How Much Does Expertise Matter?

A Barrier Walkthrough Study with Experts and Non-Experts

Yeliz Yesilada[1], Giorgio Brajnik[2] and Simon Harper[1]
[1]School of Computer Science
University of Manchester
Manchester, UK
firstname.lastname@manchester.ac.uk
[2]Dip. di Matematica e Informatica
Università di Udine
Udine, Italy
giorgio@dimi.uniud.it

ABSTRACT

Manual accessibility evaluation plays an important role in validating the accessibility of Web pages. This role has become increasingly critical with the advent of the Web Content Accessibility Guidelines (WCAG) 2.0 and their reliance on user evaluation to validate certain conformance measures. However, the role of expertise, in such evaluations, is unknown and has not previously been studied. This paper sets out to investigate the interplay between expert and non-expert evaluation by conducting a Barrier Walkthrough (BW) study with 19 expert and 51 non-expert judges. The BW method provides an evaluation framework that can be used to manually assess the accessibility of Web pages for different user groups including motor impaired, hearing impaired, low vision, cognitive impaired, etc. We conclude that the level of expertise is an important factor in the quality of accessibility evaluation of Web pages. Expert judges spent significantly less time than non-experts; rated themselves as more productive and confident than non-experts; and ranked and rated pages differently against each type of disability. Finally, both effectiveness, which is the quality of finding all and only true accessibility problems, and reliability, which is the repeatability of the outcomes when used in different context, of the expert judges are significantly higher than non-expert judges.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology, input devices and strategies, user-centred design, interaction styles*; K.4.2 [**Computers and Society**]: Social Issues—*Handicapped persons/special needs, assistive technologies for persons with disabilities*

General Terms

Human Factors, Experimentation

Keywords

Web accessibility guideline, evaluation

1. INTRODUCTION

This paper presents a study that investigates the role of expertise in the Barrier Walkthrough (BW) method. The BW was introduced in [2] and is an analytical technique based on heuristic walkthrough to evaluate accessibility of Web pages [24]. An evaluator has to consider a number of predefined possible barriers which are interpretations and extensions of well known accessibility principles [3]; they are assessed in a context so that appropriate conclusions about user effectiveness, productivity, satisfaction, and safety can be drawn, and severity scores can be derived. Barrier types are introduced for different user categories such as motor impairment, hearing impairment, low vision, blind, cognitive impairment, etc. [2, 1]. The BW, therefore, provides a systematic approach for manual accessibility evaluation.

The role of manual accessibility evaluation of Web pages has become increasingly critical with the advent of the Web Content Accessibility Guidelines (WCAG) 2.0 [3]; and their reliance on user evaluation to validate certain conformance measures. Manual evaluation would seem to be a skilled task which is valid only if performed by trained and experienced accessibility professionals. The role of evaluator, in such Web accessibility evaluations, is unknown and has not previously been studied. When we look at the usability field, we can see that a number of studies have been conducted to investigate the role of evaluator in methods such as cognitive walkthrough, heuristic evaluation and think-aloud method [13, 12, 16, 14, 20, 11]. These studies have shown that user testing methods may fail in yielding consistent results when performed by different evaluators.

To address the gap in the *Web accessibility field*, this paper presents a study with 19 expert and 51 non-expert judges that aims to investigate the role of expertise in the BW method. In particular, this study aims to answer the following research questions: 1. Is there any difference between the true barrier types identified by expert and non-expert judges? 2. Is there any difference between the severity ratings of experts and non-experts? 3. Is there any difference in the validity of the BW method when pages are evaluated by experts or non-experts?

This study shows that *expertise matters*. Expert judges spent significantly less time, they found themselves more productive and confident than non-expert judges. Expert judges, compared to non-expert judges, rank Web sites differently, and rate differently against each type of disability. When we look at the validity of the BW as a function of effectiveness and reliability, both effectiveness and reliability of the expert judges are significantly higher than non-expert judges. Effectiveness refers to how good a method is to find all and only true accessibility problems, whereas reliability refers to how repeatable are the outcomes of a method when used in different context, for instance by a different evaluator. In the wider research

context, these results suggest that for manual evaluation, for good quality evaluation, one needs to be careful about the experience level of the evaluator.

2. WEB ACCESSIBILITY EVALUATION

Different methods exist to assess the accessibility of Web pages including standards review, user testing, subjective assessments and screening techniques [10, 7, 15]. These methods differ in terms of their effectiveness, efficiency, reliability, usefulness and the evaluator effect. Unfortunately, not many Web accessibility evaluation techniques and tools are evaluated with respect to these; the case is quite different for usability evaluation methods. Several studies of *usability* evaluation methods have shown that user testing methods may fail in yielding consistent results when performed by different evaluators [22, 12] and that inspection-based methods are not free of shortcomings either [24, 25, 6, 9]. The evaluator effect has also been studied across a variety of usability evaluation techniques such as cognitive walkthrough, heuristic evaluation and think-aloud study [13, 12, 16, 14, 20, 11]. The evaluator effect occurs when different evaluators evaluating the same system detect substantially different sets of usability issues [16]. According to Hertzum and Jacobsen [13], the average agreement between usability evaluators of the same system using the same technique varied between %5 to %65.

Some researchers have also investigated the factors involved in the evaluator effect [13, 14, 18, 5]. Hertzum and Jacobsen [13] argue that the vague evaluation procedures may make evaluators focus on different things during the evaluation. Ling and Salvendy [18] show that there is a clear effect of the evaluator's cognitive style on heuristic evaluation and Catani and Bears [4] demonstrate that the individual judgements of severity were highly personal. Similar to these, Hertzum and Jacobsen [13] conclude that "the principal cause for the evaluator effect is that usability evaluation is a cognitive activity which requires that the evaluators exercise judgement". Hornbeak and Frokjear [14] shows that evaluators occasionally fail to observe the evidence of a particular problem.

Even though the phenomenon of evaluator effect has been extensively studied in the usability field, unfortunately, that is not the case for Web accessibility evaluation methods. One of the few studies touching on this issue is [21], which showed that while participants and experimenters agreed substantially on assigning severities to problems found via empirical methods, the agreement on these severities with checkpoint priorities in WCAG 1.0 was extremely poor. The same happened with respect to usability guidelines. This result suggests that it's extremely inaccurate to use fixed predefined priorities/severities. But an important function of an evaluator is to find out what the consequences of such defects are; and this could be done reliably provided that appropriate use scenarios are considered. Rating the severity of problems thus exacerbates the evaluator effect.

Mankoff *et al* [19] compare different accessibility evaluation methods; they found, for example, that Web developers using the screen monitor together with the screen reader were able to find a fraction of the true problems that is comparable to conformance reviews using WCAG guidelines. However, authors explicitly wanted to take the expertise factor out of the analysis.

In her review of accessibility evaluation methods, Lang [17] discusses the skill requirement of methods based on expert reviews, and for concludes that "[These methods] require evaluators to have a greater skill level to review, understand guidelines and recommend solutions".

In this paper we present a study that investigates this phenomenon for the Barrier Walkthrough (BW) method. The BW method can be used to evaluate Web pages in a specific context. Context comprises user categories (like blind users), Web page usage scenarios (like using a given screen reader), and user goals (corresponding to use cases). An *accessibility barrier* is any condition that makes it difficult for people to achieve a goal when using the Web page in the specified context. A barrier can be described in terms of i) the user category involved, ii) the type of assistive technology being used, iii) the goal that is being hindered, iv) the features of the pages that raise the barrier, and v) further effects of the barrier on payoff functions. The BW prescribes that severity is graded on the 3-point ordinal scale {minor, significant, critical}, and is a function of impact (the degree to which the user goal cannot be achieved within the considered context) and persistence (the number of times the barrier shows up while a user is trying to achieve that goal). Potential barriers to be considered are derived by interpretation of relevant guidelines [3] and principles [7]; more details are available at [1]. There are two major benefits of the BW method compared to conformance review: by listing possible barriers grouped by user categories, evaluators are more constrained in determining whether the barrier actually occurs. Secondly, by forcing evaluators to consider usage scenarios, an appropriate context is available to them for rating severity of the problems found.

Ideally, a good method is a dependable tool that yields accurate predictions of all the accessibility problems that may occur in a Web page. This is why methods can be compared in terms of such criteria as *effectiveness* (how well the method can help in identifying all and only the true problems), *reliability* (the extent to which independent evaluations produce the same results), *efficiency* (the amount of resources expended to carry out an evaluation that leads to specified levels of effectiveness and usefulness), *usefulness* (the effectiveness and usability of the produced results) and the method's *usability* (how easily it can be understood, learned and remembered by evaluators) [24, 8, 9].

3. EMPIRICAL STUDY

Since we expect the evaluator effect to be stronger in methods involving many decisions by evaluators, in the BW method the evaluator effect should occur not only because evaluators have to decide which the actual barriers in a page are, but also because they have to rate severity of such barriers.

To understand what is the overall role of expertise in the BW method, we want to investigate different research questions:

1. **True barrier types:** Is there any difference between the true barrier types identified by expert and non-expert judges? This question aims at finding out if non-expert judges are inconsistent in the barrier types they identify. If they are, then we need to revise the definition of these barrier types to minimise the gap between experts and non-experts.
2. **Severity ratings:** Is there any difference between the severity ratings of experts and non-experts? This aims to investigate if barriers are rated differently by expert judges compared to non-experts. If they are, then judges need to be trained for severity rating before they conduct such studies; or better rating methods should be devised.
3. **Validity of the BW method:** Is there any difference in the validity of the BW method when pages are evaluated by experts or non-experts? Ideally effectiveness and reliability should not change. If they do, then we can conclude that interpreting barriers is more complex, more error prone and more subjective to expertise level. In this way, practitioners wanting to adopt the BW method are aware that perhaps the BW method used in one context is more reliable than others, and know what possible error and agreement rates are; simi-

larly for instructors teaching how to use the BW method; or managers reading reports based on the BW method.

Besides these questions, more specifically, we want to test the following hypotheses, and the corresponding quantitative changes induced by expertise:

H1 expertise does not affect overall assignment of barriers to pages. In other words, we believe that effects of expertise in a sense are equally distributed across pages, and do not expect to see big differences;

H2 expertise does not affect the way in which severity ratings of barriers are distributed across pages;

H3 expertise improves effectiveness;

H4 expertise improves reliability;

H5 expertise interacts with different user categories; we believe that expertise of evaluators may be unbalanced: for example, with respect to vision-related impairments, expertise of an evaluator may be high, while being lower with motor impairments.

3.1 Participants

Nineteen expert judges (15 male and 4 female) took part to the study, which were aged between 27-72 with a mean of 40 and standard deviation of 11.4. Expert judges were highly experienced in testing websites for accessibility: several of them were recruited among attendees of the 10th ACM Conference on Computers and Accessibility, ASSETS 2008. In average, their subjective ratings of knowledge in Web accessibility is 4.6 (sd=0.6), 47% worked as a Web accessibility consultant and 67% tested more than 10 websites in the last six months. Fifty two non-expert judges (39 male and 11 female; 2 didn't submit the demographic questionnaire) took part to the study; they were aged between 21-46 with a mean of 24 and standard deviation of 3.9. Non-expert judges were mainly students who at that time were attending a course about Web accessibility and usability evaluation. In average, their subjective ratings of knowledge in Web accessibility is 2.3 (sd=0.9), none of them worked as a Web accessibility consultant and only 2% (i.e. 1 person) tested more than ten websites in the last six months.

With global data, when we look at the relationship between rating of knowledge of Web accessibility, being a Web accessibility consultant and the number of websites tested, we can see relationships between these. There is a significant association between being a consultant and the number of websites tested ($\chi^2(1) = 20.3, p < 0.0001$, Cramer's $\phi = 0.53$); there is a stronger significant association between the subjective rating of knowledge of Web accessibility and the number of websites tested ($\chi^2(4) = 39.3, p < 0.0001, \phi = 0.74$) and there is a significant relationship between knowledge of accessibility rating and being a consultant ($\chi^2(4) = 40, p < 0.0001, \phi = 0.75$).

3.2 Materials

Both expert and non-expert judges were asked to apply the BW method to the following pages: 1. "I love God Father movie" Facebook group; 2. The Godfather at IMDB; 3. Hall's Harbour Quilts, Halifax; 4. Sam's Chop House Manchester. These pages were chosen because they are typical and represent both professionally designed and hand-crafted pages. Facebook and the Internet Movie Database (IMDB) pages are in the top 100 most widely used pages ranked by Alexa. Even though the last two pages are not in the top 100, these pages are typical long-tail pages [23]. They are not as widely used as Facebook or IMDB but they are within the interests of a small community.

In this study, each judge evaluated one page, except for two expert judges who evaluated two pages each; the pages assigned to judges were randomised [1]. Barriers tested in the study can be found at [1]. Each judge was given a sheet with a randomized list of barriers to counterbalance order effects. The same list was repeated for each of the user categories considered.

3.3 Procedure

When participants accepted to take part in this study, they were given a judge number and asked to follow the instructions on the experiment Web page[1]. They completed the evaluation in their own time and working environment. They were asked to follow a procedure with the following three stages:

1. **Introduction:** Participants were asked to read an information sheet[1] and asked to answer some screening questions about demographics and expertise.
2. **Main:** By using the given judge number, participants were first asked to download the corresponding barriers sheet and evaluate the appropriate Web page by filling in that sheet. They were allowed to use any evaluation tool, browser extension or technique they liked. Participants were asked to evaluate each barrier with respect to blind, low vision, motor impaired users or users of small mobile devices. For each barrier and user category, they were asked to check whether that barrier exists. If it did not exist then they were asked to enter 0 or leave blank; if it existed they were asked to specify the severity based on three point scale (1=minor, 2=significant, 3=critical) and also explain the rationale for their rating.
3. **3. Conclusion:** Participants were asked to fill in a post evaluation questionnaire. This questionnaire aims to capture how long it took to complete the evaluation, the tools and techniques used and participants' subjective rating of the level of effort, productivity required and their confidence in their evaluations.

4. RESULTS

In total there were 21 evaluations by our expert judges and 52 by our non-expert judges. Table 1 summarises the total number of evaluations along with the mean values and standard deviations of subjective ratings and completion time of our judges. As suggested by the table:

- Our expert judges spent significantly less time ($M = 107$ vs. $M = 298$ min., two tailed t-test $T(53.2) = 8.6, p < 0.0001$ with a large effect size $\mathbf{d} = 2.22$) than non-expert judges;
- Expert judges found themselves slightly more productive ($M = 3.3$ vs. $M = 2.8$ in a range from 1=very low to 5=very high, $T(31) = 2.26, p = 0.0307, \mathbf{d} = 0.59$) than non-expert judges;
- Expert judges had more confident ($M = 3.9$ vs. $M = 2.5$, $T(39.2) = 6.3, p < 0.0001, \mathbf{d} = 1.63$) than non-expert judges;
- No difference emerges for the perceived effort.

4.1 True Barrier Types

First of all, we need make an important design decision on how to identify correct answers provided by participants; we need this in order to identify the set of true barriers, given a page and a user category. Luckily, this time we could tap on opinions and judgements provided by 19 leading experts in the field. However, as often happens with accessibility, there is also disagreement. To cope with such an unavoidable subjectivity, we adopted a majority rule to determine when a barrier was correctly identified. More specifically,

[1] http://hcw.cs.manchester.ac.uk/research/riam/experiments/riam-samba.php

Judge	Total evaluations	Time (min.)	Effort	Confidence	Productivity
Experts	21	107 (75)	3.4 (0.9)	3.9 (0.9)	3.3 (0.9)
Non-experts	52	298 (108)	3.4 (0.8)	2.5 (0.9)	2.8 (0.7)
All	73	243 (133)	3.4 (0.8)	2.9 (1.1)	2.9 (0.8)

Table 1: Simple means of the subjective ratings of the judges (1-very low, 5-very high; standard deviations are in parenthesis).

given a page and a user category, a barrier is *correctly identified* if the majority of experts who rated it agreed on whether its severity was 0 or greater than 0 (regardless of whether they said 1, 2 or 3).

The overall method can be summarised as follows: for each user category, we take the union of all barriers identified on each page for that user category; for instance, for users with low vision impaired users, we take the union of all the true barrier types identified on each page for this user category (`LV.all` = IMDB.LV ∪ Quilts.LV ∪ Facebook.LV ∪ Sams.LV).

With this method, Table 2 shows the list of true barrier types identified for each user category. As can be seen from this table, expert judges correctly identified 27 barrier types for low vision users, 26 barrier types for blind users, 19 for motor impaired users and 26 for mobile users. When we compare this data with the non-expert judges data, non-expert judges missed one barrier type for mobile users: "New windows", three for motor impaired users: "Inflexible page layout", "Page size limit" and "New windows", one for low vision: "Forms with no label tags" and two for blind users: "Data tables with no structural relationship" and "Page size limit".

When we also take a look at each Web page and user category, we can also see some differences between the barrier types identified by experts and non-experts; non-expert judges missed some barrier types. In summary, they missed most barrier types on Sams and they almost identified all barrier types that the experts identified on Quilts. On Facebook, they only missed the "No stylesheet support" (for low vision and mobile users). On IMDB, they missed the following three barrier types: 1. "Data tables with no structural relationship" (for blind users); 2. "Page size limit" (for blind users); 3. "Forms with no label tags" (for low vision users). On Quilts, they only missed the barrier type called "too many links" (for mobile users). Finally, on Sams, they missed the following five barrier types: 1. "Language markup" (for blind users); 2. "New windows" (for low vision, motor impaired and mobile users); 3. "Inflexible page layout" (for motor impaired users); 4. "Internal links are missing" (for motor impaired users); 5. "Functional images lacking text" (for mobile users).

4.2 Severity Ratings

To compare ratings provided by judges we use weighted severity: severity levels are categorical values {`none`, `minor`, `significant`, `critical`} encoded into {0, 1, 2, 3}. To compare pages it is convenient to transform them into numeric values, and this can be done arbitrary (provided that weights are increasing values). For simplicity here we used the weights $\{0, 1, 2, 3\}$, which implies that, for example, a critical barrier is 3 times more "important" than a minor one. Thus an average weighted severity of 0.4 for a page, for example, means that the page has the equivalent of $0.4 \cdot 61 \cdot 4 = 97.6$ minor barriers (61 barrier types times 4 user categories).

Figure 1 illustrates the values that are given in detail by Table 3. The mean weighted severity for experts is 0.344 (sd=0.747), whereas for non-experts it is 0.323 (sd=0.764). Zero is the most frequent severity rating given by judges, which means that weighted severity is positively skewed and hence not normally distributed. The difference between experts and non-experts is significantly according to Wilcoxon's test ($W > 10^8, p < 0.0001$). A systematic pairwise comparison (with adjusted p-values) shows that all pairs of user categories differ except for low vision vs. mobile, whose difference is not significant at the 5% level. Similarly for pages, where the only two pairs whose difference is not significant are Facebook vs. Quilts, and IMDB vs. Sams.

Barrier Type	Low vision	Blind	Motor	Mobile
1. Ambiguous links	√	√	√	√
2. Cascading menu	√	√	√	√
3. Data tables with no structural relationships		√		
4. Data tables with no summary		√		
5. Dynamic menu in Javascript	√	√	√	√
6. External Resources				√
7. Forms with no LABEL tags	√	√		
8. Functional images lacking text	√	√	√	√
9. Images used as titles	√	√		√
10. Inflexible page layout	√		√	√
11. Insufficient visual contrast	√			
12. Internal links are missing	√	√	√	√
13. Language markup		√		
14. Large graphics				√
15. Layout tables	√	√		√
16. Links/button are too small	√		√	
17. Links/button too close to each other			√	
18. Long URIs	√	√	√	√
19. Minimize markup	√	√	√	√
20. Missing layout clues	√	√		√
21. Mouse events	√	√	√	√
22. Moving content	√			
23. New windows	√	√	√	√
24. No cookies support				√
25. No keyboard shortcuts	√	√	√	
26. Non separated links		√		
27. No page headings	√	√	√	√
28. No stylesheet support	√			√
29. Page Size Limit	√	√	√	√
30. Rich images embedded in the background		√		
31. Rich images lacking equivalent text		√		√
32. Scrolling	√		√	√
33. Skip links not implemented	√	√	√	√
34. Stylesheet size				√
35. Text cannot be resized	√			
36. Too many links	√	√	√	√
37. Using stylesheets	√	√		√
38. Valid markup	√	√	√	√

Table 2: True Barrier Types for all users on all pages.

4.3 Validity

Validity of the BW method is a function of effectiveness and of reliability: *effectiveness* refers to how good a method is to find all and only the *true accessibility problems*, whereas *reliability* refers to how repeatable are the outcomes of a method when used in slightly different contexts, like different evaluators or different times. Both properties contribute to validity: a method that if effective but not consistent, i.e. not reliable would not be valid; similarly for a consistent method that is not effective.

4.3.1 Effectiveness

On the basis of the notion of correct identification of a barrier we introduced in Section 4.1, we can define several ways to measure the effectiveness of an evaluation (given a judge, a page and a user category). One way is to use the *accuracy rate*, i.e. the percent-

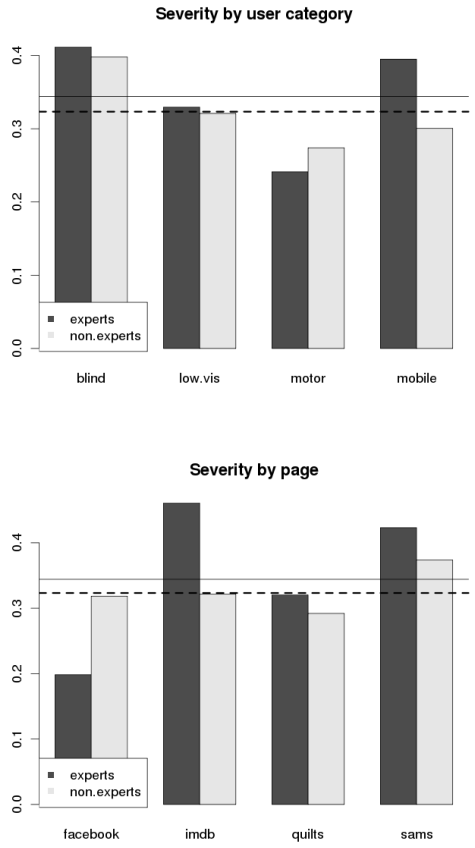

Figure 1: Weighted severities by user category (top) and by page (bottom); solid lines are the means for experts, whereas the dashed lines are for non-experts.

age of ratings that are correct over the entire set. See Table 4 for detailed values.

Over a total of 17812 ratings, judges correctly rated 81.3% of them; for experts accuracy is 86.4%, whereas for non-experts it is 79.2%; the difference is significant ($\chi^2(1) = 122, p < 0.0001$) and the 95% confidence interval of the difference is $[0.060, 0.084]$.

Accuracy rate is not the only way to measure effectiveness. Given a page and a user category, we define the *true barriers* (TB) as the set of all correctly identified barriers with severity>0 that judges found, and given a page, a user category and a judge, the *found barriers* (FB) are the set of barriers with severity>0 reported by that judge (regardless whether they are correctly identified or not).

These sets can be used to define three indexes:

	blind	low.vis	motor	mobile
Experts	0.41	0.33	0.24	0.40
Non-experts	0.40	0.32	0.27	0.30
	facebook	imdb	quilts	sams
experts	0.20	0.46	0.32	0.42
non.experts	0.32	0.32	0.29	0.37

Table 3: Mean weighted severities by user category (top) and Web page (bottom) on a scale 0 to 3.

	Accuracy %	Conf. int.
Global	81.3	$[0.807, 0.818]$
Experts	86.4	$[0.854, 0.873]$
Non-experts	79.2	$[0.785, 0.799]$

Table 4: Accuracy rates and 95% confidence intervals of accuracy: globally, for experts, for non-experts.

	facebook	imdb	quilts	sams	Sum
Wrong	141	203	195	160	699
Correct	1079	1017	1513	816	4425
Sum	1220	1220	1708	976	5124
Correct %	88	83	89	84	86
Wrong	609	696	584	750	2639
Correct	2563	2476	3076	1934	10049
Sum	3172	3172	3660	2684	12688
Correct %	81	78	84	72	79

Table 5: Accuracy for experts (top) and non-experts (bottom) across pages.

Correctness $C = \frac{|TB \cap FB|}{|FB|}$ is the proportion of reported barriers that are also correct.

Sensitivity $S = \frac{|TB \cap FB|}{|TB|}$ is the proportion of all the true barriers that were reported.

F-measure $F = \frac{2C \cdot S}{A+S}$ is the harmonic mean of A and S, which is a balanced combination of A and S summarizing the effectiveness of an evaluation.

If we group all the available ratings by judge and user category, we obtain 292 (i.e. $73 \cdot 4$) evaluations; on each of them we can compute the three indexes just defined. F-measure is normally distributed; Table 6 shows the values under different situations.

The overall mean value is 48.7%, experts reach 59% while non-experts drop to 44.5%; this difference is significant ($T(123) = 6.52, p < 0.0001$) and marked (the 95% confidence interval of the difference is $[0.101, 0.189]$, and the effect size is $\mathbf{d} = 0.843$). We can also notice that the standard deviation for experts is higher than for non-experts (18.4% rather than 13.8%).

ANOVA shows that there is main effect of user category ($F = 4.86(3), p = 0.0026$), of page ($F = 3.76(3), p = 0.0114$), and of judge type ($F = 56.41(1), p < 0.0001$); there is also interaction between page and judge type ($F = 3.51(3), p = 0.0159$); no interaction exists between judge type and user category. A pairwise T test (with p value adjustment) on page shows that only IMBD differs from SAMS ($p < 0.075$).

4.3.2 Reliability

In order to operationalize reliability of the BW method, we use reproducibility and agreement.

Reproducibility is related to the variability of ratings of a barrier type and can be defined as follows (the same definition was used in

	Mean	Std dev	Conf. int.
Global	0.487	0.166	$[0.468, 0.506]$
Experts	0.590	0.184	$[0.550, 0.630]$
Non-experts	0.445	0.138	$[0.426, 0.464]$

Table 6: F-measure means, standard deviations and 95% confidence intervals around the means: globally, for experts, for non-experts.

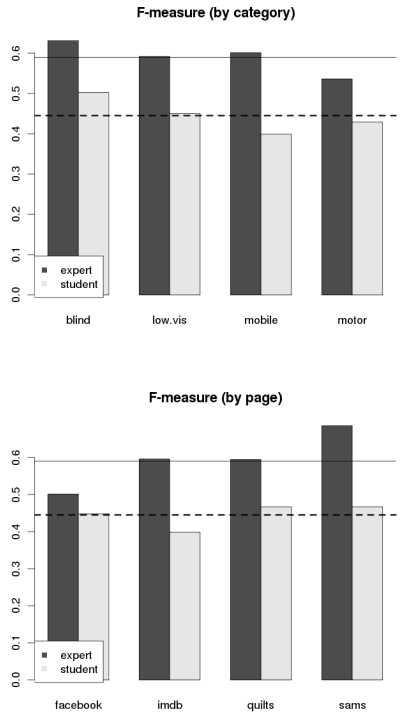

Figure 2: F-measure by user category (top) and page (bottom); solid line is the mean for experts; dashed line is the mean for non-experts.

[24]): *reproducibility* of a barrier type, given a page, a user category and a set of ratings by different judges for that barrier on that page with respect to that user category is $r = \max\{0, 1 - \frac{sd}{M}\}$, where M is the mean of weighted severity and sd is the standard deviation ($\frac{sd}{M}$ is often called *coefficient of variation*). When reproducibility is close to 1, the standard deviation is very small compared to the mean; in our case this implies that the variability of weighted ratings between our judges is low.

Reproducibility is computed on the weighted severity corresponding to each triple (barrier type, user category, page) over different judges once using the global data, once using data of only non experts, and finally once with data of only experts. Then appropriate aggregations (eg. means, medians and standard deviations) are computed. Results are shown in Figure 3; Table 7 provides numerical data. Reproducibility is not normally distributed (showing two peaks, corresponding to the two saturation points, 0 and 1).

	blind	low.vis	motor	mobile	all
Experts	0.54	0.54	0.61	0.51	0.55
Non-experts	0.39	0.34	0.45	0.36	0.38
	facebook	imdb	quilts	sams	all
Experts	0.61	0.44	0.61	0.55	0.55
Non-experts	0.38	0.38	0.44	0.34	0.38

Table 7: Mean reproducibility by user category and by page.

ANOVA tells us that there are main effects on reproducibility by judge type ($F(1) = 61.6, p < 0.0001$), by user category ($F(93) = 4.01, p = 0.0074$), and by page ($F(3) = 6.13, p = 0.0004$).

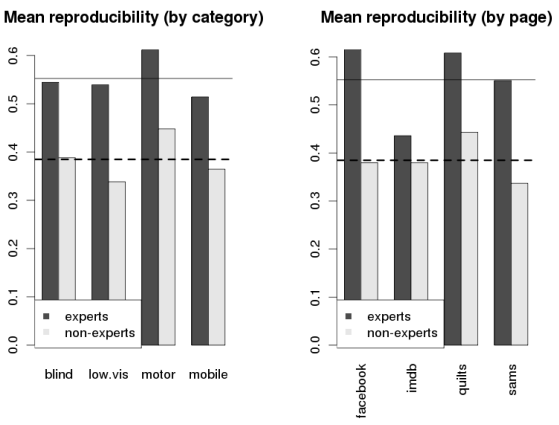

Figure 3: Mean reproducibility by user category (left) and page (right). Solid line is the mean for experts, and the dashed one for non-experts.

There is interaction between judge type and page ($F(3) = 3.48, p = 0.0153$); no interaction exists between judge type and user category. Overall reproducibility for experts is 0.55 ($\text{sd} = 0.48$), whilst dropping to 0.38 ($\text{sd} = 0.47$) for non-experts; this difference is significant (Wilcoxon test $W > 10^6, p < 0.0001$). We can also notice that experts systematically lead to more reproducible results, across user categories and across pages (the difference in mean reproducibility ranges from 6% to 24%). Pairwise Wilcoxon comparisons with adjusted p-values show that reproducibility differ significantly between Facebook and IMDB, IMDB and Quilts, Quilts and Sams; similarly, it differs significantly between low vision and motor impairments, motor impairments and mobile.

Therefore we can conclude that reproducibility is definitely affected by expertise (experts lead to more reproducible results by 17% in our sample); it also depends on the specific page being evaluated and the specific user category with respect to which the evaluation is carried out. The effect of expertise differs page by page.

There is a strong negative correlation between mean values of weighted severity and reproducibility (for experts: Spearman's $\rho = -0.827, S > 10^8, p < 0.0001$; for non-experts: $\rho = -0.7, S > 10^8, p < 0.0001$), meaning that higher severities correspond to lower reproducibility. Notwithstanding this correlation, judge type (expert/non-expert) is a factor that has a major effect on reproducibility: analysis of covariance shows that if we add "judge type" to "mean severity" as predictors in a linear model to predict "reproducibility", then the new model fits significantly better the data (i.e. "judge type" reduces the residual sum of squares by 16 and helps explaining 27% of the variance of reproducibility: multiple $R^2 = 0.272$). Therefore experts achieve higher reproducibility not only because they might have rated barriers with lower severities than non experts, but indeed because of a smaller variation of their ratings.

We introduce a second independent measure of reliability which is the level of *agreement* between judges. Rather than simply computing the mean of the correlation between pairs of judges (which would reflect only the relative agreement), we computed the *intraclass correlation coefficient* on the ratings that the set of judges gave to *all* the barriers with respect to a given page and a given user category. This index measures both the relative and absolute agreement between judges and provides a different measure of reliability

	facebook	imdb	quilts	sams	**mean**
blind	0.28	0.40	0.48	0.31	0.37
low.vis	0.29	0.23	0.28	0.33	0.28
motor	0.14	0.36	0.33	0.29	0.28
mobile	0.30	0.27	0.44	0.29	0.33
mean	0.25	0.31	0.38	0.31	0.31
blind	0.35	0.30	0.40	0.34	0.35
low.vis	0.25	0.21	0.26	0.27	0.25
motor	0.22	0.34	0.37	0.29	0.31
mobile	0.15	0.16	0.27	0.25	0.21
mean	0.24	0.25	0.33	0.29	0.28
blind	−0.07	0.10	0.08	−0.02	0.02
low.vis	0.04	0.02	0.02	0.06	0.04
motor	−0.08	0.02	−0.04	−0.00	−0.03
mobile	0.15	0.11	0.17	0.03	0.12
mean	0.01	0.06	0.05	0.02	0.04

Table 8: ICC values for experts (first table), non-experts (second table) and differences between non-experts/experts. Positive numbers mean an increase in agreement when moving from non-experts to experts. Means are taken by column or by row; the grand mean is shown in the bottom-right corner.

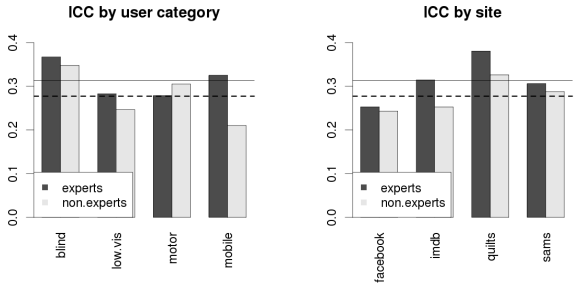

Figure 4: ICC values by user category (left) and page (right). Solid line gives the mean for experts, and the dashed one for non-experts.

than reproducibility. Details are given by Table 8 and Figure 4.

We can see that experts consistently achieve the highest agreement (overall mean is 0.31 compared to 0.28). This indicates that they are slightly more consistent in the way they rate barriers. The highest agreement for experts and non experts as well as with respect to blind user category. For both types of judges the level of agreement depends on the specific page, and the specific user category, suggesting that we should expect that specific circumstances (i.e. page and target category) are in general likely to affect agreement. Expertise does not mitigate this. Notice that for *motor impairment*, non experts reach a higher level of agreement than experts.

5. DISCUSSION

In brief, we can see from our results that expertise matters. In the previous section we have investigated different parameters to compare experts and non-experts judges, and we can see the differences due to expertise in all of these parameters.

True barrier types Non-expert judges missed some of the barrier types that were identified by the expert judges. This is the case across pages and across user categories. Therefore, we can only support **H1** partially because of the different sets of true barrier types that were identified by the two groups of judges. There could be several reasons why non-expert judges missed certain types of barrier: experts and non-experts focus on different things [13], lack of knowledge, lack of technique to capture these barriers, different cognitive skills used to interpret these barrier types [18], the definition of these barrier types were not good enough [13] or mental habits to use or adopt certain guidelines.

Severity ratings The difference between severities occurs globally and shows up also among different user categories and among pages. We have also seen that the effect of expertise on weighted severity is not constant across pages nor across user categories. This means that experts compared to non-experts will rank websites differently, and they will rate differently websites against each type of disability. Therefore, **H2** is not supported, since expertise affects the distribution of ratings.

Accuracy rate While expected, the drop in accuracy for non-experts compared to experts is relatively low (between 6 and 8.4%); such a drop tends to be consistent across different user categories, meaning that expertise tend not to interact with user categories. Accuracy also depends on page, for both kinds of judges; means can vary by 6% in both cases. It is also remarkable the relatively high level of accuracy that was achieved (which could range from 78% to 80% for non experts, and from 85% to 87% for experts). This suggest that the BW method is a good manual accessibility evaluation method as non-experts without much experience can do pretty much as good a job as experts.

F-measure It's worth noting the marked difference between judge types. The difference in F-measure between judge types ranges from 10.1 to 18.9%, which is substantial and consistent. F-measure depends also on Web page, which indicates that effectiveness of the BW method depends on which pages are being analyzed. There is interaction between judge type and pages, but not between judge type and categories: effects of expertise depend on pages, but not on user categories. We can also see that, with respect to the global mean value of F-measure (48.7%), experts can increase it to 60.1%, and non-experts can drop it to 44.5%. F-measure is a mean of correctness and sensitivity, which respectively are a function of false positives (incorrect answers produced by judges) and true negatives (correct ratings that were missed by judges); a change of 15% in F-measure means that expertise reduces both kinds of wrong answers by that amount. It is remarkable that F-measure scores are not higher than 60.1%. In other words, even experts under the best situation are bound to produce false positives and/or to miss true problems. We also see (Figure 2) that experts are systematically better than non-experts across sites and across user categories.

Effectiveness F-measure and accuracy rate are the two parameters that we have investigated for measuring effectiveness. We can see that both F-measure and accuracy are significantly different for experts and non-experts. Therefore, we can conclude that **H3** is supported by our data; expertise improves effectiveness.

Reproducibility This is affected by expertise, by user category and by the page being evaluated. The effect of expertise depends on the actual page, but not by the user category. Experts lead to more reproducible results (with an increase of 17%). Reproducibility also depends on the specific page being evaluated and the specific user category with respect to which the evaluation is carried out. The effect of expertise differs page by page, but not category by category.

Agreement Agreement (measured through ICC) of experts is 3% higher, a small difference. While agreement does depend on pages (for experts and non-experts alike) and on user categories, the effect of expertise is consistent (see rows and columns marked with "mean" in Table 8, third table). We can conclude therefore that

expertise and agreement do interact only weakly.

Reliability Reproducibility and agreement are the two parameters investigated for measuring reliability. The results show that experts have higher reliability and agreement, therefore we can conclude that **H4** is supported by our data; expertise improves reliability.

Finally, **H5** is partially supported, since the effect of expertise across categories can be noticed on reliability (reproducibility and agreement) but not on effectiveness.

6. CONCLUSIONS & FUTURE WORK

In this paper, we have presented a BW method study with 19 experts and 51 non-experts. This study shows that *expertise matter*. Expert judges spent significantly less time than non-experts; rated themselves as more productive and confident than non-experts; and ranked and rated pages differently against each type of disability. When we look at the validity of the BW method as a function of effectiveness and reliability, both effectiveness and reliability of the expert judges are significantly higher than non-expert judges. Effectiveness refers to how good a method is to find all and only true accessibility problems, whereas reliability refers to how repeatable are the outcomes of a method when used in different context, for instance by a different evaluator.

The differences could be attributed to a number of reasons: experts and non-experts focus on different things [13], lack of knowledge, different cognitive skills used to interpret these barrier types [18] or the definition of these barrier types were not good enough [13]. However, to confirm these we need to conduct further controlled studies.

In conclusion, if we consider the results presented in this paper in the wider manual accessibility evaluation context, these results suggest that since experience in accessibility is an important factor for the quality of the results, appropriate training, education of evaluators is necessary.

7. REFERENCES

[1] G. Brajnik. Barrier walkthrough: Heuristic evaluation guided by accessibility barriers. `http://users.dimi.uniud.it/~giorgio.brajnik/projects/bw/bw.html`.

[2] G. Brajnik. Web accessibility testing: When the method is the culprit. In *In Proc. of ICCHP 2006*, 2006.

[3] B. Caldwell, M. Cooper, L.G. Reid, and G. Vanderheiden. Web Content Accessibility Guidelines 2.0 (WCAG 2.0). W3C, 2008. `http://www.w3.org/TR/WCAG20/`.

[4] M.B. Catani and D.W. Biers. Usability evaluation and prototype fidelity: Users and usability professionals. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 1998.

[5] J. Cegarra and J. Hoc. Cognitive styles as an explanation of experts' individual differences: A case study in computer-assisted troubleshooting diagnosis. *Int. J. Hum.-Comput. Stud.*, 64(2):123–136, 2006.

[6] G. Cockton and A. Woolrych. Understanding inspection methods: lessons from an assessment of heuristic evaluation. In *People & Computers XV*, pages 171–192. Springer, 2001.

[7] DRC. Formal investigation report: web accessibility. Disability Rights Commission, `www.drc-gb.org/publicationsandreports/report.asp`, April 2004. Visited Jan. 2006.

[8] W. Gray and M. Salzman. Damaged merchandise: a review of experiments that compare usability evaluation methods. *Human - Computer Interaction*, 13(3):203–261, 1998.

[9] T. S. Andre H. R. Hartson and R. C. Williges. Criteria for evaluating usability evaluation methods. *Int. Journal of Human-Computer Interaction*, 15(1):145–183, 2003.

[10] S.L. Henry and M. Grossnickle. *Just Ask: Accessibility in the User-Centered Design Process*. Georgia Tech Research Corporation, Atlanta, Georgia, USA, 2004.

[11] M. Hertzum and N. Jacobsen. The evaluator effect during first-time use of the cognitive walkthrough technique. In *Proceedings of HCI International on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I*, pages 1063–1067, 1999.

[12] M. Hertzum, N. Jacobsen, and R. Molich. Usability inspections by groups of specialists: perceived agreement in spite of disparate observations. In *CHI '02 extended abstracts on Human factors in computing systems*, pages 662–663. ACM, 2002.

[13] M. Hertzum and N. Ebbe Jacobsen. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human Computer Interaction*, 13(4), 2001.

[14] K. Hornbeak and E. Frokjear. A study of the evaluator effect in usability testing. *Human-Computer Interaction,*, 23(3):251 – 277, 2008.

[15] M. Ivory and M. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computer Survey*, 33(4):470–516, 2001.

[16] N. Jacobsen, M. Hertzum, and B. John. The evaluator effect in usability tests. In *CHI '98*, pages 255–256. ACM, 1998.

[17] T. Lang. Comparing website accessibility evaluation methods and learnings from usability evaluation methods. `http://www.peakusability.com.au/about-us/pdf/website_accessibility.pdf`, Visited May 2008, 2003.

[18] C. Ling and G. Salvendy. Effect of evaluators' cognitive style on heuristic evaluation: Field dependent and field independent evaluators. *Int. J. Hum.-Comput. Stud.*, 67(4):382–393, 2009.

[19] J. Mankoff, H. Fait, and T. Tran. Is your web page accessible?: a comparative study of methods for assessing web page accessibility for the blind. In *CHI'05*, pages 41–50. ACM, 2005.

[20] J. Nielsen. Finding usability problems through heuristic evaluation. In *CHI '92*, pages 373–380. ACM, 1992.

[21] H. Petrie and O. Kheir. The relationship between accessibility and usability of websites. In *Proc. CHI 2007*, pages 397–406, San Jose, CA, USA, 2007. ACM.

[22] I. Curson S. Butler E. Kindlund D. Miller R. Molich, N. Bevan and J. Kirakowski. Comparative evaluation of usability tests. In *In Proc. of the Usability Professionals Association Conference*, 1998.

[23] B. Schwartz. *The Paradox of Choice: Why More Is Less*. Harper Perennial, 2005.

[24] A. Sears. Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, 9:213 – 234, 1997.

[25] A. Woolrych and G. Cockton. Assessing heuristic evaluation: mind the quality, not just the percentages. In *In Proc. of HCI 2000*, pages 35–36, 2000.