A Systematic Review on the Effectiveness of Web Usability Evaluation Methods

Adrian Fernandez, Silvia Abrahão, Emilio Insfran
ISSI Research Group, Department of Information Systems and Computation
Universitat Politècnica de València, 46022, Valencia, Spain
{afernandez, sabrahao, einsfran}@dsic.upv.es

Background: Usability evaluation methods have become critical in the Web domain to ensure the success of Web applications. Aim: Since a large number of proposals have been presented during the last few years, a question arises: Which usability evaluation methods have proven to be the most effective in the Web domain? Method: This paper presents a systematic review that was motivated by previous results obtained from a systematic mapping study in the Web usability evaluation field. Results: A total of 18 studies were selected from an initial set of 206 in order to extract, code, and synthesize empirical data concerning the effectiveness of usability evaluation methods for the Web. Conclusions: We detected a need of more empirical studies and more standardized effectiveness measures for comparing usability evaluation methods. Our results suggest several evaluation methods which may be useful in allowing researchers and practitioners to perform effective Web usability evaluations.

Keywords: Usability Evaluation; Web Development; Method Effectiveness; Systematic Review.

I. MOTIVATION

Usability is considered to be one of the most important quality factors for Web applications [14]. Usability evaluation methods (UEMs) which are specifically crafted for the Web have therefore become critical for Web development companies to ensure the success of their products. This has led to the need to select which UEMs are suitable in each specified context. This issue has been addressed in several studies aimed at comparing UEMs for Web development (e.g., [1]). These studies often compare a limited number of evaluation methods, and the selection of methods is normally driven by the researcher’s expectations. There is thus a need for a more systematic identification of those UEMs which have proven to be the most effective when applied to Web development in comparison to others (i.e., which UEMs obtained better results in the detection of usability problems).

In previous research [6], we conducted a systematic mapping study in order to investigate what usability evaluation methods have been employed to evaluate Web artifacts, and how have these methods been used. This research question was used to construct a search string by including synonyms and variations of the terms: Web, usability, and evaluation in order to retrieve potential papers.

After applying inclusion and exclusion criteria, a total number of 206 selected papers were classified by considering several data extraction criteria: origin of the UEM, underlying usability definition; type of UEM; type of evaluation performed by the UEM; phase(s) and Web artifacts in which it is applied; feedback provided by the UEMs; and type of empirical study used to validate the UEM.

Upon considering the knowledge obtained from our systematic mapping study, more concrete research questions related to the empirical validations of UEMs arose, such as how many individual Web usability evaluation methods have been validated and which usability evaluation methods have proven to be the most effective in the Web domain. Since the goal of systematic mapping studies is more oriented towards categorizing the selected papers at a high level of granularity and identifying representative studies than performing evidence aggregations of empirical results, the results of papers presenting empirical validations should be analyzed by considering a systematic review protocol [2][10].

This paper presents a systematic review whose aim is to analyze which usability evaluation methods have proven to be the most effective in the Web domain. The papers selected from our previous systematic mapping study were used as potential papers to be included in the review.

This paper is structured as follows. Section 2 describes the process followed in order to perform the systematic review. Section 3 discusses the results obtained. Finally, Section 4 presents our conclusions and further work.

II. RESEARCH METHOD

We performed a systematic review by considering the guidelines that are provided in [10]. The following subsections describe its stages: establishment of the research question, search strategy, selection of primary studies, quality assessment, data extraction, and synthesis strategy.

A. Establishment of the Research Question

The goal of our study is to examine the effectiveness of usability evaluation methods in Web development from the point of view of the following research question: “Which usability evaluation methods have proven to be the most effective in the Web domain?” This will allow us to aggregate the current empirical knowledge to provide useful
information for researchers and practitioners in the selection of UEMs in Web development projects. As suggested by guidelines for performing systematic reviews [10][15], the research question has been structured by following the PICOC criteria:

- **Population**: Web applications.
- **Intervention**: Usability evaluation methods (UEM).
- **Comparison**: Different usability evaluation methods.
- **Outcome**: Effectiveness of the UEM.
- **Context**: Research papers.

### B. Search Strategy and Selection of Primary Studies

In these stages, we reused the set of 206 papers included in our previous systematic mapping study as the potential set of papers to be included in the review. This rationale was based on the fact that reutilization is possible since our research question is a specialization of our previous systematic map’s research question. In fact, composing a new search string including terms such as effectiveness and comparison may considerably restrict the set of relevant papers. These 206 papers were obtained after applying a validated search strategy in relevant bibliographic sources from the years 1996 to 2009, along with several inclusion and exclusion criteria in order to obtain a relevant set of papers concerning the use of UEMs in the Web domain. Further details of this review protocol can be found in [6].

The initial set of 206 papers was evaluated by the authors of this paper in order to decide whether or not each paper should be included as a primary study. The discrepancies were solved by consensus. The studies that met both of the following inclusion criteria were included:

- Papers presenting surveys, case studies, or experiments concerning the empirical validation of usability evaluation methods. These kinds of studies are the most representative ones to gather empirical data [5].
- Papers comparing the effectiveness of two or more usability evaluation methods. We selected this kind of studies since comparisons among UEMs allow empirical data aggregation from different sources.

After applying these inclusion criteria, a total of 28 studies were selected. The reliability of inclusion of a candidate study in the systematic review was assessed by applying Fleiss’ Kappa as an agreement measure [7]. We asked three independent raters to classify a random sample of 20 studies, 10 of which had previously been included in the systematic review and 10 of which had not. The Fleiss’ kappa obtained was 0.96, which indicates an acceptable level of agreement among raters.

### C. Quality Assessment

A three-point Likert-scale questionnaire was designed to provide a quality assessment of the selected empirical studies as suggested by [10]. This quality assessment was performed independently by the three authors and its objective was to ensure, at least to some extent, that our results would be based on good quality empirical studies. The questionnaire contained five subjective closed-questions:

1. Is the paper based on research and is not merely a “lessons learned” report based on expert opinion?
2. Is there a clear statement of the aims of the research?
3. Is there an adequate description of the context in which the research was carried out?
4. Is there an adequate description of the usability evaluation methods to be compared?
5. Is there an adequate description of the measures intended to assess the UEM effectiveness?

The first three questions, which were extracted from the questionnaire proposed in Dybå and Dingsøyr [4], are based on principles of good practice for conducting empirical research in Software Engineering [11]. The others were specifically crafted for our review with the aim to assess the quality of the data provided to researchers and practitioners. The possible answers to these questions were: “Yes (+1)”, “Borderline (0)”, and “No (-1)”. Each of the studies selected had a score for each closed-question which was calculated as the arithmetic mean of all the individual scores from each reviewer. The sum of the five closed-question scores of each study provided a final score (an integer between -5 and 5). Papers with a total score of less than or equal to 3 were excluded from the review. This threshold was arbitrarily established with the aim to select high-quality papers which have obtained; at least, three closed-questions with the maximum score and the other two with border-line score.

After applying the quality assessment, a total of 18 studies were finally selected to be included in the review. The complete list of selected studies is shown in Annex A, whereas the intermediate results are available for perusal at http://www.dsic.upv.es/~afernandez/EASE12/results.html.

### D. Data Extraction and Synthesis Strategy

We extracted the following information for each of the studies selected:

- **a)** The aim and type of the empirical study.
- **b)** The usability evaluation methods that were evaluated and their type of method based in the taxonomy proposed in [16]: Testing, Inspection, Inquiry, Analytical modeling, and Simulation.
- **c)** The measures that were employed to assess the effectiveness of the usability evaluation methods.
- **d)** The Web artifacts that were evaluated (e.g., conceptual models, mock-ups, prototypes, final application).
- **e)** The context of the empirical study (e.g., participant or evaluators profile, number of participants or evaluators).

The data extracted was coded to facilitate the interpretation of empirical evidence from different empirical studies. We followed an aggregation strategy similar to that presented in Dieste et al. [3]. The studies selected were coded as Sx, effectiveness measures were coded as M, (where ‘x’ and ‘i’ signify sequential numbers), and the UEMs used in the experiments were coded with acronyms. Once all the data had been identified and coded, we built expressions as follows:
For instance, if study S01 shows that both usability evaluation methods “UEM1” and “UEM2” detected more usability problems (M1) than another usability evaluation method “UEM3”, the expression built was:

\[ [01 | [S01] | M1] (UEM1, UEM2) > UEM3 \] (1)

Expression (1) is worded as “The evidence 01, which is supported by the study S01, shows that UEM1 and UEM2 are more effective than UEM3 according to the number of usability problems detected”. It should be noted that “≈” could be used instead of “>” if the effect to be expressed is equally effective (no significant differences). In addition, the expressions obtained can be aggregated to summarize the results. The merging process can only take place if the effectiveness measures are the same. For instance, expression (1) can be merged with expressions (2) and (3):

\[ [02 | [S02] | M1] UEM3 > UEM4 \] (2)
\[ [03 | [S03] | M1] UEM3 \approx UEM5 \] (3)

Finally, the result of the merging process is expression (4). These expressions are useful in order to rank UEMs in different levels based on their effectiveness (e.g., UEM1 and UEM2 are the most effective methods at the first level). Note that in (4), the evidence IDs involved are provided instead of ID studies in order to maintain the traceability among previous evidences:

\[ [04 | [01, 02, 03] | M1] (UEM1, UEM2) > (UEM3 \approx UEM5) > UEM4 \] (4)

III. RESULTS

The analysis of the extracted data provided us with the following results for each criterion listed in Section II (D).

With regard to the aim and type of empirical studies (criterion (a)), the results show that 50% of the selected studies were intended to empirically validate a usability evaluation method which had been specifically proposed for the Web domain. On the other hand, the other 50% were intended to perform comparative studies among well-known UEMs in order to provide guidance to researchers and practitioners. In addition, experiments were the most common type of empirical study found (around 83%). This is owing to the fact that experiments provides a high level of control and are useful for comparing usability evaluation methods in a more rigorous manner. Case studies and surveys account for 12% and 6% of the selected studies, respectively.

With regard to the usability evaluation methods that were evaluated (criterion (b)), the UEMs most frequently used in the comparisons were Heuristic Evaluation (HE) [13], Think-Aloud Protocol (TAP), Cognitive Walkthrough (CW), and the Metaphor of Human-Thinking (MOT). Table I shows the complete list of the UEMs that were found in the systematic review by including also their type of method and their attached empirical studies. Any UEM defined as a new modified version of other one has been considered as a separated UEM when these modifications pursued the improvement of the UEM (e.g., Heuristic Evaluation vs. Heuristic Evaluation Plus).

<table>
<thead>
<tr>
<th>Acro.</th>
<th>Usability Evaluation Method</th>
<th>Type</th>
<th>Empirical Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>Automated Summative Evaluation</td>
<td>Testing</td>
<td>[S18]</td>
</tr>
<tr>
<td>CDL</td>
<td>Co-discovery Learning</td>
<td>Testing</td>
<td>[S11]</td>
</tr>
<tr>
<td>CTP</td>
<td>Conceptual Tool for Predicting</td>
<td>Inspec.</td>
<td>[S13]</td>
</tr>
<tr>
<td>CW</td>
<td>Cognitive Walkthrough</td>
<td>Inspec.</td>
<td>[S01][S07][S11]</td>
</tr>
<tr>
<td>CWW</td>
<td>Cognitive Walkthrough for the Web</td>
<td>Inspec.</td>
<td>[S01]</td>
</tr>
<tr>
<td>EE</td>
<td>Expert Evaluation</td>
<td>Inspec.</td>
<td>[S13]</td>
</tr>
<tr>
<td>ESE</td>
<td>End-Survey Evaluation</td>
<td>Inquiry</td>
<td>[S14]</td>
</tr>
<tr>
<td>EYE</td>
<td>Eye-tracking</td>
<td>Testing</td>
<td>[S06]</td>
</tr>
<tr>
<td>GPP</td>
<td>Gerhardt-Powals Principles</td>
<td>Inspec.</td>
<td>[S10]</td>
</tr>
<tr>
<td>HE</td>
<td>Heuristic Evaluation</td>
<td>Inspec.</td>
<td>[S02][S04][S05][S06][S08][S10][S11][S14][S15]</td>
</tr>
<tr>
<td>HEP</td>
<td>Heuristic Evaluation Plus</td>
<td>Inspec.</td>
<td>[S02]</td>
</tr>
<tr>
<td>INT</td>
<td>Interviews</td>
<td>Inquiry</td>
<td>[S17]</td>
</tr>
<tr>
<td>LBT</td>
<td>Lab-Based Testing</td>
<td>Testing</td>
<td>[S18]</td>
</tr>
<tr>
<td>LSP</td>
<td>Logic Scoring Preference</td>
<td>Inquiry</td>
<td>[S03]</td>
</tr>
<tr>
<td>MOT</td>
<td>Metaphor of Human-Thinking</td>
<td>Inspec.</td>
<td>[S07][S08][S09]</td>
</tr>
<tr>
<td>QUE</td>
<td>Questionnaire</td>
<td>Inquiry</td>
<td>[S03][S17]</td>
</tr>
<tr>
<td>RUT</td>
<td>Remote Usability Testing</td>
<td>Testing</td>
<td>[S16]</td>
</tr>
<tr>
<td>SUE</td>
<td>Systematic Usability Evaluation</td>
<td>Inspec.</td>
<td>[S05]</td>
</tr>
<tr>
<td>TAP</td>
<td>Think-Aloud Protocol</td>
<td>Testing</td>
<td>[S09][S11][S12][S17]</td>
</tr>
<tr>
<td>TUT</td>
<td>Traditional Usability Testing</td>
<td>Testing</td>
<td>[S15][S16]</td>
</tr>
<tr>
<td>WDP</td>
<td>Web Design Perspectives</td>
<td>Inspec.</td>
<td>[S04]</td>
</tr>
</tbody>
</table>

With regard to the measures that were employed to assess the effectiveness of UEMs (criterion (c)), the most common measure employed was the ratio of usability problems detected (M1). This measure is also known as thoroughness and is defined as the ratio between the number of problems identified and the total number of existing problems. In some studies, such as [S02][S10][S11], this measure is weighted by the validity measure in order to provide a more rigorous effectiveness measure. Validity is defined as the ratio between the real problems identified (i.e., problems which are not false positives) and the total number of problems identified. Table II shows the complete list of the measures that are involved in the studies. The variety of measures employed to assess the effectiveness of UEM makes it difficult to summarize empirical data from different studies.

<table>
<thead>
<tr>
<th>Code</th>
<th>Measure Name</th>
<th>Empirical Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Ratio of usability problems detected</td>
<td>[S01][S02][S04][S05][S06][S07][S08][S09][S10][S11][S13][S14][S15][S16]</td>
</tr>
<tr>
<td>M2</td>
<td>Severity and quality of problems</td>
<td>[S07][S08][S09][S12][S17]</td>
</tr>
<tr>
<td>M3</td>
<td>Ratio of task success</td>
<td>[S12][S18]</td>
</tr>
<tr>
<td>M4</td>
<td>Usability scores</td>
<td>[S03]</td>
</tr>
<tr>
<td>M5</td>
<td>Number evaluators</td>
<td>[S05]</td>
</tr>
<tr>
<td>M6</td>
<td>Number of evaluator utterances</td>
<td>[S12]</td>
</tr>
<tr>
<td>M7</td>
<td>Number of comments elicited</td>
<td>[S17]</td>
</tr>
</tbody>
</table>
With regard to the Web artifacts that were evaluated (criterion (d)), all the selected studies used a final Web application as the evaluation object. However, a few studies also used other Web artifacts to support the usability evaluations. For instance, in [S05] Hypermedia Design Models are also used, and [S09] also uses prototypes for evaluating and redesigning user interfaces.

With regard to the context of the empirical studies (criterion (e)), we observed that the majority of studies used graduate students as both evaluators to perform usability inspections (e.g., heuristic evaluations, cognitive walkthroughs) and participants in experimental sessions (e.g., think-aloud protocol, remote user testing). However, replications of experiments, which are needed to strengthen the empirical results obtained and to generalize them under certain conditions, are less common than expected.

Finally, as a result of the data synthesis, Table III presents the empirical evidences extracted from the selected studies that were coded according to the representation proposed in Section II (D).

<table>
<thead>
<tr>
<th>TABLE III. EVIDENCES EXTRACTED AND AGGREGATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Evidences from Empirical Studies</td>
</tr>
<tr>
<td>[01] [S01] M.1 CW &gt; CW [10] [S10] M.1 HE = GPP</td>
</tr>
<tr>
<td>[03] [S03] M.1 LSP &gt; QUE [12] [S12] M.1 TAP(ES) = TAP(B&amp;R)*</td>
</tr>
<tr>
<td>[04] [S04] M.1 WDP &gt; HE [13] [S13] M.1 CTP = EE</td>
</tr>
<tr>
<td>[05] [S05] M.1 SUE &gt; HE [14] [S14] M.1 HE &gt; ESE</td>
</tr>
<tr>
<td>[06] [S06] M.1 HE &gt; EYE [15] [S15] M.1 HE = TUT</td>
</tr>
<tr>
<td>[07] [S07] M.1 MOT &gt; CW [16] [S16] M.1 TUT = RUT</td>
</tr>
<tr>
<td>[08] [S08] M.1 MOT &gt; HE [17] [S17] M.1 TAP &gt; (INT, QUE)</td>
</tr>
<tr>
<td>[09] [S09] M.1 MOT &gt; TAP [18] [S18] M.1 ASE = LBT</td>
</tr>
<tr>
<td><em>(2 variants of TAP)</em></td>
</tr>
</tbody>
</table>

Aggregated Evidences

[19] [E02, 04, 05, 06, 08, 10, 11, 14, 15, 16] M.1] (CDL, HEP, MOT, SUE, WDP) (GPP=HE=TUT=RUT) > (EYE, ESE)

Our results suggest that the following UEMs can be considered as the most effective methods with which to perform Web usability evaluations: CW, HEP, MOT, SUE and WDP as inspection methods; and TAP and CDL as testing methods. However, it is important to note that more empirical evidences are needed to strengthen these results.

IV. LIMITATIONS OF THE SYSTEMATIC REVIEW

The main limitations of this systematic review are related to publication bias, selection bias, and inaccuracy in data extraction and synthesis. Since our initial set of candidate papers was provided by our previous systematic mapping, we had already assured to scan relevant special issues of journals and conference proceedings in order to alleviate the publication bias. However, our systematic mapping study neither considered grey literature (i.e., industrial reports or PhD theses), unpublished results, nor papers published after 2009. This may have affected the validity of our results to some extent.

We attempted to alleviate the selection bias (at least to some extent) by defining our inclusion criteria in order to gather the largest possible amount of studies presenting empirical evidences about UEM effectiveness; and by validating the inclusion strategy through assessing the agreement level among three independent raters. Although the quality assessment is intended to select high-quality empirical studies, this may have also affected the validity of our results regarding the final number of selected papers.

In order to alleviate the inaccuracy in data extraction and synthesis (at least to some extent), these stages were conducted by all three authors. In addition, all the discrepancies that appeared were solved by consensus.

We have also detected other limitation related to the systematic review procedure itself which is intended to be addressed in further work. Since the goal of this study is only based on the effectiveness of UEMs, we have not considered other attributes other performance characteristics which may be interesting for both researchers and practitioners.

V. CONCLUSIONS AND FURTHER WORK

We have presented a systematic review to analyze which Web usability evaluation methods have proven to be the most effective. A total of 18 out of 206 empirical studies regarding UEM comparisons were selected. Empirical evidences from these studies were extracted, coded and aggregated in order to discover which UEMs have been proven to be more effective than others.

This systematic review has some implications for research and practice. For researchers, the review identifies two issues: 1) low number of empirical studies; and 2) different measures to quantify the effectiveness of a UEM.

The first issue shows that there is a clear need for more empirical studies of comparing Web usability evaluation methods, not only in number but also in quality. This limitation is in line with the systematic review performed in Web Engineering field by Mendes [12], in which it is claimed that the majority of empirical studies cannot be considered to be methodologically rigorous.

The second issue shows that there is a need of a standard effectiveness measure for the comparison of Web usability evaluation methods. This is in line with studies performed in the Software Engineering field such as Gray and Salzman [8] and Hartson et al. [9] 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.

For practitioners, this review shows empirical evidences of UEMs which can be proven to be effective for evaluating the usability of Web applications. However, an important task for practitioners is not only to compare results from different UEMs, but also to collect data concerning the employment of the UEMs, that can be used to assess the usability of the UEM itself. This data can be very useful in detecting deficiencies and in re-designing evaluation methods in order for them to be more effective.

Although our results suggest that several UEMs are effective methods with which to perform Web usability evaluations, these results need to be interpreted with caution.
since they aim to guide researchers and practitioners, and are not intended to show which method is better than another since other factors such as the context of the empirical studies may affect these results.

Further work is intended to extend and update this systematic review in order to consider other performance characteristics that were discovered apart from effectiveness (e.g., efficiency, learnability, and cost-effectiveness).

ACKNOWLEDGMENT
This research work is funded by the MULTIPLE project (TIN2009-13838) and the FPU program (AP2007-03731) from the Spanish Ministry of Science and Education.

REFERENCES


ANNEX A: LIST OF SELECTED STUDIES


