Univ Access Inf Soc (2011) 10:33–49
DOI 10.1007/s10209-010-0186-8

LONG PAPER

Quantitative assessment of mobile web guidelines conformance

Markel Vigo · Amaia Aizpurua · Myriam Arrue · Julio Abascal

Published online: 19 May 2010

Abstract Conformance metrics for the mobile web can play a crucial role as far as engineering mobile websites are concerned, especially if they are automatically obtained. In this way, developers can have an idea in numeric terms of how suitable their developments are for mobile devices. However, there are a plethora of devices with their own particular features (screen size, formats support, etc.) that restrict a unified automatic assessment process. This paper proposes a tool-supported method for device-tailored assessment in terms of conformance with Mobile Web Best Practices 1.0, including the definition of five quantitative metrics for automatically measuring mobile web conformance: Navigability, Page layout, Page definition, User input and Overall score. The behaviour of these metrics was analysed for different devices and different web paradigms, both mobile web pages and their equivalent desktop pages. As expected, the results show that mobile web pages on more capable devices score higher. In addition, 20 users took part in an experiment aimed at discovering how conformance-based scores relate to usability dimensions. The results demonstrate that automatic scoring approaches strongly correlate with usability scores obtained by direct observation, such as task completion time and user satisfaction. This correlation is even stronger for the device-tailored assessment than the one that assumes a general profile for all devices. For instance, results show a strong negative correlation between Overall score and task completion time: $\rho\ (9) = -0.81$, $(p < 0.05)$ for the generalist approach and $\rho\ (9) = -0.88$ for the device-tailored one, entailing that mobile web guidelines and the metrics based on their conformance capture usability aspects. This result challenges the widely accepted belief that conformance to guidelines does not imply more usable web pages, at least for web accessibility conformance.

Keywords Mobile web · Usability · Metrics · Device-tailored evaluations

M. Vigo (✉) · A. Aizpurua · M. Arrue · J. Abascal
Computer Science School, Department of Computer Architecture and Technology, University of the Basque Country, Manuel Lardizabal 1, 20018 Donostia, Spain
e-mail: markel@si.ehu.es

A. Aizpurua
e-mail: scpaiaga@ehu.es

M. Arrue
e-mail: myriam.arrue@ehu.es

J. Abascal
e-mail: julio.abascal@ehu.es

1 Introduction

Mobile phones, PDAs and video game consoles have become much more widespread over the last few years. They are more affordable than they used to be, even though they are currently more powerful in terms of computing capacity, storage and network connectivity. Although legacy devices will remain, especially in the developing world [18], current devices already support (X)HTML flavoured browsers. This is one of the factors, amongst others, which is enabling the advent of the ubiquitous and mobile World Wide Web. A huge part of the world's population owns a mobile device. According to the International Telecommunication Union, at the beginning of 2009, there were around 4,000 million mobile line subscribers [17]. This rapid increase in the number of mobile owners (from 12% of the world's population in 2000 to 61% in 2009) has contributed towards making the Web

more accessible, although not all users have access to data networks, mainly due to non-affordability or lack of broadband access. Although the mobile web experience is a relatively new field, the rapid growth of potential customers has led to ergonomic aspects and best practices for the mobile web being considered.

Small displays, low text input rate, lack of a pointing device or low bandwidth, amongst others, constrain the browsing experience in the mobile web [19]. Buchanan et al. [4] proposed several guidelines aimed at developing simple websites in order to avoid scrolling and minimize navigation and keystrokes. In order to automatically overcome such problems, some proposed browser-based solutions [5], while Cui and Roto [9] propose a widget-based solution. Church et al. [7] conducted an extensive study into user behaviour on the mobile web in order to ascertain browsing and searching trends, concluding that screen limitations and low text input performance incline users to browsing rather than to searching. Shrestha [34] conducted a user study to explore the differences between desktop and mobile web usability. As expected, desktop users performed much better, mainly due to the above-mentioned device limitations. As a result, several design best practices are proposed. Therefore, for an optimal user experience of the mobile web, not only are browser-dependent solutions required but also the content should satisfy design guidelines.

In this context, the Mobile Web Initiative (MWI) from the W3C released the Mobile Web Best Practices, MWBP 1.0 [30]. These best practices or design principles aim at providing guidance for the development of mobile websites that enhance the ease of use. Similarly, the ISO 9241-11 [16] standard defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Consequently, it can be hypothesized that there is a strong relationship between conformance to the MWBP and usability.

Regarding automatic evaluation of the MWBP, two tests sets have been released so that they can be deployed in semi-automatic guideline review tools: mobileOK Basic [29] and mobileOK Pro [33]. While the former is a mature comprehensive set of evaluation tests, the latter, which supposedly is more demanding, is still a Working Draft. In this context, some review tools that implement mobileOK Basic tests have been developed. Among other features, several evaluation tools, such as TAW mobileOK Basic checker[1] by Fundación CTIC and ready.moby[2] by mTLD, check the mobileOK Basic test. The W3C Mobile Web Initiative released an open-source downloadable checker as well as its web interface.[3] Taking advantage of this library,[4] Garofalakis and Stefanis [13] developed MokE, an online checker. Although tools are not going to replace human judgement, they are of a great help in saving time. Manually checking websites against design guidelines is a time-demanding task that can be cumbersome and frustrating without the help of guideline review tools.

According to Roto [32], when accessing the WWW with mobile devices, four aspects determine the user experience: the accessing device, the browser, the Internet connection and the website itself. A holistic usability evaluation would require considering them as interdependent attributes. As far as device characteristics are concerned, the MWBP and mobileOK tests only take into consideration the so-called Default Delivery Context[5] (DDC), which the MWBP document refers to as "the minimum delivery context specification necessary for a reasonable experience of the Web". Mobile devices are extremely diverse, and most of them deviate from the DDC. None of the above-mentioned tools contemplate device-specific evaluation, even though the MWBP are full of statements in this regard. Therefore, these automatic checkers will yield a great number of false positives for those devices that have more and better features than the DDC, as well as false negatives for those devices with fewer features and less support. While the W3C evaluation library is useful in several scenarios when effectiveness and accuracy are required, tools that rely on it and its underlying philosophy are useless. Vigo et al. [41] addressed this problem by retrieving device features from heterogeneous data repositories and filling in slots in flexible guideline definition languages.

When it comes to measurement, software metrics have traditionally been applied in engineering processes in order to manage critical situations that arise in the development and maintenance stages. Measurable and quantifiable assessment is needed to identify errors and for decision-making support, cost prediction and quality rating. The statement made by software engineering practitioners that "you can't control what you can't measure" [10] captures perfectly the purpose of metrics. With regard to the WWW, web metrics specifically assess the conformance of web applications with the requirements of quality models such as 2QCV3Q [25]. These quality models often consider usability as one of the key attributes for developing high-quality products. In this sense, sound methods and tools to semi-automatically assess web application usability by means of quantitative metrics have been proposed [28].

The objective of this paper is twofold: first, it provides a method for an accurate mobile web conformance

[1] http://validadores.tawdis.net/mobileOK/

[2] http://ready.mobi/

[3] http://validator.w3.org/mobile/

[4] http://dev.w3.org/cvsweb/2007/mobileok-ref/

[5] http://www.w3.org/TR/mobile-bp/#ddc

measurement that considers the specific hardware, software and user-agent features of a particular mobile device. To this end, data from device-tailored evaluation reports are automatically exploited. Secondly, it aims at discovering the relationship between user behaviour and automatically obtained mobile web guidelines conformance scores, thus enabling to ascertain whether the scores can be used as usability predictors. Since the MWBP guidelines can be understood as usability principles, those pages that conform to this set of guidelines are more usable.

The rest of the paper is organized as follows: Sect. 2 describes the application scenarios where the framework for device-tailored assessment can be used. Sections 3 and 4 (partially) are based on prior work [40]. The former discusses the relationship between mobile web usability and accessibility, and the latter describes the architecture of the evaluation framework for specific mobile devices and reporting issues. Section 5 goes deeper into the metrics for mobile web guidelines conformance, and Sect. 6 shows a case study that compares scores obtained when evaluating the desktop and mobile version of 102 web pages with different devices. In addition, Sect. 6 also addresses how the metrics behave when changing the evaluation paradigm, i.e. whether evaluations are device-tailored or DDC-based. User testing and the obtained results are reported in Sect. 7, and finally, conclusions are drawn in Sect. 8.

2 Application scenarios for device-tailored assessment

The assessment framework presented herein can be used in diverse situations that benefit both developers and end-users.

2.1 Engineering mobile web applications

There are four approaches for accessing device-tailored web content [3]. While the first and second approaches focus on the development of content that caters for the specific characteristic of each device, the third and fourth make use of external agents that adapt content automatically:

- Device-specific authoring entails developing web content for a particular device.
- Multiple-device authoring is similar to device-specific authoring, but in this case, a range of devices is identified and an adaptation process is carried out based on automatic transformations upon user request.
- Client-side navigation relies on the capabilities of the user agent to adapt or transform the content.
- Automatic re-authoring. A unique web document is transformed taking into account the characteristics of the client device. These transformations are applied on the client, on the server side or on a proxy.

The application scope of the framework presented herein would fit in the first approach. Web developers can take advantage of the evaluation framework and create content for a set of mobile devices. In addition, it can serve as a sound tool to validate the content created by multi-device authoring environments. The assessment framework produces a numeric value to measure to what extent a web page conforms to mobile web guidelines. Therefore, it also serves to keep track of the conformance or quality assurance of prototypes in the iterative development lifecycle. Since frequent updates characterize the Web and even more so the emerging Web 2.0, where users play an active role as content creators, mobile web metrics can be used for maintenance and monitoring purposes.

Generally, web pages that are specifically designed with mobile devices in mind are close to being MWBP conformant. However, not all websites have their equivalent mobile version and users often find it quite challenging interacting with desktop websites with their mobile devices. Automatic metrics can also be useful to accurately know beforehand the usability level of a website in order to estimate the effort to transform a desktop website into a mobile friendly one.

2.2 Adaptive navigation support

Mobile Web scores are indicators of how suitable a page's content is for a determined mobile device. Scores can be used in the Information Retrieval processes to sort search engine results according not only to the best query match but also to device suitability. In this context, scores can also be deployed in order to annotate links according to the suitability of the content they point to. It can be hypothesized that by annotating links in this way, user orientation and satisfaction will increase when browsing mobile web content. A similar approach was successfully followed for blind users [43]. Users found annotated links useful when links were related to the same topic. However, in the case of the mobile web, it is necessary, first and foremost, to demonstrate that conformance scores adequately capture the suitability of content in terms of usability for a determined mobile device (see Sect. 8).

3 The mobile web and web accessibility

The problems encountered while interacting with the WWW in a mobile context can be referred to as accessibility barriers for the able-bodied [14] and reinforce the statements that accessibility content benefits all users [37].

In this sense, Trewin [38] highlighted the existence of an overlap between mobile web usability recommendations and guidelines for physically impaired users. While this study also emphasizes that these users have a more extreme range of requirements, Mankoff et al. [24] state that by applying accessibility-related good practices, navigation in the Web can be enhanced for a wider audience. Putting together the previously mentioned studies and an initiative to formalize this overlap by Yesilada et al. [44], the problems that the able-bodied encounter while browsing the WWW with mobile devices and the barriers found by users with physical, sensorial and cognitive disabilities while interacting with the WWW on a desktop computer are related to device features as follows:

- The *small display* results in excessive scrolling, which causes disorientation of the user in a similar way to that in which the lack of context disorientates visually impaired users with desktop computers [22].
- *Lack of a pointing device* forces the user to rely on reduced keyboards, which significantly slows down the navigation. Thus, there is an information overload and excessive sequencing when reading the information. Visually impaired users find similar barriers with desktop computers when navigation bars and menus are read time and again until they get to the piece of information they are looking for [22].
- *Low text input rate* filling out a form or typing a URL with a reduced keyboard can be a tedious task and the error rate increases considerably, resulting in frustrating browsing sessions. Users with motor disabilities face an analogous problem when accessing the Web with an alternative input device.
- *Low bandwidth* an unaffordable high-speed connection or low-speed connection leads users not to load pictures in their browsers. If there is no alternative description for visual content, there is an information loss that affects mobile device users in the same way as blind users when accessing a web resource in the WWW.
- *Colour* if the device has a monochrome display (which is not very frequent) or limited colour support, information conveyed by images can be lost. Colour-blind users face similar problems.
- *Lack of support* for mark-up, scripting or data formats. While the previous items in this list are related to the accessibility of the access device, this last item refers to the accessibility of the content itself. In desktop environments, assistive technologies such as screen readers cannot handle certain mark-up or scripting formats. Similarly, user agents on mobile devices have analogous support problems, and usually, the user is not able to access the information.

Mobile devices can help people with disabilities to lead a more autonomous life, and therefore, they also make use of these devices. However, they also have to deal with accessibility barriers [31]: "The main difference I find while accessing the web on a mobile phone rather than a desktop is that the screen reader fails to identify web page elements such as alternatives for images, headings, lists." These sorts of problems produced by user agents and more specifically by assistive technologies were identified as negative dependencies [40]. No matter if a web page conforms to accessibility guidelines such as providing alternatives to visual content or labelling headings adequately, accessibility barriers will still remain due to the lack of support by assistive technologies.

4 Device-tailored web guidelines evaluation

Taking into consideration all the specific features of all mobile devices in the Mobile Web Best Practices and its derived mobileOK documents is not a very effective practice, as the existing large amount of devices would lead to unmanageable documents. Consequently, all assumptions, guidelines, best practices and evaluation tests depend on the Default Delivery Context, which is a useful mechanism, for the reader, to understand how device features impact on the suitability of content on mobile devices. Yet it has been shown in Sect. 1 how inefficient it is to consider the DDC in some scenarios, particularly when guideline review tools are used. Therefore, those best practices that have a dependency on device features such as hardware, software, user agents or even HTTP headers have also been identified. An RDF-based vocabulary has been created in order to univocally refer to these concepts.

4.1 Evaluation framework

Figure 1 depicts the architecture of the evaluation framework for specific mobile devices, which is described in more detail in previous work [41]. The components and processes work as follows:

1. HTTP headers are manipulated in order to simulate a particular device requesting a web resource. As a result, if there is a mobile web version of the requested web page and the server is configured to redirect to the mobile URL, the source code for this particular mobile device is retrieved. In addition, some HTTP headers are gathered as they contain relevant data to complete evaluation tests.

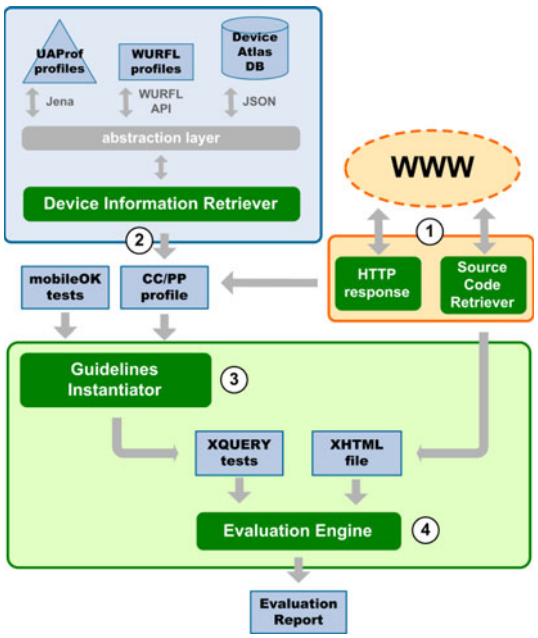

Fig. 1 Architecture of the evaluation framework for specific mobile devices

2. Data about the particular device's features are retrieved from heterogeneous repositories (UAProf,[6] WURFL[7] and DeviceAtlas[8]) and are put together in a profile along with data from the previously obtained HTTP response headers. This dynamically created profile extends the Composite Capability/Preference Profiles (CC/PP) vocabulary for profiling [21] in order to gain expressivity and to be able to specify those terms that are not defined by UAProf or CC/PP vocabularies.
3. Best practices such as mobileOK tests are specified in a machine-understandable way that follows an RDF-enriched XML Schema, as proposed by Hunter and Lagoze [15]. In this language, RDF statements are used as values of XML attributes and refer to the device-dependent issues identified in the MWBP and mobileOK documents, i.e. those device features, such as the specific screen size or scripting support, which in the mobile web guidelines are assumed to correspond to those defined by the DDC. The CC/PP profile that captures device features and HTTP traffic data also makes use of this notation, so that matching between profiles and filling in slots in guidelines becomes straightforward. Figure 2 shows how the evaluation framework behaves at this stage with a test case for the *Scrolling* operation. In this case, the CC/PP contains information regarding the width of the screen, which is expressed with the RDF property `access:screenWidth`, with its value equal to 200 pixels. Although the specification of the test case in the guidelines definition language is fixed, it contains a slot to be filled in with the information from the CC/PP profile. After the matching process, the guideline is completed with data from the profile and conveys the following instruction: "Check whether there is any IMG element that contains a width attribute with a value greater than 200 pixels".
4. Finally, XQuery statements are inferred on-the-fly and automatically created from the guidelines that have previously been completed with device features data and HTTP traffic headers. Evaluation is performed directly by executing XQuery statements against the web resource to be evaluated, which has been transformed into an XML document.

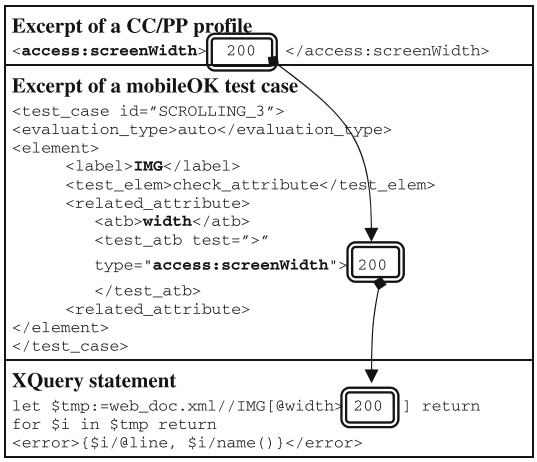

Fig. 2 How slots in the guidelines are filled in from CC/PP profiles and the automatic creation of XQuery statements

4.2 Reporting

XQuery statements deal with reporting issues and yield information regarding the specific test cases that have been violated, where in the source code the problem is located, which element, attribute or value has caused the issue, etc. This information is of paramount importance for developers to be able to find, check and repair problems. As it is explained in the next section, for measurement purposes both the number of issues and the number of potential points of failure are necessary. As the main goal is to obtain usability scores using data in reports produced by automatic review tools, retrieving the types of problems is also crucial, as it has underlying implications for weighting issues found. Thus, a fine-grained approach to reporting is proposed by extending the report issues in MWBP and

[6] http://w3development.de/rdf/uaprof_repository/
[7] http://wurfl.sourceforge.net/
[8] http://deviceatlas.com/

mobileOK tests. The evaluation framework identifies 3 different issues and, at most, produces two of them simultaneously:

In contrast to mobileOK tests, a more fine-grained approach is proposed that aims at meeting W3C reporting standards such as EARL [1]. Single best practices are an `earl:TestRequirement` and are implemented by evaluation techniques, which at the same time are decomposed into atomic evaluation test cases (`earl:TestCase`). When applying the herein-defined mobileOK tests, those test cases that can be fully evaluated automatically (`earl:automatic`) produce the following issues:

- *Auto* is a fully automatic test and not satisfying this type of test case violates mobile web principles. It produces a *pass* (`earl:passed`) if it is met and a *fail* (`earl:fail`) otherwise.
- *Recommendations* are informative warnings that are raised by fully automatable tests, but they are not reliable evidence for indicating whether the violation of the test case entails a major issue in the interaction quality. For instance, it is not clear if the use of `accesskeys` benefits the user, as there might be an overlap with other shortcuts. Moreover, since the nomenclature for `accesskeys` is not standardized, if they were implemented the user is seldom aware of it.

Semi-automatic issues or warnings (`earl:semi-automatic`) are also identified:

- *Warnings* are raised by semi-automatic tests that partially check one test case. The remainder of the best practice, the statement that cannot be automatically checked, should be evaluated by an expert. It is equivalent to `earl:cantTell`.

In order to automatically measure the usability level of a web resource for a specific mobile device, a component that obtains the above-mentioned data (issue type, number of errors, etc.) from evaluation reports has been integrated into the architecture of Fig. 1.

5 Mobile web quantitative metric (MWQM)

The Web Accessibility Quantitative Metric (WAQM) was defined in [39]. This metric catered for the specific reporting characteristics of WCAG 1.0 guidelines [6] and was adjusted to the EvalAccess evaluation tool's peculiarities. The Mobile Web Quantitative Metric, which is tailored to the evaluation framework and guidelines presented herein, has two main characteristics: (1) it yields quantitative scores and (2) they are automatically obtained. Thus, it fits into the application scenarios devised in Sect. 2.

However, the above-mentioned characteristics lead to make several assumptions:

- **Assumption 1**: it is assumed that test cases and the best practices on which they are based are valid. Therefore, it is accepted that the non-fulfilment of any of them will cause a loss in the quality of interaction and usability.
- **Assumption 2**: it is assumed that all test cases impact on the usability in a similar way. Thus, all best practices have the same severity.
- **Assumption 3**: usability scores will produce a normalized value. As discussed later, this is a precondition for applying the Logic Scoring Preferences method. Thus, usability can be described in percentage terms.
- **Assumption 4**: test cases that require human judgement are considered actual problems since the objective is to obtain mobile web conformance automatically. Therefore, results will produce a lower-bound conformance score.

Three metrics are identified for different test cases:

- **Failure rate (fr)**: this measures the ratio between actual errors and potential errors. For example, the *Non_text_alternatives* test case checks whether each picture has an alternative description. Thus, 10 pictures out of 100 would obtain $fr = 0.1$, while 5 images out of 25 would obtain $fr = 0.2$. Therefore, the normalized conformance score is 1-*fr*. Sullivan and Matson [36] followed a similar approach for measuring web accessibility.
- **Accept/reject**: while test cases to be measured by the failure rate are checked every time, a determined mark-up label or attribute appears, and some test cases are evaluated just once. For instance, the *Balance* test case states that links should be used conservatively. Therefore, following the mobileOK tests, this test produces one error if more than 30 links are used. This metric can be understood as a particular case of failure rates when the range from 0 to 1 is covered just by integer values, and thus, only two values are possible. If one test case of this type fails, the conformance will decrease proportionately to the number of test cases in a guideline. For instance, if *Background_Image_Readability* fails, the overall score for *Page Layout and Content* guideline will decrease by 25%, as there are just four test cases.
- **Absolute number of errors**: test cases in this category are those related to technology support (be it either software or hardware) by a specific device, such as table rendering or *querty* keyboard support. If a determined device does not support a given technology, all related mark-up elements will produce an error. As an example, if there is no support for tables (as described by the *Tables_suppor*t best practice), since

tables may not degrade gracefully all tables fail to meet this test. In this case, the total number of (X)HTML labels will divide the total number of tables in order to obtain a percentage value that shows the impact of this test case. As structural elements are not related to content, they are left out of the total number of tags. A similar approach was followed by Bayley and Burd [2] in order to measure web accessibility.

As can be observed in Table 1, some test cases produce both errors and warnings, while others contemplate more than one metric. This is due to the fact that numerous techniques are divided into test cases. The last two columns on the right refer to which of the two sets of mobileOK tests corresponds to the test case. As there is an overlap, sometimes both sets can specify the same test case, although mobileOK Pro tends to be more demanding and stricter. In such a case, mobileOK Pro will be preferred.

5.1 Adapting logic scoring preferences for mobile web guidelines

Traditional scoring techniques work as follows: a number of components (n) are independently evaluated; in the particular case of the mobile web evaluation framework $n = 34$, which is the number of best practices. Evaluation results are a set of normalized scores $E_1,\ldots,E_n$, where $0 \leq E_i \leq 1$. When evaluated components have a different impact on the measurement, positive normalized weights are associated with each evaluation result $W_1,\ldots,W_n$ where $0 < W_i < 1$ and $\sum_i W_i = 1$. As a result, the global score would be $E = W_1E_1 + \cdots + W_iE_\mathrm{i} + \cdots + W_nE_n$, $0 \leq E \leq 1$. This approach does not work when a number of elements are used, because the distribution of weights is not effective and results in the following limitations:

- Mandatory requirements cannot be modelled. If $E_i = 0$, E will never be equal to zero.
- If the number of components is very high, the consequence of a low score for a given component is not very significant.
- If components are significant and thus have a high weight, the impact of low-weighted components is irrelevant.

Logic Scoring Preferences, LSP by Dujmovic [12], is an aggregation model that overcomes the above-mentioned limitations. LSP can also be understood as a preferential neural networks model. Its strength relies in its capacity for evaluating complex systems including numerous subsystems that can be composed of further subsystems and elements. Similarly, the MWBP are composed of four general guidelines containing a number of best practices that at the same time are decomposed into several test cases for evaluation purposes. The high number of subcomponents and the fact that they can be grouped according to best practice and guideline membership lead to believe that LSP is appropriate for mobile web usability measurement. Moreover, LSP was successfully applied in the context of the measurement of web applications usability [28]. Using the *weighted power mean*, the drawbacks of traditional aggregation systems are overcome.

$$E = \left(W_1E_1^{p(d)} + \cdots + W_iE_i^{p(d)} + \cdots + W_nE_n^{p(d)}\right)^{1/p(d)}$$

The values of $\rho(d)$ are predefined elsewhere [11], and they are selected based on the required logical relationship between elements of the system, being different levels of conjunction and disjunction. The output of the $\rho(d)$ function changes depending on the number of elements to measure and d, which is the degree of disjunction. The value of d ranges from total disjunction ($d = 1$), arithmetic mean ($d = 0.5$), to conjunction ($d = 0$) in steps of 1/16. When simultaneity in satisfying the requirements (mobile web best practices in this case) is necessary, conjunction or similar is applied. In contrast, if the objective is to penalize the main component only if all subcomponents fail, disjunction is applied. Normally intermediate values are preferred, as extreme cases do not apply. These intermediate ranges of values are ($0 < d < 0.5$) for *quasi-conjunctions* and ($0.5 < d < 1$) for *quasi-disjunctions*. Depending on the value of d, relationships between elements can be weak, medium or strong. More details on the mathematical background can be found in [12].

LSP is useful when components in a system are hierarchically arranged and there are numerous items. As can be observed in Table 1, MWBP 1.0 can be decomposed into guidelines and best practices. In addition, these best practices can also be decomposed into test cases, resulting in a number of requirements to be met in order to fully conform to MWBP 1.0. The relationships between the components in the system are determined by their reporting type and the location within the hierarchy (be it guideline, best practice or test case). The following paragraphs describe these relationships in more detail:

- Single techniques are understood as the basic requirements that describe a particular mobile web conformance issue. As for automatic evaluation purposes, techniques are often decomposed into test cases, which are the atomic pieces of evaluation. It is thus required that all its test cases are met in order to satisfy a technique. In other words, it is mandatory to satisfy all test cases simultaneously. Thus, low input values will strongly determine the final result. This idea of simultaneity fits perfectly with the conjunction logical relationship and can be clearly explained by the "a

Table 1 Mobile web best practices that can be (semi-)automatically evaluated and the type of results each best practice yields, as well as the computed metric

Mobile web best practice	Type	Metric	Basic	Pro
Navigation and links				
Balance	Auto	Accept/reject	x	√
Access_keys	Recommendation	fr	x	√
Link_target_format	Auto	fr	√	x
Link_target_id	Auto	fr	x	√
Auto_refresh	Recommendation	Accept/reject	√	√
Redirection	Auto	Accept/reject	√	√
Image_maps	Auto	Count | fr	√	x
Pop-ups	Auto	fr	√	x
Page layout and content				
Scrolling	Auto	fr	x	√
Graphics_for_spacing		fr	√	√
Use_of_color	Warning	fr	x	√
Background_Image_Readability	Warning	Accept/reject	x	√
Page Definition				
Page_title	Auto | warning	Accept/reject	√	√
No_frames	Auto	Count	√	x
Structure	Recommendation	fr	x	√
Character_encoding_support	Auto	Accept/reject	√	x
Tables_support or tables_alternatives	Auto	Count	√	√
Tables_layout	Recommendation | warning	fr	√	√
Tables_nested	Auto	fr	√	x
Non_text_alternatives	Auto | recommendation	fr	√	√
Objects_or_scripts	Auto | recommendation	fr | count	√	√
Style_sheets_use	Auto | recommendation	Count	√	x
Style_sheets_support	Recommendation	Count	√	√
Content_format_support or Content_format_preferred	Auto	fr	√	√
Cookies	Auto | recommendation	Accept/reject	x	√
Caching	Auto | recommendation	Accept/reject	√	x
Fonts	Recommendation	Count	x	√
Images_specify_size	Auto	fr	√	x
Images_resizing	Auto	fr	√	x
User input				
Default_input_mode	Auto | recommendation	fr	√	x
Avoid_free_text	Warning	Count	x	√
Provide_defaults	Recommendation | warning	fr	√	√
Tab_order	Warning	fr	x	√
Control_labelling	Auto	fr	x	√

The last two columns refer to whether each evaluation set (basic for mobileOK Basic; pro for mobileOK Pro) implements a best practice (√: it is implemented; x: it is not implemented)

chain is only as strong as its weakest link" statement. However, as the typology of test cases may vary regarding their fulfilment certainty, it is crucial to define their relationship and the degree of conjunction or *quasi-conjunction* applied:

- **Case 1:** *auto* vs. *recommendation*. Both issues are fully automatable tests, but while those issues yielded by automatic test cases will certainly cause a decrease in the conformance level, those test cases that produce *recommendation* issues just have a strong likelihood of causing a decrease in the usability if a web document fails to meet them. Therefore, as some uncertainty is introduced, the strong *quasi-conjunction* (C+) is applied, where $d = 0.125$.

- **Case 2:** *auto* vs. *warning*. Generally, the evaluation framework test cases that yield *warning* issues are 50% automatable. An expert should check the remaining 50% of the statement. Following the reasoning of the previous item and due to the incompleteness of the latter component, the medium conjunction (CA) is applied, where $d = 0.25$.
- **Case 3:** *recommendation* vs. *warning*. This is the relationship that entails more uncertainty. Therefore, the weakest conjunction (C−) is applied, where $d = 0.375$.

Figure 3 shows where all the aforementioned cases are located in the range of logic relationships that LSP provides:

The best practice *Auto_refesh* can be considered to illustrate the method. Two test cases implement the best practice. Let us assume that an evaluation of a given web page reports the following:

– Test case 1 checks the `meta` element, the `http-equiv` attribute and the URL it points to. In this case, there is no such label and thus $\text{score}_1 = 1$
– Test case 2 checks whether the HTTP response header contains a command to do refresh and if so, it checks whether it is redirected to the current URL. In this case, a problem is found and $\text{score}_2 = 0$.

Since there is a need for simultaneity, the C+ logical function is applied obtaining score_1 C+ $\text{score}_2 = 0$ for *auto_refresh* best practice. This process applies for all the test cases in each best practice.

- Among those best practices that are members of a guideline: single techniques can be considered as elemental usability indexes. In addition, sets of these singles are grouped together in order to satisfy higher-level usability principles, in the particular case that concerns this paper: *Navigation and Links*, *Page Layout and Content*, *Page Definition* and *User Input*. Since simultaneity is also a requirement among best practices within a guideline, the C− logical function is chosen, where d is equal to 0.4375. This function is located between C- and the arithmetic mean (A).

Now let us consider the *Navigation and Links* guideline or principle to illustrate the second step. The following two test cases implement the best practice. Let us consider the following scenario:

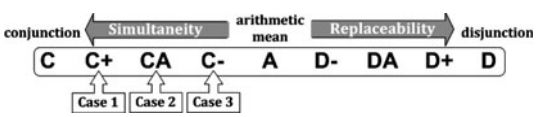

Fig. 3 Subtest case location in the LSP value range

– There are less than 30 links, so $\text{score}_1 = 1$ for *Balance*
– `accesskeys` are not implemented, so $\text{score}_2 = 0$.
– There are two linked PDF documents, and the device has no support. As there is a total of 10 linked elements, the score based on the failure rate for *Link_target_format* is $\text{score}_3 = 0.80$.
– Links are adequately labelled and therefore *Link_target_id* gets $\text{score}_4 = 1$.
– As previously explained, *Auto_refresh* gets $\text{score}_5 = 0$.
– There is no *Redirection* to other pages, so $\text{score}_6 = 1$.
– There are no *Image_maps, so* $\text{score}_7 = 1$.
– There are three *Pop-ups* out of 10 links, and thus, the score will be $\text{score}_8 = 0.7$

Applying the C− function to these best practices, a 0.74 score is obtained for *Navigation and Links.*

- Among the four principles for assessing mobile web usability, each guideline is weighted by the number of test cases it contains divided by the total number of test cases with weight $0 \leq W_i < 1$ and $\sum_i W_i = 1$. In the context of the mobile web evaluation framework, the overall score is $E = W_{Nav}E_{Nav} + W_{Lay}E_{Lay} + W_{Def}E_{Def} + W_{Inp}E_{Inp}$, $0 \leq E \leq 1$.

6 Case study

With the aim of measuring and observing the impact of the assessment method, a case study assessing 102 mobile websites' home pages[9] was conducted. These web pages had an analogous desktop version, which it was assumed would provide similar content and analogous functionalities. In other words, the objective was to observe the behaviour of the automatically obtained metrics for the desktop and mobile versions of a web page.

Two different mobile devices were used in order to show how usability scores vary depending on the device's features. One of them was a legacy device, while the other had more features and more software support than the Default Delivery Context. The former (D_1), a Nokia 3590, has fewer functionalities than the DDC, and the latter (D_2), a Sony Ericsson P990, has more capabilities than those specified in the DDC. Table 2 shows these devices' characteristics in terms of the device-dependent best practices and compared to the Default Delivery Context. The notation for specifying device features is the same as is declared in the device profile: self-describing RDF vocabularies.

One hundred and two desktop web pages and their equivalent mobile web pages were evaluated against the MWBP 1.0 and the aforementioned mobile devices. HTTP

[9] These websites can be found at http://cantoni.mobi

Table 2 Features of mobile devices with respect to their dependencies in the MWBP are boolean, e.g. `cssSupport`, using √ (supported), x (not supported), N/A notation while others refer to numeric characteristics, e.g. `screenWidth`, or supported formats enumeration, such as `picFormatSupport`

Feature	D_1	DDC	D_2
`access:xhtmlSupport`	√	√	√
`access:cssSupport`	√	√	√
`access:picFormatSupport`	GIF, WBMP	GIF, JPEG	BMP, ICO, GIF, JPEG, PNG, SVG + XML, TIFF, WBMP, X-BMP, X-EPOC-MGM, X-WMF
`prf:CcppAccept-Charset`	US-ASCII, ISO-8859, UTF-8, ISO-10646-UCS-2	UTF-8	US-ASCII, ISO-8859-1, ISO-8859-2, ISO-8859-4, ISO-8859-5, ISO-8859-7, ISO-8859-9, ISO-10646-UCS-2, KOI8-R, csKOI8R, UTF-8, UTF-16
`prf:FramesCapable`	x	N/A	x
`prf:JavaScriptEnabled`	x	x	x
`prf:JavaAppletEnabled`	x	N/A	x
`prf:TablesCapable`	√	N/A	√
`access:cookiesSupport`	√	N/A	√
`access:pntSupport`	x	x	x
`access:quertyKeyboard`	x	x	√
`access:screenWidth`	96	120	240

headers were manipulated so that the system retrieved the appropriate web content for each device. Box plots in Fig. 4 show the descriptive statistics, and the following conclusions can be drawn.

- Between D_1 and D_2 in mobile web pages: mean usability value is 57 (SD = 6.5) for D_1, maximum value is 83, minimum is 44 and median is 58; whereas for D_2 mean is 59 (SD = 6.3), the maximum reaches 83 again, minimum is 47 and median 59. Apparently, the values are very similar; actually, D_2 obtains higher scores due to its better features. For instance, *Page layout* obtains higher values in D_2 (mean = 98, SD = 6.1) compared to D_1 (mean = 84, SD = 25). This means that D_2 renders web content much better, mainly because it has a wider screen size, while D_1 shows a higher variability. Looking carefully at the scores obtained by different guidelines, it can also be concluded that *Navigation* and *Page definition* obtain very poor results in both cases, while *Overall*, *Page layout* and *User input* get acceptable values.

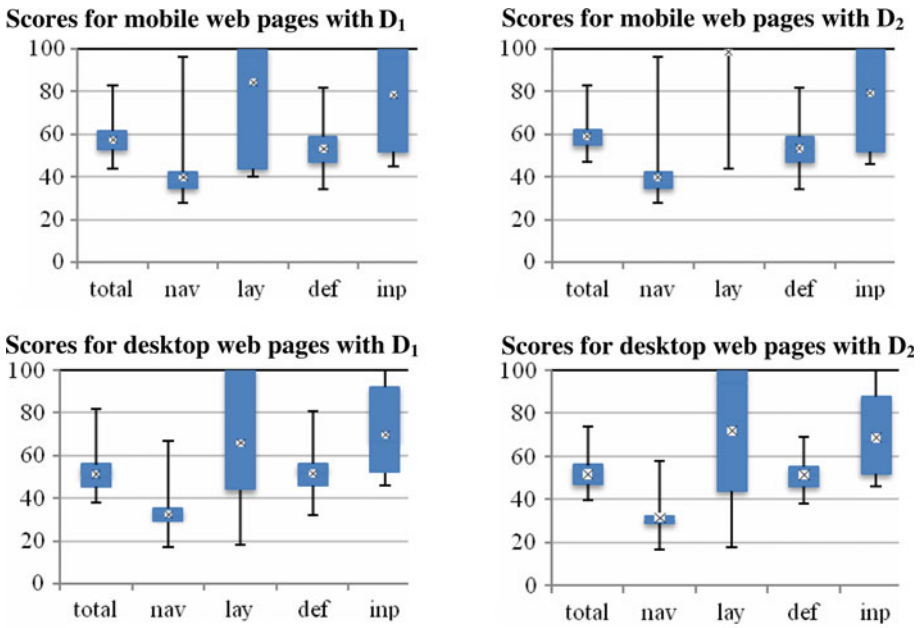

Fig. 4 Scores for D_1 and D_2 when mobile and desktop web pages are measured

- Between D_1 and D_2 in desktop web pages: again, *Navigation* obtains very poor results with a maximum of 67 and a mean of 32 (SD = 7.4) for D_1 and a maximum of 58 and a mean of 32 (SD = 6.3) for D_2. These low values are caused due to the vast amount of information that desktop web pages tend to have and the fact that the lack of a pointing device forces the user to navigate sequentially with reduced keyboards. *Page layout* shows a lot of variability and spans almost the entire range (mean = 66, SD = 28 for D_1; mean = 69, SD = 19 for D_2). Consequently, extreme values are obtained, high values and very low ones (a maximum of 100 and a minimum of 18 in both cases). This means that some pages that were developed keeping in mind that desktop computers adapt quite well to legacy mobile devices while others do not. Something similar happens with *User input* in that it reaches a mean value of 69 (SD = 19) for both cases.
- Between mobile pages and desktop pages: *User input* and *Page layout* get much lower values for desktop pages. For D_1, *User input* is 13 points lower in desktop web pages and it is even worse for *Page layout*, which obtains a mean of 18 points less for D_1 and 26 for D_2.

Histograms of absolute frequency in Fig. 5 show that for those pages developed with mobile devices in mind (H_2 and H_4) scores concentrate around the 50–60 mark, although D_2 obtains substantially more values in the 60–70 range. Conversely, web pages for desktop computers are positively skewed and values concentrate in the 40–50 range, although again D_2 has more values in the 50–60 range than D_1 has. Thus, higher values are obtained by H_4 and lower values are yielded by H_1. The scores empirically confirm that D_2, the device with better features, will provide a better experience in the mobile web, although its behaviour in desktop web pages is also acceptable. This last statement is reinforced by Nielsen's studies [27].

Paired t tests were carried out between the scenarios above in order to discover the significance of the differences between the scores. When comparing desktop and mobile web pages, it was found that for D_1 there was a significant effect for *Overall* $t(101) = 4.09$, $p < 0.000$; *Navigation* $t(101) = 4.99$, $p < 0.000$; *Layout* $t(101) = 2.68$, $p < 0.01$; and *Input* $t(101) = 3.14$, $p < 0.003$. In addition, D_2 also shows a similar behaviour for *Overall* $t(101) = 8.95$, $p < 0.000$; *Navigation* $t(101) = 7.34$, $p < 0.000$; *Layout* $t(101) = 9.17$, $p < 0.000$; and *Input* $t(101) = 3.88$, $p < 0.000$. That is, mobile web pages score higher than desktop web pages, and this difference is greater if the device has more capabilities. These results, although expected, show that the metric behaves adequately. Therefore, it can be concluded that the paradigm (whether web pages are to be deployed on desktop computers or mobile devices) strongly determines the scores as long as *Overall*, *Navigation*, *Layout* and *Input* attributes are measured with the method proposed.

Taking a look at the evaluation details, the mobile web pages contain a mean of 58 HTML elements, while the desktop sites have a mean of 617. The mobile web pages contain almost 10% of the HTML elements of their equivalent desktop pages. The results show that the more elements a web page has, the more likely it is to violate mobile web best practices. However, statistical data show that desktop web pages obtain only 10 points less than mobile web pages. This behaviour can be explained by the fact that failure rates are used in metrics rather than the absolute number of errors. The scores might seem to be low, but as the metrics are automatically calculated those test cases that require human judgement are considered as actual problems, as stated in *Assumption 4*. Consequently, the scores obtained by this evaluation framework give a lower-bound level of conformance. Mobile web pages are simpler than their equivalent desktop pages, but some problems such as providing `accesskeys` and the usage of non-recommended XHTML labels, such as those that can be used for styling purposes (`b`, `font` or `centre`), as well as the problems related to lack of support of content format, produce most of the problems. In addition, those issues that overlap with web accessibility, such as providing alternate descriptions for pictures or form labelling, are often not met.

7 User testing

The conducted user testing was targeted to verify the following hypotheses: (1) The higher a mobile website scores in terms of mobile web best practices, the better users perform, as long as the task involves information search and (2) Scores obtained by device-tailored evaluations are a more accurate approach to users' performance than those relying on the Default Delivery Context.

7.1 Method[10]

7.1.1 Participants

Twenty users (15 men and 5 women) whose average age was 29 years (SD = 2.5) took part in the experiment. All

[10] Further information on the experimental settings can be found at http://sipt07.si.ehu.es/UAIS09/index.html

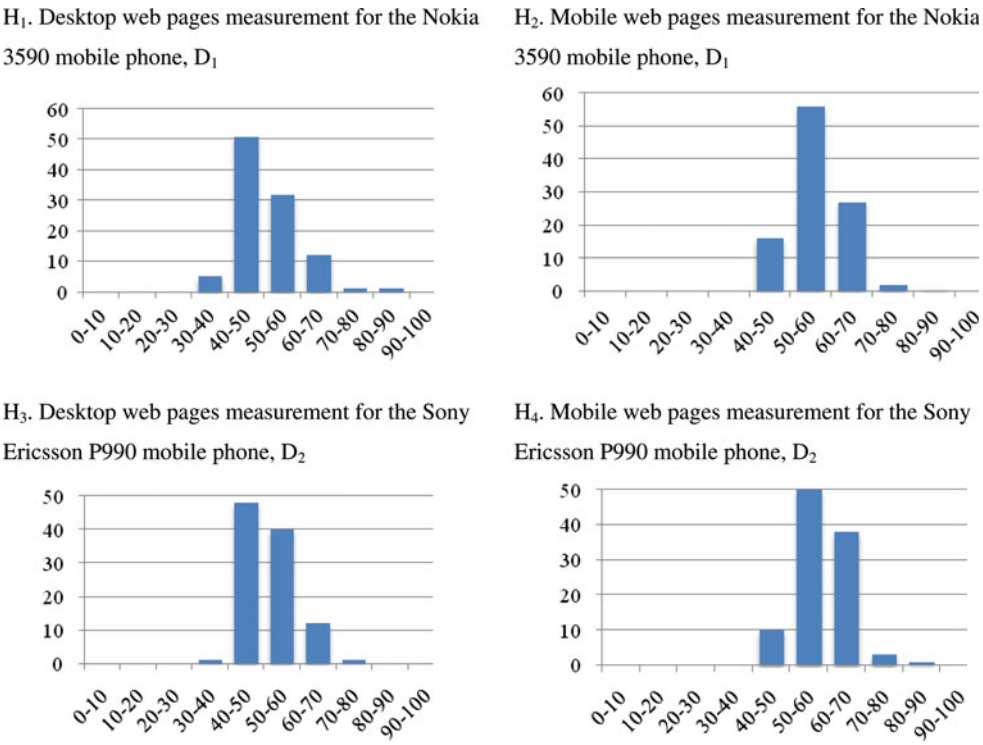

Fig. 5 Scores for D_1 and D_2 when mobile and desktop web pages are measured

of them were Spanish native speakers, although 90% of them had good–excellent English reading skills. The participants were members of staff and students of the Computer Science School at the University of the Basque Country (60% researchers, 25% lecturers and 15% students). Therefore, 75% of the users spend more than 15 h a week browsing the Web. However, they were not familiar with using mobile devices for accessing the Internet as only 15% used them almost every day (mainly for checking e-mail), while 35% seldom used them and 50% had never used mobile devices for browsing the Web.

7.1.2 Materials

Ten mobile websites were selected from the case study in Sect. 6. Due to download inconsistencies, one site was removed from the list and the experiment was thus carried out with 9 websites. The selected sites were information centric rather than interactive or functionality oriented ones and their content was in English (see "Appendix" for a complete list of URLs). Although the search functionality was available on several sites, users were not allowed to use it when searching for a target. WinHTTrack Website Copier was used for retrieving the aforementioned sites on 23rd November, 2008. Since this tool allows manipulation of the HTTP request header, this header was replaced by the corresponding header of the mobile device used in the experiment. The websites were uploaded to a web server in order to keep the data coherent. The websites were accessed using a Dell Axim X30 PDA, running Pocket Internet Explorer on Windows Mobile 2003. Table 3 shows the specific device features that a device-tailored evaluation of MWBP guidelines requires.

7.1.3 Measures

Amongst others, effectiveness, efficiency and satisfaction are considered as usability dimensions by the ISO 9241 usability standard. In addition, an extensive review of the literature on mobile usability [8] concludes that the

Table 3 Features of mobile devices with respect to their dependencies in the MWBP, as explained in Table 2

Feature	
`access:xhtmlSupport`	√
`access:cssSupport`	√
`access:picFormatSupport`	JPG, GIF, BMP, PNG, XBM
`prf:CcppAccept-Charset`	UTF-8
`prf:FramesCapable`	√
`prf:JavaScriptEnabled`	√
`prf:JavaAppletEnabled`	x
`prf:TablesCapable`	√
`access:cookiesSupport`	√
`access:pntSupport`	√
`access:quertyKeyboard`	√
`access:screenWidth`	240

aforementioned are the most frequently measured usability attributes in the mobile environment. In the context of the conducted user test, these attributes are measured as follows:

- Effectiveness is measured in terms of successfully **completed task rate;** 1 if the target is found and 0 otherwise.
- Efficiency is measured in terms of **task completion time**. Even if users had unlimited time to find their goals, those tasks that exceeded 60 seconds were regarded as incomplete or not completed.
- **Satisfaction** is measured using Lewis's after-scenario questionnaire [23]. Specifically, the following 4 questions were selected:

1. The organization of information on the website is clear.
2. The interface of this website is pleasant.
3. It was easy to learn to use this website.
4. Overall, I am satisfied with how easy it is to use the website.

A 10-point Likert scale ranging from 0 (strongly disagree) to 9 (strongly agree) was used to collect users' impressions after each task.

On the other hand, automatic usability measurement was taken using the metrics proposed in this paper. Although users were able to navigate through entire websites, only home pages were measured. Even if it would be desirable to define metrics for websites, a study by Nielsen and Tahir [26] states that the usability of the homepage predicts the usability of the whole site. Vigo et al. [42] provide empirical data supporting such a statement, at least in the context of automatic accessibility measurement.

7.1.4 Procedure

User testing was conducted in a controlled laboratory, and users were videotaped during the experiment in order to collect interaction data afterwards. Users gave their signed consent to being videotaped. First of all, users were introduced to the experiment's objectives and were given instructions. Before starting with the experiment, all of them trained with the PDA and its browser in order to get accustomed to the system, which took them 10 min at most. In this familiarization, users had to find two goals targets in scenarios that were analogous to those appearing in the experiment. After the training stage, users had to complete 9 tasks consisting of finding one target in each of the 9 sites. All the goals were located at depth level 1 in the website, where depth level 0 was the homepage. Therefore, for an optimum execution task, users had to click two links in a site, one at each depth level. Users were able to read the statement corresponding to a task's goal before clicking on the link leading to the task. However, this statement was permanently displayed on the upper part of the screen as the websites were embedded in a two horizontal frame structure. While the upper frame displayed the statement of the task, the main frame below contained the website. This might hinder interaction, but all sites contained this top frame so the potential disturbance was the same for all sites. Regarding assessment, there was no problem with such a frame because it did not have an impact on horizontal scrolling, which is the one addressed by mobile web guidelines. Task execution order was randomized for each user in order to remove the learning effect. Users knew that they had found the objective as goal pages were replaced by a web page stating they could continue with the next task. After completing or giving up each task, users filled out the above-mentioned questionnaire. On average, the whole experiment took half an hour to be completed and users were rewarded with 5€.

7.2 Results

Table 4 shows the automatic scores obtained for each home page with different evaluation approaches, the one based on the Default Delivery Context features and the device-tailored one.

A paired t test for the *overall* attribute shows that the difference between DDC-dependent and device-tailored measurements is definitely significant, $t(8) = 3.11$, $p < 0.05$. In addition, the difference between the *navigation* attribute for both approaches is also significant, $t(8) = 3.91$, $p < 0.00$. This section aims at shedding some light on these different measurements, by investigating the impact of these two measurement approaches on user performance and usability measured in terms of user interaction.

In a post-test videotape analysis, objective usability data, such as success rate and task completion time, were collected. Subjective usability data were obtained from post-test questionnaires. Table 5 shows both objective and subjective usability measures. The first column again contains the website where the users performed their tasks. The following three columns correspond to effectiveness, task completion time and user satisfaction, respectively.

The objective is to determine whether automatically obtained conformance scores correlate with objective usability measured by direct observation and subjective usability. In addition, if the aforementioned premise is fulfilled, it is ascertained whether usability scores obtained from device-tailored evaluations are better predictors than those scores obtained from DDC-based evaluations. Table 6 contains the correlation matrix between automatic

Table 4 Automatically obtained usability scores for each site (column 1) for each MWBP attribute and an overall score for DDC-based (columns 2–6) and device-tailored (columns 7–11) measurements

Site	$Overall_{DDC}$	Nav_{DDC}	Lay_{DDC}	Def_{DDC}	Inp_{DDC}	$Overall_{tailored}$	$Nav_{tailored}$	$Lay_{tailored}$	$Def_{tailored}$	$Inp_{tailored}$
dallasnews	71	73	96	55	100	72	79	96	55	100
edmunds	66	50	94	57	100	73	81	94	57	100
popurls	62	35	44	67	100	75	63	100	67	100
nypost	68	50	100	60	100	75	81	100	60	100
drinkboy	58	48	44	54	100	71	77	100	54	100
wapedia	58	36	100	63	45	65	63	100	63	46
houston	56	32	96	52	74	56	34	96	52	74
elpasotimes	54	33	100	53	52	54	36	100	53	52
bostontimes	53	31	100	53	52	54	34	100	54	52

Table 5 Usability values obtained by direct observation: completed task rate gets 1 if the target is found and 0 otherwise; Task completion time is measured in seconds and is <60; while satisfaction ranges from 0 (strongly dissatisfied) to 9 (very satisfied)

Site	Completed task rate	Task completion time	Satisfaction
dallasnews	0.95	16.11	7.49
edmunds	1	15.35	7.65
popurls	0.8	18.88	5.89
nypost	1	20.95	6.9
drinkboy	1	20.4	6.775
wapedia	0.9	37.5	7.11
houston	0.9	32.78	6.06
elpasotimes	0.45	44.89	5.03
bostontimes	0.95	34.95	5.85

conformance scores and usability in terms of effectiveness, task completion time and satisfaction.

Overall score and task completion time shows a very strong negative correlation, $\rho(9) = -0.81$, $p < 0.00$ for DDC-based and $\rho(9) = -0.88$, $p < 0.00$ for device-tailored measurements, which is 7 percentage points stronger. This means that the higher scores a web page, the less time it takes to find a target. Similarly, there is a strong correlation between *Navigation* and task completion time $\rho(9) = -0.70$, $p < 0.05$ for DDC and $\rho(9) = -0.82$, $p < 0.00$ for device-tailored, 12 percentage points stronger and a higher confidence level. This would imply that providing navigation mechanisms reduces task completion time. There is almost a negative linear correlation between *Input* and task completion time $\rho(9) = -0.94$, $p < 0.00$ for both measurement methods. This might happen due to the fact that controls such as forms, input buttons, text boxes and similar occupy large areas of the often small screens. Taking up large areas of the screen reduces visibility, forcing the users to scroll in order to get the overview of a web page, which is always a time-consuming task. Regarding subjective measurements such as user satisfaction, this also shows a strong correlation with *Overall* score $\rho(9) = 0.74$, $p < 0.03$, while the device-tailored score correlates $\rho(9) = 0.67$, $p < 0.05$. Finally, *Navigation* and satisfaction correlate with

Table 6 Correlation matrix between scores and usability values (* $p < 0.05$; ** $p < 0.03$; *** $p < 0.00$)

Metric	Paradigm	Effectiveness	Task completion time	Satisfaction
Overall	DDC-based	0.45	−0.81***	0.74**
	Device-tailored	0.5	−0.88***	0.67*
Navigation	DDC-based	0.42	−0.70*	0.73**
	Device-tailored	0.55	−0.82***	0.82***
Layout	DDC-based	−0.09	0.44	0.08
	Device-tailored	−0.31	0.43	−0.52
Page definition	DDC-based	0.07	−0.28	0.18
	Device-tailored	0.08	−0.27	0.17
Input	DDC-based	0.47	−0.94***	0.49
	Device-tailored	0.48	−0.94***	0.50

$\rho(9) = 0.73$, $p < 0.03$ for the DDC-based evaluation and $\rho(9) = 0.82$, $p < 0.00$ for the device-tailored one.

8 Conclusions

This paper has presented an assessment framework for the mobile web. Mobile Web Best Practices are the criteria used for evaluation purposes, and the Logic Scoring Preferences measurement method has been adapted to the particular features of the evaluation framework and reporting issues. The assessment framework can be deployed in several situations that range from engineering mobile web content to using scores for the adaptive web, demonstrating that the framework benefits both developers and end-users.

The heterogeneity of existing mobile devices causes inaccuracies in evaluation results that rely on mobile web guidelines. This is the motivation for designing an assessment framework that not only considers mobile web guidelines, such as Mobile Web Best Practices 1.0, but also the specific hardware and software characteristics of a particular device. Based on the Logic Scoring Preferences method, 5 metrics are defined: one for each MWBP 1.0 guideline (*Navigation*, *Page layout*, *Definition* and *Input*) and the *Overall* value. Therefore, exploiting data from device-tailored evaluation reports also makes it possible to automatically obtain a device-adapted usability score. Even though the relevance of each best practice has not been considered, successful results have been obtained. However, it is foreseen that future work will deal with best practice relevance and how the weighting of each checkpoint affects the scores and their relationship with usability metrics. This will entail a user test in order to obtain the perceived relevance of each checkpoint and will lead to a higher accuracy in the results. Another problem that arises is that the evaluation framework only considers default software configuration of the device. Therefore, if the user had later installed or updated a user agent or software, the related information would be outdated, as device data repositories only store the preliminary configuration of devices. Given that relying on automatic review tools and obviating human judgment introduces an error rate into the evaluation, the scores thus represent a lower-bound level of conformance.

In order to discover the behaviour of mobile web pages and desktop web pages with respect to the assessment method presented herein, 102 web pages were evaluated and measured. Two mobile devices were considered in the process: one device that has fewer capabilities than the DDC and another device that has much better features than the DDC. The results show that when it comes to mobile web pages, the scores for *Page Layout* are much higher for the device with better support, mainly due to its greater screen size. High scores are obtained for *Page Layout* and *User Input* guidelines, which indicate that the web pages display properly, forms are well structured and data input is relatively straightforward. However, the results for *Navigation* show that navigation mechanisms still remain an issue. This can be fixed by providing shortcuts and using file formats for pictures that are widely supported by most devices. Regarding page definition, XHTML labels for styling purposes should be replaced by style sheets, as most devices support them these days. Both devices obtain acceptable scores for *Overall*, which indicates that web pages developed with mobile devices in mind adhere to a certain extent to design best practices. Regarding desktop web pages, both devices score very low for *navigation*. This can be explained as there is an overload of content that cannot be correctly browsed due to the lack of a pointing device and the small screen size. The most significant difference between mobile and desktop pages concerns *User Input* and *Page Layout*, where much lower values are obtained for desktop pages. The device with less support scores 13 points lower in desktop web pages for *User Input* and it is even worse for *Page Layout*, obtaining a mean of 18 points less, while the device with better support scores 26 points higher. Statistical analysis reveals that the results are significantly different when it comes to comparing web pages for desktop and mobile devices, since mobile web pages score higher. In addition, web pages score much higher when more capable devices are considered.

Nine sites were selected to carry out a user test to discover the relationship between automatically obtained mobile web conformance scores and usability attributes. Using a PDA, 20 participants set out to find a specific target in each site. Effectiveness, task completion time and user satisfaction were measured. The objective was twofold: to find the relationship between the automatically obtained scores and usability and to check which of the two approaches (DDC-based vs. device-tailored) had a stronger relationship with usability. The scores produced by the assessment framework strongly correlate with the objective (task completion time) and subjective (satisfaction) usability metrics. While DDC-based measurement yields satisfactory results, device-tailored evaluation and measurement has shown that it adjusts better to the user's performance, as most correlations are substantially stronger. In this sense, it can also be concluded that the method proposed here to automatically assess mobile web guidelines conformance (evaluation and measurement) adequately captures the usability perceived by the user interacting with a specific device and their performance.

This means that the proposed metrics adequately capture the essence of MWBP in the proposed two evaluation scenarios, DDC-based and device-tailored. It can also be concluded that best practices in MWBP adequately capture usability issues in terms of task completion time and satisfaction.

The results obtained provide evidence that adherence to Mobile Web Best Practices produces an increase in the usability level. However, it cannot be guaranteed that this relationship is caused by conformance to the mobile web guidelines. For instance, it is not known whether the lack of usability is caused by the violation of mobile web guidelines. Alternatively, this can be explained as a consequence of the awareness of skilled mobile web developers. In other words, those who develop mobile web guidelines conformant pages also create usable mobile content and vice versa. It can also be explained that actual conformance entails usability. In this case, the empirical conclusion contradicts the general belief that conformance to web guidelines does not entail usable pages in practice, at least for web accessibility [20]. However, this contradiction could occur because this framework does capture the interaction context, as suggested by Sloan et al. [35]. It could also occur because those mobile web guidelines that do not overlap with web accessibility guidelines make a difference in favour of usability or because the user test was conducted with able-bodied users.

Appendix

See Table 7.

Table 7 URLs used in the user test

identifier	URL
dallasnews	http://www.dallasnews.com/mobile/
edmunds	http://pda.edmunds.com/
popurls	http://popurls.mobi/
nypost	http://www.nypost.com/avantgo/
drinkboy	http://www.drinkboy.com/offline/index.html
wapedia	http://wapedia.mobi/en/
houston	http://mobile.chron.com
elpasotimes	http://m.elpasotimes.com/
bostontimes	http://mobile.boston.com/

References

1. Abou-Zahra, S., Squillace, M.: Evaluation and report language (EARL) 1.0 schema. W3C Evaluation and Repair Tools Working Group. http://www.w3.org/TR/EARL10-Schema/ (2009). Accessed Sep 2009
2. Bailey, J., Burd, E.: Tree-map visualisation for web accessibility. Computer Software and Applications Conference, COMPSAC'05, pp. 275–280 (2005)
3. Bickmore, T.W., Schilit, B.N.: Digestor: device-independent access to the World Wide Web. Computer Networks and ISDN Systems **29**, 1075–1082 (1997)
4. Buchanan, G., Farrant, S., Jones, M., Thimbleby, H.: Improving mobile internet usability. International Conference on World Wide Web, WWW'01, pp. 673–680 (2001)
5. Buyukkokten, O., Garcia-Molina, H., Paepcke, A., Winograd, T.: Power browser: efficient web browsing for PDAs. SIGCHI Conference on Human Factors in Computing Systems, CHI'00, pp. 430–437 (2000)
6. Chisholm, W., Vanderheiden, G., Jacobs, I.: Web content accessibility guidelines 1.0. W3C Web accessibility initiative. http://www.w3.org/TR/WAI-WEBCONTENT/ (1999). Accessed Sep 2009
7. Church, K., Smyth, B., Cotter, P., Bradley, K.: Mobile information access: a study of emerging search behavior on the mobile internet. ACM Trans. Web **1**(1), article 4 (2007)
8. Coursaris, C.K., Kim, D.J.: A Research agenda for mobile usability. 26th Conference on Human Factors in Computing Systems, CHI'07, pp. 2345–2350 (2007)
9. Cui, Y., Roto, V.: How people use the web on mobile devices. International Conference on World Wide Web, WWW'08, pp. 905–914 (2008)
10. DeMarco, T.: Controlling Software Projects: Management, Measurement and Estimates. Prentice Hall, Upper Saddle River, NJ (1986)
11. Dujmovic, J.J.: Neural networks—concepts, applications, and implementations. In: Antognetti, P., Milutinovic, V. (eds.) Preferential Neural Networks, pp. 155–206. Prentice Hall, Upper Saddle River, NJ (1991)
12. Dujmovic, J.J.: A Method for evaluation and selection of complex hardware and software systems. 22nd International Computer Measurement Group Conference, pp. 368–378 (1996)
13. Garofalakis, J., Stefanis, V.: MokE: a tool for Mobile-ok evaluation of web content. International Cross-Disciplinary Conference on Web Accessibility, W4A'08, pp. 57–64 (2008)
14. Harper, S.: Mobile web: reinventing the wheel? ACM SIGACCESS Access. Comput. **90**, 16–18 (2008)
15. Hunter, J., Lagoze, C.: Combining RDF and XML schemas to enhance interoperability between metadata application profiles. International World Wide Web Conference, WWW'01, pp. 457–466 (2001)
16. ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability. International Organization of Standardization (1998)
17. International Telecommunication Union, ITU: Worldwide mobile cellular subscribers to reach 4 billion mark late 2008. http://www.itu.int/newsroom/press_releases/2008/29.html (2008). Accessed Sep 2009
18. Jain, R.: The mobile web in developing countries. W3C Workshop on the Mobile Web in the Developing Countries (2006)
19. Kaikkonen, A., Roto, V.: Navigating in a mobile XHTML application. SIGCHI Conference on Human Factors in Computer Systems, CHI 2003, pp. 329–336 (2003)
20. Kelly, B., Sloan, D., Phipps, L., Petrie, H., Hamilton, F.: Forcing standardization or accommodating diversity?: a framework for applying the WCAG in the real world. International Cross-Disciplinary Workshop on Web Accessibility, W4A'05, pp. 46–54 (2005)
21. Klyne, G., Reynolds, F., Woodrow, F., Ohto, H., Hjelm, J., Butler, M., Tran, L.: Composite capability/preference profiles

(CC/PP): structure and vocabularies 1.0. W3C Device Independence Working Group. http://www.w3.org/TR/2004/REC-CCPP-struct-vocab-20040115/ (2004). Accessed Sep 2009
22. Leporini, B., Paternò, F.: Applying web usability criteria for vision-impaired users: does it really improve task performance? Int. J. Human-Comput. Interact. **24**(1), 17–47 (2008)
23. Lewis, J.R.: IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int. J. Human-Comput. Interact. **7**(1), 57–78 (1995)
24. Mankoff, J., Dey, A., Batra, U., Moore, M.: Web accessibility for low bandwidth input. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'02, pp. 17–24 (2002)
25. Mich, L., Franch, M., Gaio, L.: Evaluating and designing web site quality. IEEE Multimed **10**(1), 34–43 (2003)
26. Nielsen, J., Tahir, M.: Homepage Usability: 50 Websites Deconstructed. New Riders Publishing, Indianapolis (2000)
27. Nielsen, J.: Mobile Web 2009 = Desktop Web 1998. Jakob Nielsen's Alertbox. http://www.useit.com/alertbox/mobile-usability.html (2009). Accessed Sep 2009
28. Olsina, L., Rossi, G.: Measuring web application quality with WebQEM. IEEE Multimed. **9**(4), 20–29. IEEE Computer Society Press (2002)
29. Owen, S., Rabin, J. W3C mobileOK Basic Tests 1.0. W3C Mobile Web Initiative. http://www.w3.org/TR/mobileOK-basic10-tests/ (2008). Accessed Sep 2009
30. Rabin, J., McCathieNevile, C.: Mobile web best practices 1.0. W3C Mobile Web Initiative. http://www.w3.org/TR/mobile-bp/ (2008). Accessed Sep 2009
31. Rohra, P.: Re: People with disabilities using mobile devices to interact with the Web. WAI-IG mailing list. http://lists.w3.org/Archives/Public/w3c-wai-ig/2008AprJun/0099.html (2008). Accessed Sep 2009
32. Roto, V.: Web browsing on mobile phones—characteristics of user experience. Dissertation, Helsinki University of Technology (2006)
33. Scheppe, K.: mobileOK Pro Tests Version 1. W3C Mobile Web Initiative http://www.w3.org/2005/MWI/BPWG/Group/TaskForces/mobileOKPro/drafts/ED-mobileOK-pro10-tests-20080610 (2008). Accessed Sep 2009
34. Shrestha, S.: Mobile web browsing: usability study. International Conference on Mobile Technology, Applications, and Systems, MC'07, pp. 187–194 (2007)
35. Sloan, D., Heath, A., Hamilton, F., Kelly, B., Petrie, H., Phipps, L.: Contextual web accessibility—maximizing the benefit of accessibility guidelines. International Cross-Disciplinary Workshop on Web Accessibility, W4A'06, pp. 121–131 (2006)
36. Sullivan, T., Matson, R.: Barriers to use: usability and content accessibility on the Web's most popular sites. ACM Conference on Universal Usability, CUU'00, pp. 139–144 (2000)
37. Thatcher, J., Burks, M.R., Heilmann, C., Henry, S.L., Kirkpatrick, A., Lauke, P.H., Lawson, B., Regan, B., Rutter, R., Urban, M., Waddell, C.D.: Web Accessibility: Web Standards and Regulatory Compliance. Springer, New York (2006)
38. Trewin, S.: Physical usability and the mobile web. International Cross-Disciplinary Workshop on Web Accessibility, W4A'06, pp. 109-112 (2006)
39. Vigo, M., Arrue, M., Brajnik, G., Lomuscio, R., Abascal, J.: Quantitative metrics for measuring web accessibility. International Cross-Disciplinary Workshop on Web Accessibility, W4A'07, pp. 99–107 (2007)
40. Vigo, M., Kobsa, A., Arrue, M., Abascal, J.: User-tailored web accessibility evaluations. ACM Conference on Hypertext and Hypermedia, HYPERTEXT'07, pp. 95–104 (2007)
41. Vigo, M., Aizpurua, A., Arrue, M., Abascal, J.: Evaluating web accessibility for specific mobile devices. International Cross-Disciplinary Conference on Web Accessibility, W4A'08, pp. 65–72 (2008)
42. Vigo, M., Brajnik, G., Arrue, M., Abascal, J.: Tool independence for the web accessibility quantitative metric. Disabil. Rehabil. Assist. Technol. **4**(4), 248–263 (2009)
43. Vigo, M., Leporini, B., Paternò, F.: Enriching web information scent for blind users. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS'09, pp. 123–130 (2009)
44. Yesilada, Y., Chuter, A., Henry, S.L.: Shared web experiences: barriers common to mobile device users and people with disabilities. W3C Web accessibility initiative. http://www.w3.org/WAI/mobile/experiences (2008). Accessed Sep 2009