in the evaluation of Web artifacts. The low ratio of replicated usability problems obtained by WUEP is owing to the metric operationalization, since metrics are applied in one type of Web artifact at a higher level of abstraction (analysis or design model) rather than in the final code. In addition, with regard to the evaluators' perceptions, the participants were more satisfied when they applied WUEP, and they also found it easier to use than HE.

The results of the qualitative analysis suggest that WUEP could be greatly improved with a tool that automates most of the tasks involved in the method, and would support the calculation of some metrics, also allowing the generation of usability reports.

From a research perspective, the family of experiments was a valuable means to obtain feedback with which to improve our Web Usability Evaluation Process. As far as we know, this is the first empirical study that provides evidence of the usefulness of a usability evaluation method for a model-driven Web development process. This empirical study is intended to contribute to Web Engineering research through its proposal of a well-defined framework that can be reused by other researchers in the empirical validation of their Web usability evaluation methods.

From a practical perspective, we are aware that this study provides only preliminary results on the usefulness of our Web Usability Evaluation Process in practice. Although the experimental results provided good results as regards the performance of our proposal as a usability inspection method for Web applications developed using model-driven development, these results need to be interpreted with caution since they are only valid within the context established in this family of experiments. There is a need for more empirical studies with which to test our proposal in other settings. Nevertheless, this study has value as a pilot study to test the integration of usability evaluations into model-driven Web development processes.

As future work, we intend to perform more replications in order to minimize the influence of the threats to validity identified. In particular, these replications will consider: different experimental designs conducted by external experimenters (i.e., external replications); new kinds of participants such as practitioners from industry with different levels of experience in usability evaluations; other Web artifacts from different model-driven Web development processes (e.g., WebML [Ceri et al. 2000], UWE [Koch and Kraus 2003]); and different kinds of Web applications, by also providing a predefined set of usability attributes to be evaluated which will be extracted from surveys involving usability and Web domain experts.

APPENDIX A: EXAMPLES OF BOTH USABILITY INSPECTION METHODS

This appendix presents brief examples of how both inspection methods can be applied in order to evaluate Web artifacts. Section A.1 shows an example of the WUEP execution stage when a metric is applied to a navigational model (Navigational Access Diagram from Object Oriented Hypermedia). Section A.2 shows an example of the HE execution stage when a heuristic is applied to a final user interface.

A.1 WUEP example

Let us suppose that we wish to evaluate the Cancel Support attribute (extracted from the Web Usability Model) in a Navigational Access Diagram (NAD0) which represents the contact management functionality of a Web application that has been developed by using Object Oriented Hypermedia (OO-H). A brief explanation of the modeling primitives of a NAD in OO-H is provided in Table A1 for the reader's convenience.

<< Insert Table A1 approximately here >>

Figure A1 presents the Web artifact that is going to be evaluated (NAD0). Users can retrieve the information concerning *all contacts* or they can search for a given contact by providing an *initial* or a search *string*. These functionalities are represented by the three navigational links that connect the *menu* collection and the *Contact Details* navigational class. Users can also create a new user by executing the *New* method in the *Create Contact* navigational class, and can modify an existing contact by executing the *Modify* method in the *Contact Details* navigational class.

<< Insert Figure A1 approximately here >>

Table A2 presents the operationalization of the *User Operation Cancellability* metric (associated with the *Cancel Support* attribute) in order for it to be applied to Navigational Access Diagrams (NADs) from Object Oriented Hypermedia (OO-H).

<< Insert Table A2 approximately here >>

As is shown in Figure A2, when the UOC metric is applied to NAD0, we obtain the value UOC (NAD) = 1/2 = 0.5, since only the Service Link associated with the Modify method provides a return Target Link to the previous navigation step. Considering the established thresholds in the metric operationalization, this is a medium usability problem (UP001) which is reported as shown in Table A3.

<< Insert Figure A2 approximately here >>

<< Insert Table A3 approximately here >>

A.2 HE example

Let us suppose that we wish to evaluate the final user interface (FUI0 presented in Figure A3) of the same Contact Management functionality previously mentioned in A.1 by applying the *Visibility of the System Status* heuristic.

<< Insert Figure A3 approximately here >>

The Visibility of the System Status (VSS) heuristic states that: "The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. The two most important things that users need to know at your site are probably *where am I?* and *where can I go next?*. So it is important to keep users informed about what is happening. To prove this, look for feedback on each user interaction. Make sure each page is branded and that you indicate which section it belongs to. Links to other pages should be clearly marked. Since users could be jumping to any part of your site from somewhere else, you need to include this status on every page. For example, when a user clicks on a 'Send' link in an order form, feedback is needed that tells you about whether your order has been received by the site. This information may appear as a different page or a popup that contains a link back to the main site".

As is shown in Figure A4, when the VSS heuristic is applied to FUI0, we detect four elements with the capability of providing feedback about the current status of the Web application. These elements are: the tabs that present the main menu, the textbox that shows the connected user, and both titles for the left menu and for the main content. Since the tabs do not provide feedback about which section has previously been selected and the title “List of contacts” does not provide feedback about what criteria was used to filter the contacts, we can consider the usability principle that is represented by the heuristic as partially supported by the Web artifact. There is therefore a usability problem that is reported as shown in Table A4. In this case, the severity level of the usability problem is based on the evaluator opinion. It is also important to note that, in a model-driven Web development approach, this usability problem may be corrected by considering the previous Web artifacts that were employed to automatically generate this final user interface (i.e., the Abstract Presentation Diagram that defines the user interface and the code generation rules that transform NADs and APDs into the final UI source code).

<< Insert Figure A4 approximately here >>

<< Insert Table A4 approximately here >>

APPENDIX B: EXCERPTS FROM THE EXPERIMENTAL MATERIAL

This appendix presents excerpts from all the different experimental materials. Section B.1 presents an excerpt from both the WUEP and HE appendixes that contain the operationalized metrics and heuristics to be applied, respectively. Section B.2 shows an example of a Web artifact to be evaluated: the Abstract Presentation Diagram (Web artifact APD1) for the Task Management functionality extracted from the Experimental Object 1 (which is included in the data gathering documents: WUEP-O1 and HE-O1). SectionB.3 collects the experimental tasks to be carried out when WUEP and HE are applied to APD1 (these tasks are also included in the data gathering documents: WUEP-O1 and HE-O1). Section B.4 shows the template which was employed to report usability problems in WUEP and HE. The original materials have been translated into English for the reader’s convenience. The original experimental material and the raw data are available for download at http://www.dsic.upv.es/~afernandez/JSS/familyexp.html.

B.1 Examples of operationalized metrics and heuristics

B.1.1. Operationalized Metrics.

Metric	Depth of the Navigation (DN)
Usability attribute	Appropriateness recognisability / Navigability/ Reachability
Generic description	Level of depth in the user navigation, in other words, the longest navigation path which is needed to reach any content/feature (without loops) from the Web app by the user.
Scale	Integer greater than 0
Interpretation	The higher the value, the more difficult it is for the user to reach the content/feature.
Operationalization	This metric can be calculated for each Navigational Access Diagram (NAD) by considering the number of navigation steps from the longest navigation path. Where: Navigation step: when a Target Link exists between two nodes (any modeling primitive and/or more than one modeling primitives connected by Automated Links and/or Source Links) Longest navigation path: The path with the greatest number of navigation steps, which begins in the first Navigational Class or Collection where the navigation starts, and which ends in the last Navigational Class or Service Link, from which it is not possible to reach another modeling primitive previously visited. The calculation formula is therefore: DN(NAD) = Number of navigation steps from the longest navigation path
Thresholds	[1 ≥ DN ≤ 4]: No usability problem.

	[5 ≤ DN ≤ 7]: Low usability problem. [8 ≤ DN ≤ 10]: Medium Usability Problem. [DN ≥ 10]: Critical Usability Problem.

Metric	Proportion of links without meaningful names (PLM)
Usability attribute	Learnability / Predictability / Meaningful links
Generic description	Ratio between the number of links without a meaningful name and the total number of links.
Scale	Ratio between 0 and 1.
Interpretation	The higher the value, the worse the predictability that is provided, since the user may experience difficulties in predicting the target and results of his/her actions.
Operationalization	This metric can be calculated in all the abstract pages belonging to an Abstract Presentation Diagram (APD) by considering the proportion of non-proper names used by APD links. The calculation formula is therefore: $PLM(APD) = \frac{\text{Number of Links without a meaningful name}}{\text{Total number of Links in the APD}}$
Thresholds	[PLM = 0]: No usability problem. [0 < PLM ≤ 0.3]: Low usability problem. [0.3 < PLM ≤ 0.6]: Medium Usability Problem. [0.6 < PLM ≤ 1]: Critical Usability Problem.

Metric	Headings according to the target of the links (HAT)
Usability attribute	Ease of use / Consistency / Heading consistency
Generic description	Number of headings whose name is not in accordance with the link name from which the heading was reached.
Scale	Integer greater than 0.
Interpretation	The higher the value, the worse the consistency that exists in the Web application content, thus affecting the ease of use.
Operationalization	This metric can be calculated in the final user interface (FUI) by considering the names of the links and the headings of the content reached by these links. The calculation formula is therefore: HAT(FUI) = Number of headings that are not in accordance with the link name which was followed to reach the current content.
Thresholds	[HAT = 0]: No usability problem. [1 ≤ HAT ≤ 3]: Low usability problem. [4 ≤ HAT ≤ 6]: Medium Usability Problem. [HAT ≥ 7]: Critical Usability Problem.

B.1.2. Heuristics.

Heuristic	Match between system and the real world.
Description	The system should speak the users' language, with words, phrases and concepts that are familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order. On the Web, you have to be aware that users will probably come from diverse backgrounds, so figuring out their "language" can be a challenge. An example of a real-world concept that is applied to Web applications may be the icons employed to distinguish between errors, warnings, or advice. Another example would be the shopping cart metaphor. In many Web stores, customers usually click once to select an element (equivalent to taking it off the shelf in a real store), click again to "add to cart" (equivalent to placing the item in their real cart) and then add a third click to confirm their purchase intention (equivalent to approaching the cashier in order to pay for it).

Heuristic	User control and freedom
Description	Users often choose some functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. It is important to provide control operations such as: cancel, undo and redo. Many of the "emergency exits" are provided by the browser, but there is still plenty of room on the site to support user control and freedom. Or, there are

	many ways authors can take away user control that are built into the Web. A "home" button on every page is a simple way to let users feel in control of the site. Be careful when forcing users into certain fonts, colors, screen widths or browser versions. And watch out for some of those "advanced technologies": user control is not usually added until the technology has matured. One example is animated GIFs. Until browsers let users stop and restart the animations, they can do more harm than good.

Heuristic	Recognition rather than recall
Description	Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate. Good labels and descriptive links are also crucial for recognition. It is best to always maintain links, menus, structures, actions and options visible to allow them to be memorized. For example, if a website has a lot of submenus, you should use a system that allows users to know which section you are at any time. This could be leaving a "trail of crumbs", or the Web application could use a color scheme that makes it possible to differentiate between the sections.

B.2 Example of a Web artifact to be evaluated

Figure B1 shows the Abstract Presentation Diagram (APD1) by including its six abstract pages. Detailed information about the content of these abstract pages is provided as follows. Elements marked with '(*)' are attributes from the *Navigational classes* and their display text in the final Web application will be the values from the attribute:

The first abstract page (Figure B1 (a)) represents the access to the different existing folders: predefined, created, user-specific. It contains:

– 1 label: *"Folder"*.
– 1 image: portfolio icon with one tick.
– 7 links: *"New folder", "All tasks", "Pending tasks", "Ended tasks", "Task out of date", "folder_name(*)", "user_name(*)".*

The second abstract page (Fig. B1 (b)) represents the task list which is filtered by the selected folder. It contains:

– 3 labels: "Task list", "folder_name ( )", "description (*)", "!", "Description", "End date"
– 2 images: folder icon, portfolio icon.
– 2 links: "New Task", "name and status (*)"

The third abstract page (Fig. B1 (c)) represents the warning message that appears when the selected folder does not contain any attached task.

– 1 label: "<b>NOTICE</b> The selected …"
– 1 image: exclamation icon

The fourth abstract page (Fig. B1 (d)) represents the detailed task information in conjunction with the available operations: attribute modification, ended percentage update, and user assignment:

– 21 labels: "Task detail", "EN1 (*)", Task title, Begin date, End Date, etc.
– 4 links: "aIe", "parent_folder (*)", Modify, Reassign.

The fifth abstract page (Fig. B1 (e)) represents the creation of a new task. Form fields refer to the attributes from the Task class that was defined in the Class Model:

– 7 labels: "New Task", "Task name", "description", "priority", "assigned user", "begin date", "End date (deadline)".
– 1 link: "New"

The sixth abstract page (Fig. B1 (f)) represents the creation of a new folder. Form fields refer to the attributes from the Folder class that was defined in the Class Model.

– 3 labels: “New Folder”, “Folder name”, “Folder description”.
– 1 link: “OK”

<< Insert Figure B1 approximately here >>

B.3 Examples of experimental tasks

B.3.1. Experimental tasks for applying WUEP to APD1.

1. Using as support the list of operationalized metrics:
    a. Select the metrics that can be applied to the APD that is shown in Figure B1.
    b. Apply each metric in order to obtain its value.
    c. Classify the value obtained according to the threshold established for each metric.
2. For each detected usability problem (low, medium, critical), fill in the required fields provided by the usability report template, and write the ID of the problem in the last column.

Write starting time (hh:mm): ________

Metric Acronym	Metric calculation	Severity level of the usability problem	Usability problem ID
…	…	…	…

Write finishing time (hh:mm): ________

B.3.2. Experimental tasks for applying HE to APD1.

1. Using the list of heuristics as support, identify whether the principles that are represented by each heuristic can be applied to the APD that is shown in Figure B1. If not, mark the “Not Applicable” box.
2. For each applicable heuristic, indicate the degree to which the represented principles are supported by the heuristic (YES=Supported; P=Partially supported; NO = Not supported). Justify your decision by indicating some elements from the artifact evaluated.
3. For each heuristic whose usability principles were not supported, fill in the usability problems detected in the usability report template, and write the ID of the problem in the last column.

Write starting time (hh:mm): ________

Heuristic ID	Usability principle represented	Justification by elements of the device ID observed usability problem	Usability problem ID
	□ Not Applicable □ YES □ P □ NO		
…	…	…	…

Write finishing time (hh:mm): ________

B.4 Examples of templates for reporting usability problems

B.4.1. Template for reporting usability problems in WUEP.

Fields to complete for each usability problem identified:

- Description: Textual description of the problem identified.
- Occurrences: Number of times the usability problem is repeated in the same Web artifact evaluated (if applicable).
- Recommendations: Guidance on how to prevent and/or correct the usability problem detected.

ID	P001
Description	
Occurrences	
Recommendations	

ID	P002
Description	
Occurrences	
Recommendations	

…

B.4.2. Template for reporting usability problems in HE tasks for applying HE to APD1.

Fields to complete for each usability problem identified:

- Description: Textual description of the problem identified.
- Occurrences: Number of times the usability problem is repeated in the same Web artifact evaluated (if applicable).
- Severity level: Classification of the usability problem: critical, medium or low.
- Recommendations: Guidance on how to prevent and/or correct the usability problem detected.

ID	P001
Description	
Severity level	□ Low □ Medium □ Critical
Occurrences	
Recommendations	

ID	P002
Description	
Severity level	□ Low □ Medium □ Critical
Occurrences	
Recommendations	

…

ACKNOWLEDGMENTS

The authors would like to thank all the participants in the experiments, along with the usability experts that supported certain tasks of the evaluation design stage, and of which the control group was composed.

REFERENCES

Abrahão, S., and Insfran, E. 2006. Early Usability Evaluation in Model-Driven Architecture Environments. In *Proceedings of the 6th IEEE International Conference on Quality Software (QSIC'06).* IEEE Computer Society, 287-294.

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

Abrahão, S., and Poels, G. 2009. A family of experiments to evaluate a functional size measurement procedure for Web applications. In *Journal of Systems and Software* 82, 2, 253-269.

Abrahão, S., Juristo N., Law, E., Stage, J. 2010. Interplay between usability and software development. Journal of Systems and Software 83 (11): 2015-2018.

Abrahão, S., Insfrán, E., Carsí, J.A, and Genero, M. 2011. Evaluating requirements modeling methods based on user perceptions: A family of experiments. In *Information Sciences* 181, 16, 3356–3378.

Allen, M., Currie, L., Bakken, S., Patel, V., and Cimino, J. 2006. Heuristic evaluation of paper-based Web pages: A simplified inspection usability methodology. In *Journal of Biomedical Informatics* 39, 4, 412-423.

Basili, V.R., Selby, R.W., and Hutchens, D.H. 1986. Experimentation in Software Engineering. In *IEEE Transaction on Software Engineering*12, 7, 733-743.

Basili, V.R., and Rombach, H.D. 1988. The TAME project: towards improvement-oriented software environments. In *IEEE Transactions on Software Engineering* 14, 6, 758–773.

Basili, V.R. 1993. The Experimental Paradigm in Software Engineering. In *Proceedings of the International Workshop, Experimental Software Eng. Issues: Critical Assessment and Future Directions*, LNCS 706, Springer.

Basili, V.R. 1996. The Role of Experimentation in Software Engineering: Past, Current, and Future. In *Proceedings of International Conference on Software Engineering (ICSE'96)*, 442-449.

Basili, V.R., Shull, F., and Lanubile, F. 1999. Building Knowledge through Families of Experiments. In *IEEE Transactions on Software Engineering* 25, 456-473.

Biostat 2006. Biostat Comprehensive Meta-Analysis v2. http://www.meta-analysis.com/

Blackmon, M.H., Kitajima, M., and Polson, P.G. 2005. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'05)*, 31-40.

Bolchini, D., and Garzotto, F. 2007. Quality of Web Usability Evaluation Methods: An Empirical Study on MiLE+. In *Proceedings of the International Workshop on Web Usability and Accessibility (IWWUA'07)*, 481-492.

Briand, L., Labiche, Y., Di Penta, M., and Yan-Bondoc, H. 2005. An experimental investigation of formality in UML-based development. In *IEEE Transactions on Software Engineering* 31, 10, 833–849.

Cachero, C., Melia, S., Genero, M., Poels, G., and Calero, C. 2007. Towards improving the navigability of Web applications: a model-driven approach. In *European Journal of Information Systems* 16, 420–447.

Casteleyn, S., Daniel, F., Dolog, P., and Matera, M. 2009. *Engineering Web Applications*, Springer.

Ceri, S., Fraternali, P., and Bongio, A. 2000. Web modeling language (WebML): a modeling language for designing Web sites. In *Proceedings of the 9th World Wide Web Conference (WWW'09)*, pp. 137–157.

Chattratichart, J., and Brodie, J. 2004. Applying user testing data to UEM performance metrics. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems (CHI'04)*, ACM press. 1119-1122.

Ciolkowski, M., Shull, F., and Biffl, S. 2002. A family of experiments to investigate the influence of context on the effect of inspection techniques. In *Proceedings of the 6th International Conference on Empirical Assessment in Software Engineering (EASE'02)*, 48–60.

Cockton, G., Lavery, D., and Woolrychn, A. 2003. Inspection-based evaluations. In *The Human-Computer Interaction Handbook*, 2nd edition, J.A. Jacko and A. Sears, Eds. Lawrence Erlbaum Associates, 1171-1190.

Colosimo, M., De Lucia, A., Scanniello, G., and Tortora, G. 2009.Evaluating Legacy System Migration Technologies through Empirical Studies. In *Information and Software Technology*, 51, 12, Elsevier, 433-447.

Conover, W. J. 1998. Practical Nonparametric Statistics, Wiley, 3rd edition.

Conte, T., Massollar, J., Mendes, E., and Travassos, G.H. 2009. Web usability inspection technique based on design perspectives. In *IET Software* 3, 2, 106-123.

Costabile, M.F., and Matera, M. 2001. Guidelines for hypermedia usability inspection. In *IEEE Multimedia* 8, 1,66-69.

Cruz-Lemus, J.A., Genero, M., Caivano, D., Abrahão, S., Insfrán, E., and Carsí, J.A. 2011. Assessing the influence of stereotypes on the comprehension of UML sequence diagrams: A family of experiments. In *Information and Software Technology* 53, 12, 1391-1403.

Davis, F.D. 1989. Perceived Usefulness, Perceived ease of use and user acceptance of information technology. In MIS Quarterly 13, 3, 319-340.

Dzidek, W. J., Arisholm, E., and Briand, L. C. 2008. A Realistic Empirical Evaluation of the Costs and Benefits of UML in Software Maintenance. In *IEEE Transactions on Software Engineering*, 34, 3, 407-432.

Fenton, N. 1993. How Effective Are Software Engineering Methods?. In *Journal of Systems and Software*, 22, 2,141-146.

Fernandez, A., Insfran, E., and Abrahão, S. 2009. Integrating a Usability Model into a Model-Driven Web Development Process. In *Proceedings of the 10th International Conference on Web Information Systems Engineering (WISE'09)*, LNCS 5802, Springer, 497-510.

Fernandez, A., Abrahão, S., and Insfran, E. 2010. Towards to the Validation of a Usability Evaluation Method for Model-Driven Web Development. In *Proceedings of 4th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'10)*, ACM, New York, NY.

Fernandez, A., Insfran, E., and Abrahão, S. 2011. Usability Evaluation Methods for the Web: A Systematic Mapping Study. In *Information and Software Technology*53, 789–817.

Fernandez, A., Abrahão, S., and Insfran, E. 2011. A Web Usability Evaluation Process for Model-Driven Web Development. In *Proceedings of the 23rd International Conference on Advanced Information Systems Engineering (CAiSE'11)*, Springer, 108-122.

Fisher, R.A. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. In *Biometrika*, 10, 4, 507–521.

Gellersen, H., Wicke, R., and Gaedke, M. 1997. Web Composition: an object-oriented support system for the Web engineering lifecycle. In *Computer Networks and ISDN Systems* 29, 8-13, 865-1553.

Glass, G.V., Mcgaw, B., and Smith, M.L. 1981. *Meta-Analysis in Social Research*. Sage Publications.

Gomez, J., Cachero, C., and Pastor, O. 2000. Extending a conceptual modeling approach to Web application design. In *Proceedings of the 12th International Conference on Advanced Information Systems Engineering (CAiSE'00)*, 79-93.

Gray, W.D., and Salzman, M.C. 1998. Damaged merchandise? A review of experiments that compare usability evaluation methods. In *Human-Computer Interaction* 13, 3, 203-261.

Hartson, R.H., Andre, T.S., and Williges, R.C. 2003. Criteria for Evaluating Usability Evaluation Methods. In *International Journal of Human-Computer Interaction* 15, 1, 145-181.

Hedges, L.V., and Olkin, I. 1985. *Statistical Methods for Meta-Analysis*. Academia Press.

Hertzum, M., and Jacobsen, N.E., 2001. The evaluator effect: a chilling fact about usability evaluation methods. In *International Journal of Human-Computer Interaction* 13, 421-443.

Hollingsed, T., and Novick, D.G. 2007. Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th annual ACM international conference on Design of communication (SIGDOC'07),* 249-255, ACM Press.

Hornbæk, K., and Frøkjær, E. 2004. Two psychology-based usability inspection techniques studied in a diary experiment. In *Proceedings of the 3rd Nordic conference on Human-computer interaction (NordiCHI '04),* 3-12.

Hornbæk, K., and Frøkjær, E. 2004. Usability Inspection by Metaphors of Human Thinking Compared to Heuristic Evaluation. In *International Journal of Human-Computer Interaction* 17, 3, 357-374.

Höst, M., Regnell, B., and Wohlin, C. 2000. Using students as subjects - a comparative study of students and professionals in lead-time impact assessment. In *Proceedings of the 4th Conference on Empirical Assessment and Evaluation in Software Engineering (EASE'00)*, 201–214.

Hu,P. J., and Chau, P.Y.K. 1999. Examining the Technology Acceptance Model Using Physician Acceptance of Telemedicine Technology. In *Journal of Management Information Systems* 16, 2, 91-113.

Hvannberg, E.T., Law, E., and Lárusdóttir, M. 2007. *Heuristic evaluation: Comparing ways of finding and reporting usability problems*. In *Interacting with Computers* 19, 2, 225-240.

Hwang, W., and Salvendy, G. 2010. Number of people required for usability evaluation: the 10±2 rule. In *Communications of the ACM* 53, 5, 130-133

ISO/IEC. 2001. ISO/IEC 9126-1 Standard, Software Engineering, Product Quality – Part 1: Quality Model.

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

Ivory, M.Y. 2001. An Empirical Foundation for Automated Web Interface Evaluation. PhD Thesis, University of California, Berkeley, Computer Science Division.

Juristo, N., and Moreno, A.M. 2001. *Basics of Software Engineering Experimentation*, Kluwer Academic Publishers.

Juristo, N., Moreno, A., and Sanchez-Segura, M.I. 2007. Guidelines for eliciting usability functionalities. In *IEEE Transactions on Software Engineering* 33, 11, 744-758.

Kampenes, V., Dybå, T., Hannay, J.E., and Sjøberg, D.I.K. 2007. A Systematic Review of Effect Size in Software Engineering Experiments. In *Information and Software Technology* 49, 11-12, 1073-1086.

Kitchenham, B.A., Pfleeger, S., Hoaglin, D.C., El Emam, K., and Rosenberg, J. 2002. Preliminary guidelines for empirical research in software engineering. In *IEEE Transactions on Software Engineering*28, 8, 721–734.

Kitchenham, B.A. 2008. The role of replications in empirical software engineering - a word of warning. In *Empirical Software Engineering* 13, 2, 219-221.

Koch, N., and Kraus, A. 2003. Towards a Common Metamodel for the Development of Web Applications. In *3rd International Conference on Web Engineering (ICWE'03),* LNCS 2722, Springer.

Koutsabasis, P., Spyrou, T., and Darzentas, J. 2007. Evaluating usability evaluation methods: criteria, method and a case study. In *Proceedings of the 12th international conference on Human-computer interaction: interaction design and usability (HCI'07),* 569-578.

Leavit, M., and Shneiderman, B. 2006. Research-Based Web Design & Usability Guidelines. U.S. Government Printing Office. http://usability.gov/guidelines/index.html

Malak, G., and Sahraoui, H. 2010. Modeling Web Quality Using a Probabilistic Approach: An Empirical Evaluation. In *ACM Transactions on the Web* 4, 3, Article 9, 31 pages.

Matera, M., Rizzo, F., and Carughi, G. 2006. Web usability: principles and evaluation methods. In *Web engineering*, E. Mendes and N. Mosley, Eds. Springer, 143–179.

Maxwell, K. 2002. *Applied Statistics for Software Managers*. Software Quality Institute Series, Prentice Hall.

Mendes, E. 2005. A Systematic Review of Web Engineering Research. In *Proceedings of the International Symposium on Empirical Software Engineering (ISESE'05)*, 498–507.

Molina, F., and Toval, A. 2009. Integrating usability requirements that can be evaluated in design time into Model Driven Engineering of Web Information Systems. In *Advances in Engineering Software* 40, 12, 1306-1317.

Moreno, N., and Vallecillo, A. 2008. Towards interoperable Web engineering methods. In *Journal of the American Society for Information Science and Technology* 59, 7, 1073–1092.

Nielsen, J. 1994. Heuristic evaluation. In Usability inspection methods, J. Nielsen and R.L. Mack. Eds. John Wiley & Sons, 25-62.

Nielsen, J. 2005.Ten best intranets of 2005. Jakob Nielsen's Alertbox. http://www.useit.com/alertbox/20050228.html.

Offutt, J. 2002. Quality Attributes of Web Software Applications. In *IEEE Software* 19, 2 (Special Issue on Software Engineering of Internet Software), 25-32.

Olsina, L., and Rossi, G. 2002. Measuring Web Application Quality with WebQEM. In *IEEE Multimedia* 9, 4, 20-29.

Olson, J.R., and Olson, G.M. 1990. The growth of cognitive modeling in human-computer interaction since GOMS. In *Human-Computer Interaction* 5, 221–265.

Panach, I., Condori, N., Valverde, F., Aquino, N., and Pastor, O. 2008. Understandability measurement in an early usability evaluation for model-driven development: an empirical study. In *Proceedings of the 2nd International Symposium on Empirical Software Engineering and Measurement (ESEM'08)*, 354-356.

Ricca, F., Di Penta, M., Torchiano, M., Tonella, P., and Ceccato, M. 2010.How Developers' Experience and Ability Influence Web Application Comprehension Tasks Supported by UML Stereotypes: A Series of Four Experiments. In *IEEE Transactions on Software Engineering* 36, 1, 96-118.

Rosenthal, R. 1986. *Meta-Analytic Procedures for Social Research*. Sage Publications.

Shull, F., Carver, J. C., Vegas, S., and Juristo, N. 2008. The role of replications in Empirical Software Engineering. In *Empirical Software Engineering* 13, 2, 211-218.

Sjøberg, D.I.K., Anda, B., Arisholm, E., Dybå, T., Jørgensen, M., Karahasanovic, A., and Vokác. M. 2003. Challenges and recommendations when increasing the realism of controlled software engineering experiments. In *Empirical Methods and Studies in Software Engineering Experiences from ESERNET 2001-2003*, LNCS 2765, 24-38.

Sjøberg, D.I.K., Hannay, J. E., Hansen, O., Kampenes, V. B., Karahasanovic, A., Liborg, N., and Rekdal, A. C. 2005. A Survey of Controlled Experiments in Software Engineering. In *IEEE Transaction on Software Engineering* 31, 9, 733-753.

Sottet, J., Calvary, G., Coutaz, J., and Favre, J. 2007. A model-driven engineering approach for the usability of plastic user interfaces. In *Proceedings of the Working Conference on Engineering Interactive Systems*, 140–157.

Ssemugabi, S., and De Villiers, R. 2007. A comparative study of two usability evaluation methods using a web-based e-learning application. In *Proceedings of the annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries (SAICSIT '07)*, 132-142.

Sutcliffe, A. 2002. Assessing the reliability of heuristic evaluation for Web site attractiveness and usability. *In Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)*, 1838-1847.

Sutton, J.A., Abrams, R.K., Jones, R.D., Sheldon, A.T., and Song, F. 2001. Methods for Meta-Analysis in Medical Research. John-Wiley & Sons.

Tan, D., and Bishu, R. 2009. Web evaluation: heuristic evaluation vs. user testing. In *International Journal of Industrial Ergonomics* 39, 621-627.

Valderas, P., and Pelechano, V. 2011. A survey of requirements specification in model-driven development of web applications. In *ACM Transactions on the Web*5, 2, Article 10, 51 pages.

Venkatesh V. 2000. Determinants of perceived ease of use: integrating control, intrinsic motivations, and emotion into the technology acceptance model. In *Journal Information Systems Research,*11, 4, 342–65.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C.,Regnell, B., Andwesslen, A. 2000. *Experimentation in Software Engineering - An Introduction*, Kluwer.

VITAE

Adrián Fernández Martínez: Assistant Researcher of the Department of Information Systems and Computation (DISC). He obtained a degree in Computer Science Engineering at Facultad Informática of Universitat Politècnica de València, (Spain, 2007) and a Master's degree in Software Engineering, Formal Methods and Information Systems at Department of Information Systems and Computation (Spain, 2009). He is currently a PhD student in the Doctorate program in Computer Science. His areas of interest are: Web Engineering, Software Quality, Empirical Software Engineering, and Model-driven Engineering. His main research work focuses on the development and empirical validation of a usability evaluation process based on Model-Driven Architecture (MDA) for Web application development.

Silvia Abrahão Gonzales: Associate professor of the Department of Information Systems and Computation (DISC) at Universitat Politècnica de València in Spain. She obtained a degree in Computer Science Engineering in 1995, a M.S degree in System Analysis sponsored by IBM-Brazil in 1995 and a PhD degree in Computer Science at the Universitat Politècnica de València in 2004. Currently, she is also a research affiliate at the Software Engineering Institute at Carnegie Mellon University, USA. Her research areas of interest include software quality, empirical software engineering, quality assurance in software product lines and Web engineering.

Emilio Insfran Pelozo: Associate Professor at the Department of Information Systems and Computation (DISC) of the Universitat Politècnica de València, Spain. He received a MSc degree in Computer Science from the Cantabria University (Spain, 1994) and a PhD degree from the Universitat Politècnica de València (Spain, 2003). He is currently a Research Affiliate at the Software Engineering Institute (SEI) of the Carnegie-Mellon University (USA), and formerly performed research stays at the University of Twente (the Netherlands) and at the Brigham Young University, Utah (USA). His main research interests are requirements engineering, software product lines, model-driven development, software quality, usability evaluation and software engineering environments and tools.

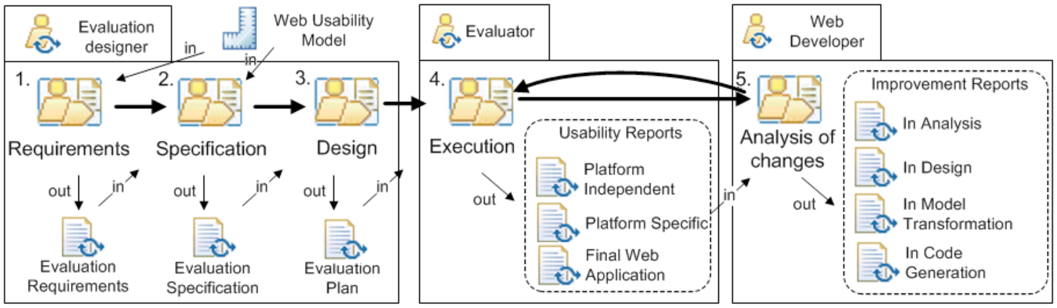


Fig. 1. Overview of the Web Usability Evaluation Process

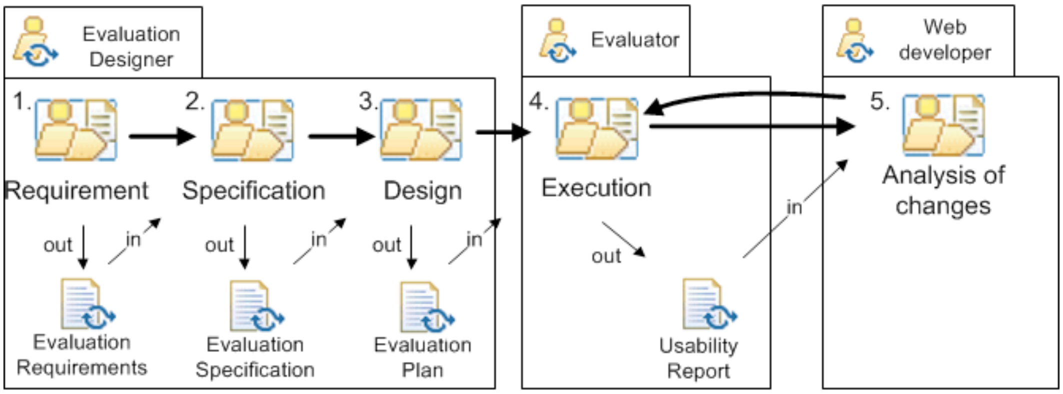
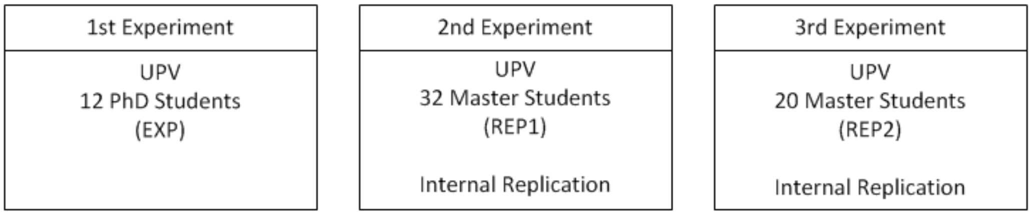


Fig. 2. Overview of the Heuristic Evaluation Process



Main factor: Method (i.e., WUEP vs HE)
Other factors: Experimental Objects (O1 and O2), Order of Experimental Objects, and Order of Method
Dependent variables: Effectiveness, Efficiency, Perceived Ease of Use, and Perceived Satisfaction of Use

Fig. 3. Overview of the family of experiments

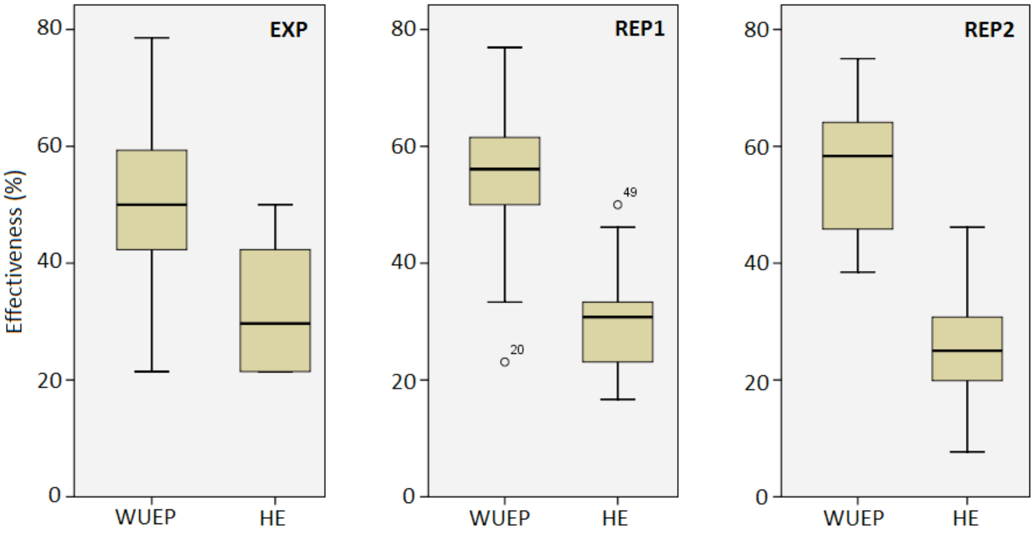


Fig. 4. Boxplots for the Effectiveness variable

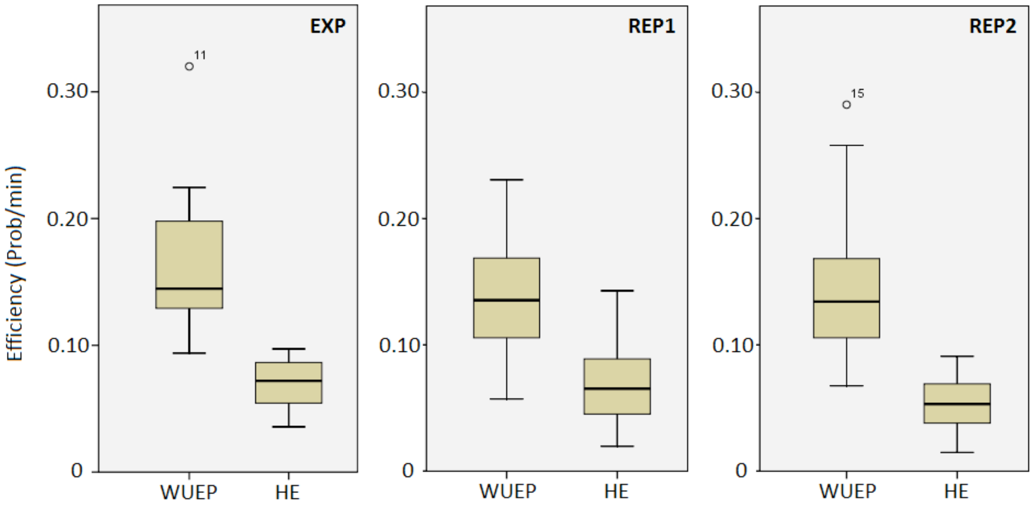


Fig. 5. Boxplots for the Efficiency variable

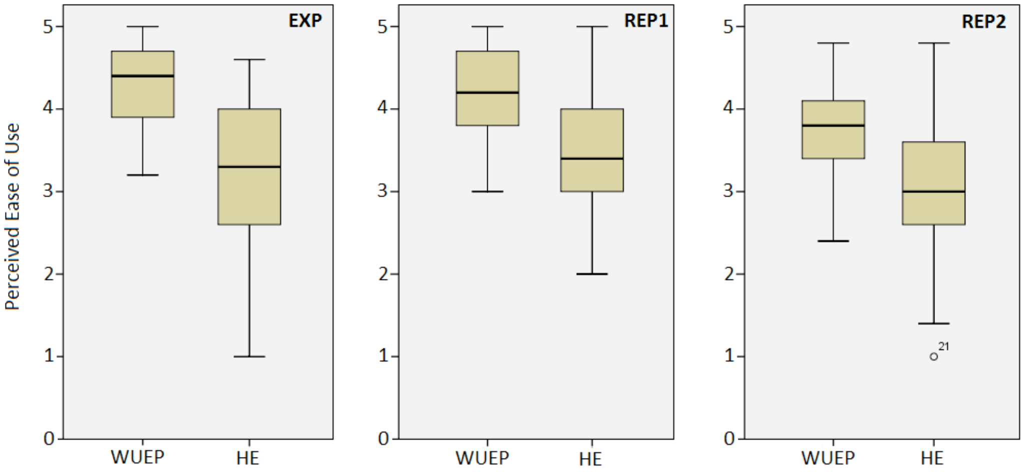


Fig. 6. Boxplots for the Perceived Ease of Use variable

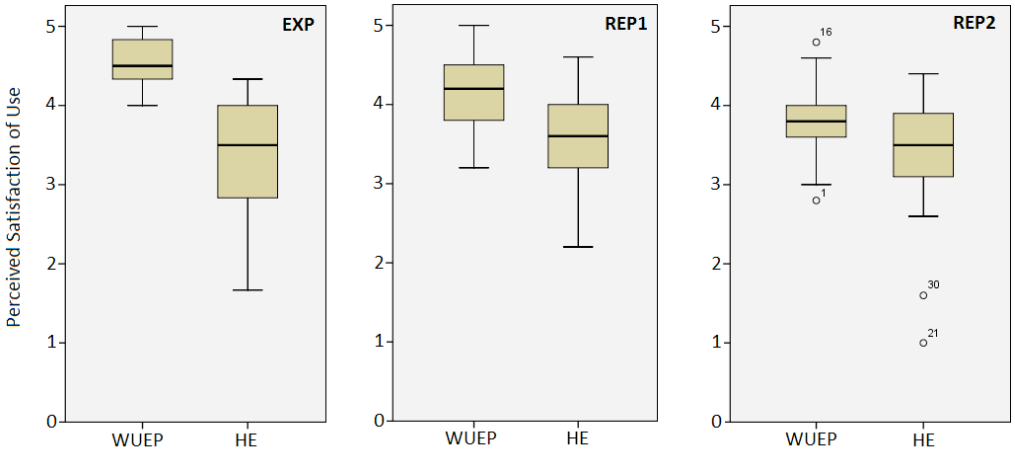


Fig. 7. Boxplots for the Perceived Satisfaction of Use variable

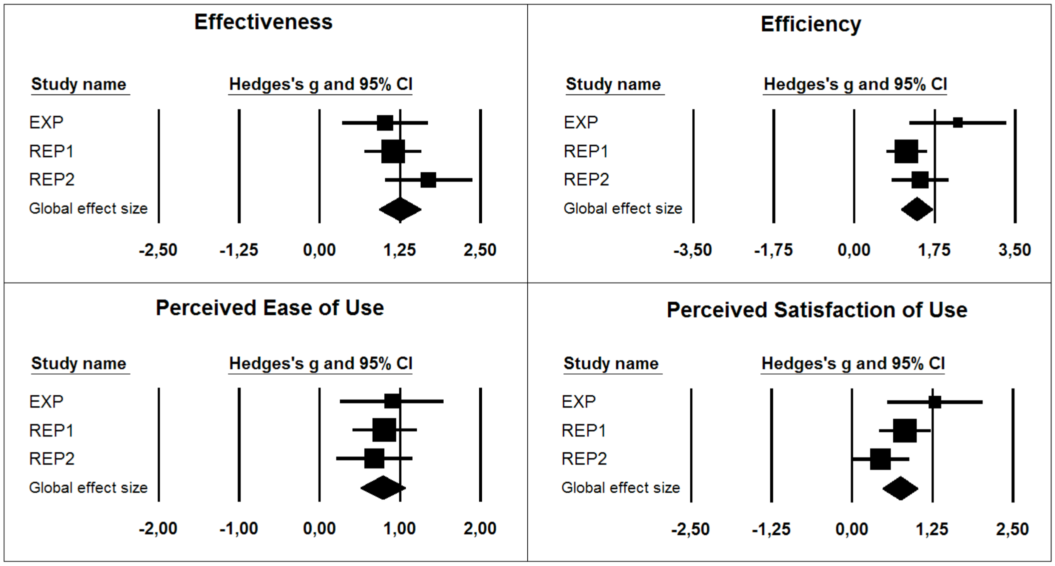


Fig. 8. Meta-analysis for all the dependent variables

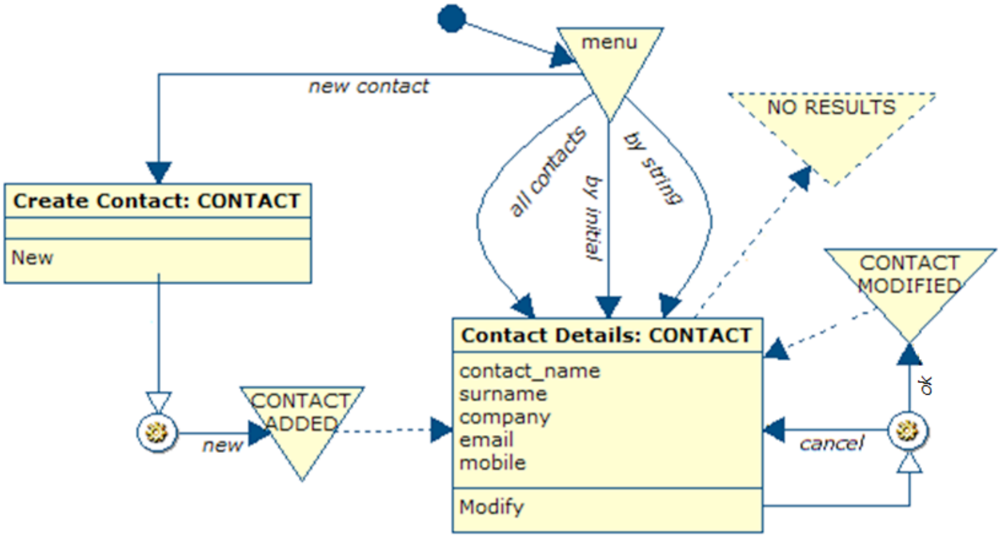


Fig. A1. Example of a NAD for contact management (NAD0)

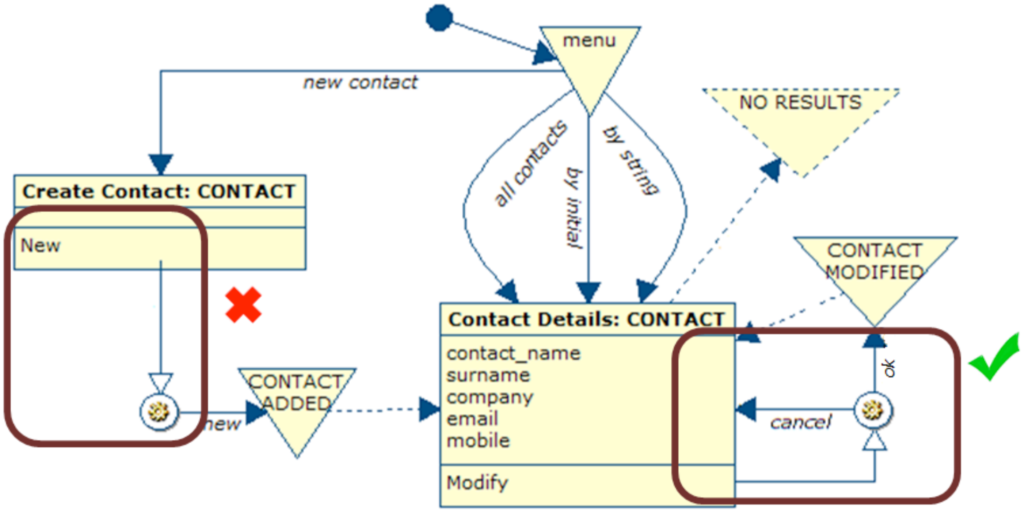
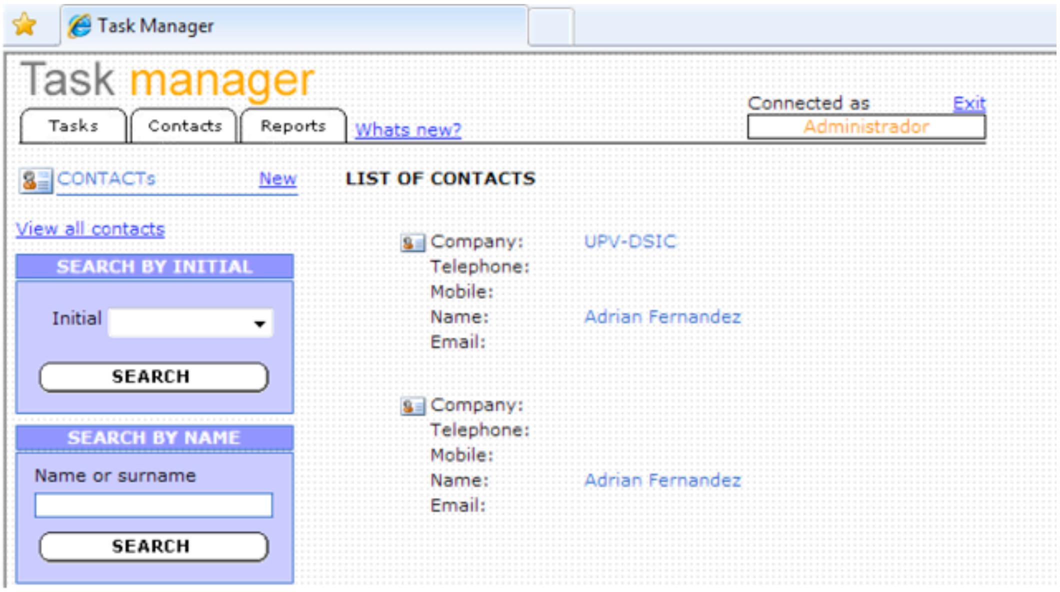


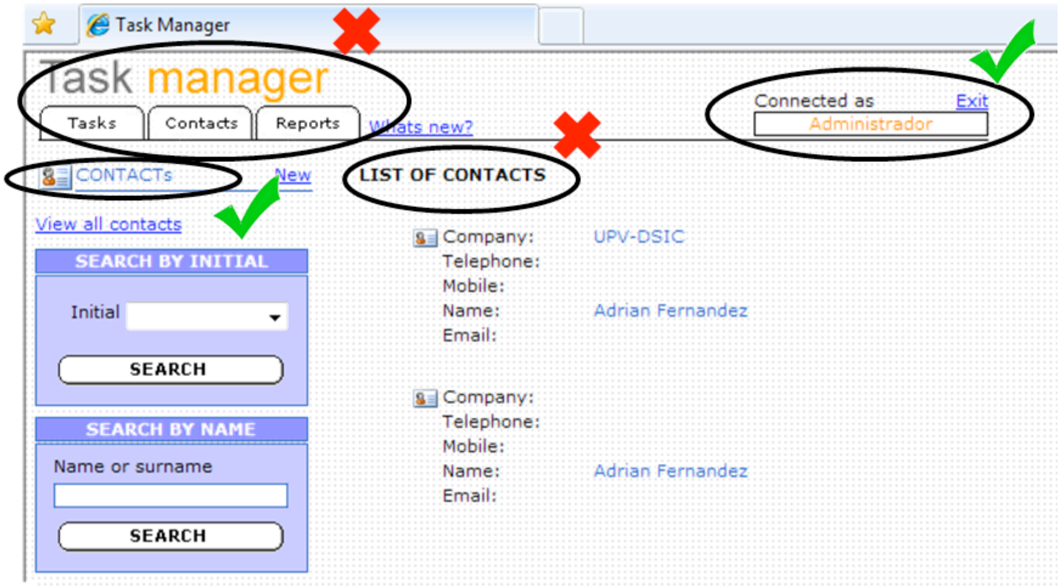
Fig. A2. Example of the UOC metric application to NAD0



Company: UPV-DSIC
Telephone:
Mobile:
Name: Adrian Fernandez
Email:

Company:
Telephone:
Mobile:
Name: Adrian Fernandez
Email:

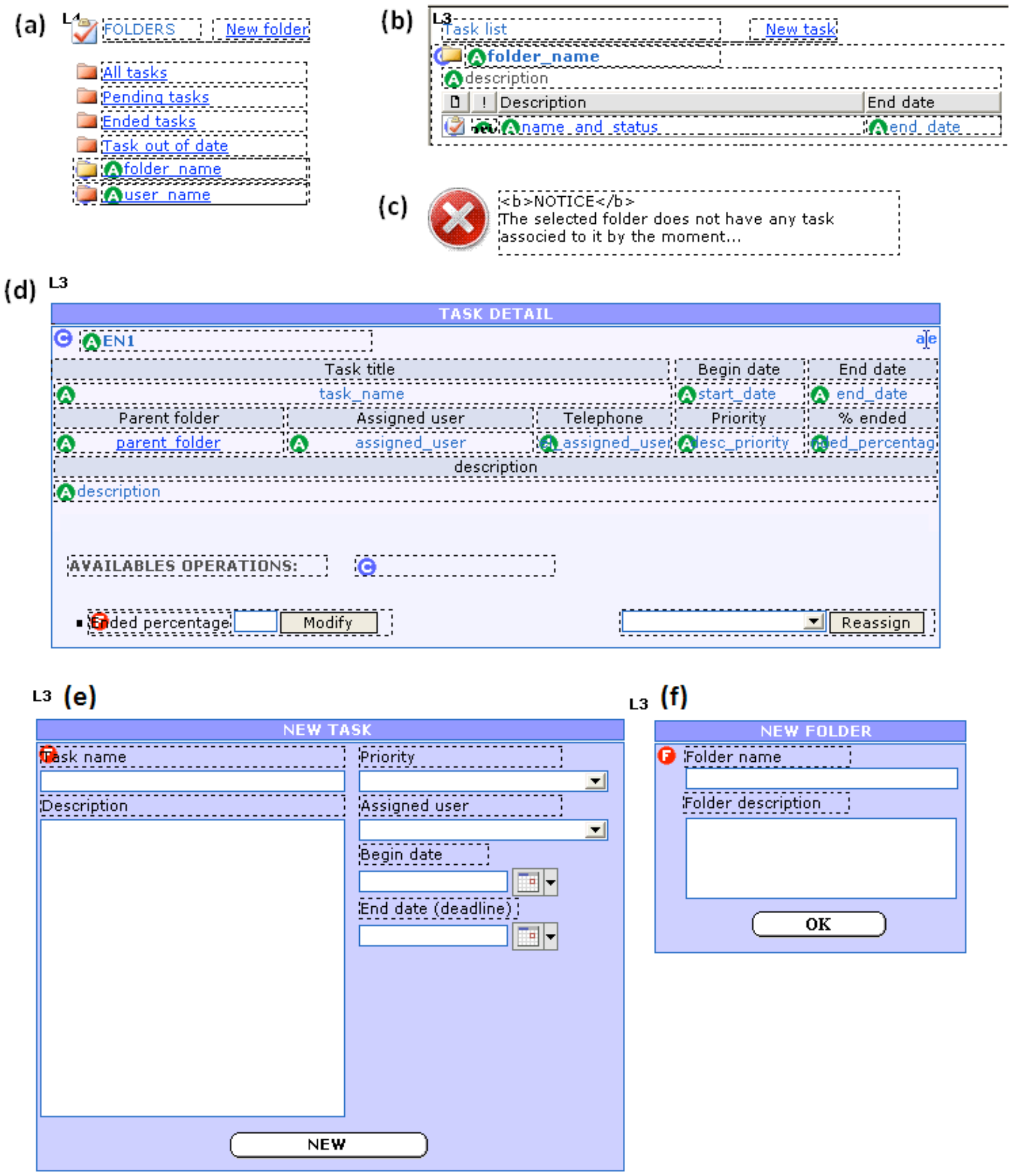
Fig. A3. Example of a FUI for contact management (FUI0)



Company: UPV-DSIC
Telephone:
Mobile:
Name: Adrian Fernandez
Email:

Company:
Telephone:
Mobile:
Name: Adrian Fernandez
Email:

Fig. A4. Example of the VSS heuristic application to FUI0



<b>NOTICE</b>
The selected folder does not have any task
associed to it by the moment...

Fig. B1. Example of the Web artifact (APD1)

Table 1. Closed-Questions to Evaluate Both Subjective Dependent Variables

Questions	Positive statement (5 points)	Negative Statement (1 point)
PEU1	The application procedure of the method is simple and easy to follow.	The application procedure of the method is complex and difficult to follow.
PEU2	I have found the evaluation method easy to learn.	I have found the evaluation method difficult to learn.
PEU3	In general terms, the evaluation method is easy to use.	In general terms, the evaluation method is difficult to use.
PEU4	The proposed metrics/heuristics are clear and easy to understand.	The proposed metrics/heuristics are confusing and difficult to understand.
PEU5	It was easy to apply the evaluation method to the Web artifacts.	It was difficult to apply the evaluation method to the Web artifacts.
PSU1	In general terms, I believe the evaluation method provides an effective manner with which to detect usability problems.	In general terms, I believe the evaluation method provides an ineffective manner with which to detect usability problems.
PSU2	The employment of the evaluation method would improve my performance in Web usability evaluations.	The employment of the evaluation method would not improve my performance in Web usability evaluations.
PSU3	I believe that it would be easy to be skillful in the use of the evaluation method.	I believe that it would be difficult to be skillful in the use of the evaluation method.

Table 2. Experimental Design Schema

	Groups (Sample size: 4*n* subjects)			
	G1(*n* subjects)	**G2** (*n* subjects)	**G3**(*n* subjects)	**G4** (*n* subjects)
1st Session	WUEP applied in O1	WUEP applied in O2	HE applied in O1	HE applied in O2
2nd Session	HE applied in O2	HE applied in O1	WUEP applied in O2	WUEP applied in O1

Table 3. Experimental Objects

Experimental Object	**User**	**Functional Feature**	**Use Cases**	**Web Artifacts to be evaluated**
O1	Project Manager	Task Management	Create/Modify/ Delete tasks, Categorize tasks, etc.	1 Navigational Access Diagram (NAD1)
				1 Abstract Presentation Diagram (APD1)
				1 Final User Interface (FUI1)
O2	Software Programmer	Report Management	Create daily reports, Access to partner reports, etc.	1 Navigational Access Diagram (NAD2)
				1 Abstract Presentation Diagram (APD2)
				1 Final User Interface (FUI2)

Table 4. Hypotheses to Test the Influence in the Order of Independent Variables

Dependent variables	**Order of Methods**	**Order of Experimental Objects**
Effectiveness	HM1$_0$: Effec_Diff(WUEP) = Effec_Diff(HE)	HO1$_0$: Effec_Diff(O1) = Effec_Diff(O2)
	HM1$_a$: Effec_Diff(WUEP) ≠ Effec_Diff(HE)	HO1$_a$: Effec_Diff(O1) ≠ Effec_Diff(O2)
Efficiency	HM2$_0$: Effic_Diff(WUEP) = Effic_Diff(HE)	HO2$_0$: Effic_Diff(O1) = Effic_Diff(O2)
	HM2$_a$: Effic_Diff(WUEP) ≠ Effic_Diff(HE)	HO2$_a$: Effic_Diff(O1) ≠ Effic_Diff(O2)
Perceived Ease of Use	HM3$_0$: PEU_Diff(WUEP) = PEU_Diff(HE)	HO3$_0$: PEU_Diff(O1) = PEU_Diff(O2)
	HM3$_a$: PEU_Diff(WUEP) ≠ PEU_Diff(HE)	HO3$_a$: PEU_Diff(O1) ≠ PEU_Diff(O2)
Perceived Satisfaction of Use	HM4$_0$: PSU_Diff(WUEP) = Effec_Diff(HE)	HO4$_0$: PSU_Diff(O1) = PSU_Diff(O2)
	HM4$_a$: PSU_Diff(WUEP) ≠ Effec_Diff(HE)	HO4$_a$: PSU_Diff(O1) ≠ PSU_Diff(O2)

Table 5. Planning for the Original Experiment (EXP)

	1st Day		**2nd Day**	
Id. Group	G3 (3 subjects)	G4 (3 subjects)	G1 (3 subjects)	G2 (3 subjects)
Training (15+20 minutes)	OO-H Introduction			
	Training with HE		Training with WUEP	
1st Session (90 minutes)	HE in O1	HE in O2	WUEP in O1	WUEP in O2
	Questionnaire for HE		Questionnaire for WUEP	
Break (180 minutes)				
Training (20 minutes)	Training with WUEP		Training with HE	
1st Session (90 minutes)	WUEP in O1	WUEP in O2	HE in O2	HE in O1
	Questionnaire for HE		Questionnaire for WUEP	

Table 6. New Closed-Questions Added to the Questionnaire

Questions	Positive statement (5 points)	Negative Statement (1 point)
PSU4	I believe the evaluation method helps to improve my skills in Web usability evaluation.	I do not believe the evaluation method helps to improve my skills in Web usability evaluation.
PSU5	I am satisfied with the use of the evaluation method, to the point that I would recommend its use in the evaluation of Web applications	I am not satisfied with the use of the evaluation method, to the point that I would not recommend its use in the evaluation of Web applications

Table 7. Planning for the Second Experiment (REP1)

	Groups			
	G1 (9 subjects)	G2 (10 subjects)	G3 (10 subjects)	G4 (9 subjects)
1st Day (60 minutes)	OO-H Introduction			
	Training with HE			
	Training with WUEP			
2nd Day (30 + 90 minutes)	OO-H Introduction			
	Training with WUEP			
	Training with HE			
	WUEP in O1	WUEP in O2	HE in O1	HE in O2
	Questionnaire for WUEP		Questionnaire for HE	
3rd Day (30 + 90 minutes)	OO-H Introduction			
	Training with HE			
	Training with WUEP			
	HE in O2	HE in O1	WUEP in O2	WUEP in O1
	Questionnaire for HE		Questionnaire for WUEP	

Table 8. Overall Results of the Usability Evaluations

		EXP (N=12)		REP1 (N=32)		REP2 (N=20)	
Statistics	Method	Mean	SD	Mean	SD	Mean	SD
Number of problems per subject	HE	4.25	1.40	3.81	1.06	3.30	1.22
	WUEP	**7.00**	2.21	**6.88**	1.64	**7.05**	1.47
False positives per subject	HE	2.08	2.15	2.28	1.57	2.50	1.76
	WUEP	**0.00**	0.00	**0.66**	0.60	**0.40**	0.60
Replicated problems per subject	HE	1.41	0.79	1.72	1.65	2.25	1.48
	WUEP	**0.00**	0.00	**0.00**	0.00	**0.10**	0.31
Duration (min)	HE	61.83	14.43	61.28	19.33	63.50	10.89
	WUEP	**44.16**	13.53	**53.56**	13.81	**53.50**	15.17
Effectiveness (%)	HE	31.63	10.89	30.53	08.63	26.28	09.13
	WUEP	**51.83**	16.09	**54.91**	12.49	**56.41**	11.45
Efficiency (Prob. / min)	HE	0.07	0.02	0.07	0.03	0.05	0.02
	WUEP	**0.17**	0.06	**0.14**	0.05	**0.14**	0.06
Perceived Ease of Use	HE	3.23	1.01	3.44	0.70	3.03	0.89
	WUEP	**4.25**	0.57	**4.16**	0.61	**3.73**	0.56
Perceived Satisfaction of Use	HE	3.36	0.84	3.56	0.64	3.32	0.84
	WUEP	**4.52**	0.36	**4.18**	0.47	**3.82**	0.49

Table 9. *p*-values obtained for the Influence of Order of Methods and Experimental Objects

Order of	Dependent variable	EXP	REP1	REP2
Methods	Effectiveness	No (0.161)	No (0.166)	No (0.275)
	Efficiency	No (0.846)	No (0.769)	No (0.536)
	Perceived Ease of Use	No (0.871)	No (0.672)	No (0.350)
	Perceived Satisfaction of Use	No (0.339)	No (0.160)[1]	No (0.579)[1]
Experimental Objects	Effectiveness	No (0.394)	No (0.642)[1]	No (0.664)
	Efficiency	No (0.910)	No (0.882)	No (0.709)
	Perceived Ease of Use	No (0.908)	No (0.734)	No (0.454)
	Perceived Satisfaction of Use	No (0.514)	No (0.270)[1]	No (0.419)

[1] Result obtained with the Mann-Whitney non-parametric test

Table 10. Summary of the Results of the Family of Experiments

Experiment	Type of subjects	Num. of subjects	Hypotheses accepted	Influence of method order	Influence of object order
EXP	PhD Students	12	$H1_a$, $H2_a$, $H3_a$, and $H4_a$	No	No
REP1	Master's Students	32	$H1_a$, $H2_a$, $H3_a$, and $H4_a$	No	No
REP2	Master's Students	20	$H1_a$, $H2_a$, $H3_a$, and $H4_a$	No	No

Table 11. The Hedges' metric values for all the dependent variables

Dependent variable	Experiment	Effect Size (Hedges' g)	Significance (*p*-value)
Effectiveness	EXP	Large (1.022)	Yes (p = 0.003)
	REP1	Large (1.146)	Yes (p < 0.001)
	REP2	Large (1.697)	Yes (p < 0.001)
	Global Effect Size	Large (1.243)	Yes (p < 0.001)
Efficiency	EXP	Large (2.261)	Yes (p < 0.001)
	REP1	Large (1.146)	Yes (p < 0.001)
	REP2	Large (1.443)	Yes (p < 0.001)
	Global Effect Size	Large (1.352)	Yes (p < 0.001)
Perceived Ease of Use	EXP	Medium (0.904)	Yes (p = 0.006)
	REP1	Medium (0.811)	Yes (p < 0.001)
	REP2	Medium (0.682)	Yes (p = 0.005)
	Global Effect Size	Medium (0.785)	Yes (p < 0.001)
Perceived Satisfaction of Use	EXP	Large (1.294)	Yes (p < 0.001)
	REP1	Medium (0.825)	Yes (p < 0.001)
	REP2	Medium (0.451)	Yes (p = 0.046)
	Global Effect Size	Medium (0.747)	Yes (p < 0.001)

Table 12. Cronbach's alphas for the reliability of questionnaires

Dependent variable	EXP	REP1	REP2
Perceived Ease of Use	Acceptable (0.909)	Acceptable (0.762)	Acceptable (0.842)
Perceived Satisfaction of Use	Acceptable (0.802)	Acceptable (0.780)	Acceptable (0.785)

Table A1. Some modeling primitives of a NAD from OO-H

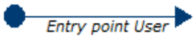
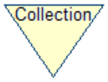
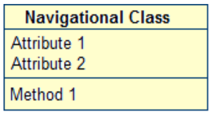
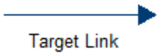
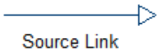
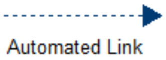
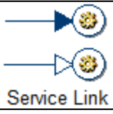
Modeling primitive	Meaning
	Each NAD has a unique *Entry Point User* that indicates the starting point of the navigation process.
	A *Collection* is a hierarchical structure that groups a set of navigational links. It is an abstraction from the menu concept.
 Navigational Class Attribute 1 Attribute 2 Method 1	A *Navigational Class* represents a view of a set of attributes and methods in a class from the UML class diagram that defines the content and the static structure of the Web application.
	A *Target Link* represents that the target node is reachable by explicit user navigation. (Depicted as a bold arrow)
	A *Source Link* represents that the target node is reachable in the same navigation step in which the source node was reached. (Depicted as an empty arrow)
	An *Automated Link* represents that the target node is reachable with no need for user navigation. (Depicted as an arrow with a broken line)
	A *Service Link* represents the execution of a method from a navigational class.(Depicted as an *Target Link* or *Source Link* with a gear icon)

Table A2. User Operation Cancellability Metric

Metric	User Operation Cancellability (UOC)
Usability attribute	Operability / Controllability / Cancel support
Generic description	Proportion between the number of operations provided that cannot be cancelled by the user prior to completion and the total number of operations requiring the pre-cancellation capability.
Scale	Ratio between 0 and 1.
Interpretation	The higher the value, the worse the controllability that appears in the WebApp owing to the fact that it is necessary to use external operations (Web browser actions) in order to return to a previous state if the user wishes to cancel the current operation.
Operationalization	This metric can be calculated for each NAD by considering the *Service Links* that have been associated with the *Navigational Classes* methods as user operations. These methods provide the cancellation if a *TargetLink* exists that returns from the *Service Link* to the previous navigation step. The calculation formula is therefore: $OC(NAD) = \frac{\text{Number of Service Links without a return Target Link}}{\text{Total number of Service Links}}$
Thresholds	[UOC = 0]: No usability problem. [$0 < UOC \leq 0.3$]: Low usability problem. [$0.3 < UOC \leq 0.6$]: Medium Usability Problem. [$0.6 < UOC \leq 1$]: Critical Usability Problem.

Table A3. Usability report for usability problem UP001

ID	P001
Description	The operation “create a new contact” cannot be cancelled by the user.
Usability attribute	Ease of use / Controllability/ Cancel support
Severity level	Medium
Evaluated artifacts	NAD (Navigational Access Diagram)
Problem source	NAD (Navigational Access Diagram)
Occurrences	1 Service Link without a return Target Link
Recommendations	Add a new *Target Link* between the *Service Link* and the *Create Contact navigational class* in order to support the cancellation.

Table A4. Usability report for usability problem UP002

ID	P002
Description	- The tabs do not provide feedback about which section has been previously selected. - The title “List of contacts” does not provide feedback about what criteria was used to filter the contacts.
Heuristic applied	Visibility of the System Status
Severity level	Medium
Evaluated artifacts	FUI (Final User Interface)
Source problem	APD (Abstract Presentation Diagram) and Code generation rules from NAD and APD to FUI
Occurrences	2 Elements that do not provide proper feedback about the Web application status.
Recommendations	- Replace the “List of contacts” title with another that is more specific (in the related APD). - Replace the code generation rule that provides the tabs in the FUI source code with another which can show the current selection.