A Comparative Test of Web Accessibility Evaluation Methods

Giorgio Brajnik
Dipartimento di Matematica e Informatica
Università di Udine
Via delle Scienze, 206 — 33100 Udine — Italy
giorgio@dimi.uniud.it

ABSTRACT

Accessibility auditors have to choose a method when evaluating accessibility: expert review (a.k.a. conformance testing), user testing, subjective evaluations, barrier walkthrough are some possibilities. However, little is known to date about their relative strengths and weaknesses. Furthermore, what happened for usability evaluation methods is likely to repeat for accessibility: that there is uncertainty about not only pros and cons of methods, but also about criteria to be used to compare them and metrics to measure these criteria.

After a quick review and description of methods, the paper illustrates a comparative test of two web accessibility evaluation methods: conformance testing and barrier walkthrough. The comparison aims at determining merits of barrier walkthrough, using conformance testing as a control condition. A comparison framework is outlined, followed by the description of a laboratory experiment with 12 subjects (novice accessibility evaluators), and its results. Significant differences were found in terms of *correctness*, one of the several metrics used to compare the methods. Reliability also appears to be different.

Categories and Subject Descriptors: H. Information Systems H.5 Information Interfaces and Presentation (I.7) H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6).

General Terms: Human Factors, Measurement, Reliability.

Keywords: Web Accessibility, Accessibility Evaluation Method, Quality Assessment.

1. INTRODUCTION

Web accessibility can be evaluated by means of different methods, including *standards review*, *user testing*, *subjective assessments* and *barrier walkthrough* [10, 5, 2, 7]; yet little is known about their properties.

Accessibility evaluation methods (AEMs) can differ in terms of their effectiveness, efficiency and usefulness, as hinted by [15, 1, 3]. If proper information about AEMs were available, then practitioners (auditors, teams involved in quality assessment, developers) could: 1) optimize resources expended in assessing and evaluating accessibility, like time, effort, skilled people, facilities; 2) aim at predictable and controllable quality of results produced by AEMs; 3) support *sustainable accessibility* processes; and 4) standardize the content of accessibility reports.

Several studies of usability evaluation methods have shown that user testing methods may fail in yielding consistent results when performed by different evaluators [17, 12, 11] and that inspection-based methods aren't free of shortcomings either [19, 22, 4, 9]. Although accessibility and usability are two different properties, there is no reason to assume that the kind of uncertainty and mishaps that apply to usability evaluation methods should not apply to AEMs as well. The most likely conclusion is that different AEMs lead to different kinds of results revealing different levels of quality, require different levels of resources, and differ for their applicability.

To date there are few experimental studies on AEMs that could confirm or contradict these *expected* properties. The main purpose of this paper is to describe results of an experiment aimed at comparing quality of two inspection-based AEMs: *standards review* (called henceforth *conformance review, CR*) and *barrier walkthrough* (BW), the former based on the Italian requirements for web accessibility [13] and the latter being a method proposed by the author [2] and based on *heuristics walkthrough* [19]. A secondary purpose is to outline a quality framework for AEMs and to describe a first benchmark for quality of AEMs.

In [1] the BW method was investigated and some preliminary, non conclusive results were obtained. In this paper we provide more significant results, based on experimental data obtained from a more focused and specific experiment.

Results show that with respect to novice evaluators, barrier walkthrough compared to conformance review markedly improves the number of problems that are correctly detected, as well as the correctness rate; no significant difference was found for sensitivity, and reliability decreases a little.

2. ACCESSIBILITY EVALUATION METHODS

Conformance reviews, called also *expert, standards, or guidelines review* or *manual inspection* [10, 21], is the AEM by far most widely used [6]. It is based on checking if a page satisfies a checklist of criteria. It is an analytic method, based on evaluators' opinions, producing failure modes (in

ASSETS'08, October 13–15, 2008, Halifax, Nova Scotia, Canada.

the form of violated checkpoints) possibly with defects and solutions.

Conformance reviews are dependent on the chosen checklist, that range from standards issued by international bodies (like the Web Content Accessibility Guidelines, WCAG, published by the W3C), to national or state-level guidelines, to individual organizations guidelines (like those issued by IBM, SUN or SAP, for example).

Other methods for evaluating accessibility include *screening techniques*, informal empirical techniques based on using a interface in a way that some sensory, motor or cognitive capabilities of the user are artificially reduced [10]; *subjective assessments*, based on a panel of users instructed to explore and use a given website by themselves, and later report feedback on what worked for them and what did not; *user testing*, based on informal empirical usability tests with disabled people adopting the think-aloud protocol [10, 5]. One final method is *barrier walkthrough*, which is discussed below in Section 2.2.

2.1 Benefits and drawbacks of AEMs

Conformance review generally leads to the following benefits: ability to identify a rather large range of diverse problems for a diverse audience (albeit this depends on the quality of the underlying checkpoints). It is relatively cost-effective, especially when coupled with automatic testing tools, and is able to identify the defects[1] underlying the checkpoint violations. The drawbacks are that it requires skillful evaluators, and that it is not likely to distinguish reliably between important and non important accessibility problems. See [14, 16, 7, 18] for a controversial list of shortcomings of accessibility guidelines and conformance reviews.

Screening techniques are easy to use, very cheap, but also very unsystematic. Sometime however such a method is capable of achieving good levels of sensitivity [16].

The benefits of subjective assessments include its low cost, and its ability to be performed remotely in space and time. In addition, participants are able to explore the areas of the website that most suite them, which is likely to increase their motivation in using it. However the drawbacks are important: it is a method that is not systematic, both regarding the pages being tested and the criteria used to evaluate them. In addition, different users with different experience and different attitudes will report very different things about the same page.

Benefits of user testing as an AEM include [15] its capability to accurately identify usability problems, that are usually experienced by real users, and that have catastrophic consequences. The drawbacks include: relatively low efficiency, inability to highlight the defects, problems may be missed if predefined scenarios are not well chosen, it will not identify low-severity problems. In addition, it is rather complicated to set up a user testing session with disabled participants (given their requirement in terms of appropriate assistive technology, including room facilities), and the results of performing user testing is likely to be a set of usability problems that are general to all the users of the website, rather than being specific to disabled ones. In other words, the method is likely to identify a number of true, but irrelevant problems. Finally, inexperienced test facilitators are likely to introduce noisy results, and in the worst case even invalidate the test. See [18] for a discussion on the relationship between usability and accessibility.

[1]I use the term *failure mode* to mean the way in which the interaction fails; the *defect* is the reason for the failure, its cause; the *effects* are the negative consequences of the failure mode.

2.2 Barrier Walkthrough

A major difference in the methods just previewed is the role that context plays during evaluations [3]. While analytic methods appear to be inexpensive compared to empirical methods, we hypothesize that their validity and reliability depend on how context is considered.

The barrier walkthrough method [1, 2] is an accessibility inspection technique where context of website usage is explicitly considered. An evaluator has to assess a number of predefined barriers which are interpretations and extensions of well known accessibility principles; they are linked to user characteristics, user activities, and situation patterns so that appropriate conclusions about user effectiveness, productivity, satisfaction and safety can be drawn, and appropriate severity scores can be consequently derived. The method is rooted on heuristics walkthrough [19] which takes into account the context of use of the website. For BW, context comprises certain user categories (like blind persons), usage scenarios (like using a given screen reader), and user goals (corresponding to *use cases*, like submitting an IRS form).

An *accessibility barrier* is any condition that makes it difficult for people to achieve a goal when using the web site through specified assistive technology (see Figure 1 for an example). A barrier is a failure mode of the web site, 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.

barrier	users cannot perceive nor understand the information conveyed by an information rich image (*e.g.* a diagram, a histogram)
defect	an image that does not have accompanying text (as an ALT attribute, as content of the OBJECT tag, as running text close to the picture or as a linked separate page)
users affected	blind users of screen readers, users of small devices
consequences	users try to look around for more explanations, spending time and effort; they may not be able to get an information; effectiveness, productivity, satisfaction are severely affected

Figure 1: Example of barrier

Severity of a barrier depends on the context of the analysis (type of user, usage scenario, user goal). The BW method prescribes that severity is graded on a 1–2–3 scale (minor, major, critical), and is a function of *impact* (the degree to which the user goal cannot be achieved within the considered context) and *frequency* (the number of times the barrier shows up while a user is trying to achieve that goal). Therefore the same type of barrier may be rated with different severities in different contexts; for example, a missing

skip-links link may turn out to be a nuisance for a blind user reading a page that has few preliminary stuff, while the same defect may show a higher severity within a page that does a server refresh whenever the user interacts with links or select boxes.

Potential barriers to be considered are derived by interpretation of relevant guidelines and principles [7, 20]. A complete list can be found in [2].

A preliminary experimental evaluation of the BW method [1] showed that this method is more effective than conformance review in finding more severe problems and in reducing false positives; however, it is less effective in finding all the possible accessibility problems.

3. QUALITY FACTORS

AEMs can be compared on the basis of a general quality framework, that includes such criteria as effectiveness, efficiency and usefulness, applied to the results obtained through application of the method [19, 8, 11, 15, 9]. In particular, in our study we consider the following criteria, some of which are further refined into more specific ones:

Effectiveness defined as the extent to which the method can be used to yield desired results with appropriate levels of accuracy and completeness. Effectiveness is further refined into:

- **Validity** defined as the extent to which the problems detected during an evaluation are also those that show up during real-world use of the system. It is further refined into:
  - **Correctness** the percentage of reported problems that are true problems (sometimes called *precision* or *validity*).
  - **Sensitivity** the percentage of the true problems that are reported (sometimes called *recall* or *thoroughness*).
- **Reliability** the extent to which independent evaluations produce the same results.

Notice that these three criteria are related but independent. An AEM that yields high correctness may be useless if too few of the actual problems are identified (*i.e.* if it features too little sensitivity). Similarly, by itself a high level of sensitivity is not enough unless a sufficient correctness is obtained: the AEM may find all the *true* accessibility problems, but they may be swamped by a large number of false positives. Finally, an unreliable AEM is not desired since even if it features high validity, this cannot be exploited in a consistent way.

Efficiency the amount of resources (time, skills, money, facilities) that are expended/used to carry out an evaluation that leads to given levels of effectiveness and usefulness.

Usefulness is the effectiveness and usability of the results produced (with respect to users that have to assess, or to fix, or otherwise to manage accessibility of a web site).

These quality factors need to be operationalized so that they can be actually measured. In our study, we made them operational through the metrics described below, in Sect. 4.1. Although usefulness is an important quality factor, it will not be discussed any further as it falls beyond the scope of this experimental study.

4. EXPERIMENTAL PLAN

The purpose of the experiment is to compare effectiveness and efficiency of BW and CR (based on the Italian requirements [13]) and see if they differ in any of the criteria outlined above, and in particular if the results obtained in [1] are confirmed by a more formal experiment. The null hypotheses are that there are no differences in correctness, sensitivity, reliability and efficiency.

Sixteen students of my course (user centered web design) showed up for the experiment, out of 35 that were invited to. During the course, students were exposed to web accessibility, guidelines, CR and BW for about 15 lecture hours, after which they were asked to analyze given web sites and write corresponding reports.

In this *within-subjects* experiment, each of them was asked to perform two evaluations, one per method whose order was counterbalanced. Each evaluation was on two preselected pages of two websites (a home page and a student-services page of two Italian universities; websites were also counterbalanced). Subjects were assigned in appearance order to a sequence of the two tasks (CR/BW applied on website A/B, and then viceversa).

Each subject was given the tasks order, a spreadsheet with a checklist of barriers or checkpoints, a workstation with Firefox and the Web Developer bar[2], links to complete BW description and description of the CR requirements, and for BW a use case (which was the same for both websites).

Subjects, when performing BW, were asked to identify as many barriers as possible, provided they were relevant to blind users of screen readers and to motor disabled users, and rate them as minor (1), major (2), or critical (3). When performing CR they were asked to mark violated requirements. When they completed an evaluation, they were instructed to write down the elapsed time and email the spreadsheet.

Twelve of the sixteen students worked in a university lab; four worked at home.

A judge (myself) identified all barriers and violated requirements on all four pages, and rated the severity of the barriers.

4.1 Independent and dependent variables

Independent variables for this experiment include the method used (CR/BW), the tested website (A/B) and the order of application of a method (1^{st} or 2^{nd}).

Collected data include the barriers reported by an evaluation, their severity (in the scale 1-2-3), the violated requirements, whose severity is conventionally set to 1, the barriers rated by the judge (with severity ranging in 0-1-2-3; 0 was given to barriers that the judge considered to be false positive, *i.e.* erroneously identified by the subject).

Appropriate aggregation of these data produced by a single evaluation leads to the following dependent variables:

Correctness $C = \frac{|\text{IDENTIFIED} \cap \text{TRUE}|}{|\text{IDENTIFIED}|}$, that is: among the barriers/requirements that were given a positive sever-

[2]version 1.1.5, available from `http://chrispederick.com/work/web-developer`

ity by the subject, the proportion that received a positive severity also by the judge.

Sensitivity $S = \frac{|\text{IDENTIFIED} \cap \text{TRUE}|}{|\text{TRUE}|}$, that is: among the barriers/requirements that were given a positive severity by the judge, the proportion that received a positive severity also by the subject.

F-measure a combination of correctness C and sensitivity S in the range $[0, 1)$: $F = \frac{CS}{\alpha C + (1-\alpha)S}$; by setting $\alpha = 0.5$ we get $F = \frac{2CS}{C+S}$, which is what was used in this study. F is a monotonic and symmetric function of both arguments, and is normally used to represent a balanced combination of correctness and sensitivity.

Reliability Following [19], reliability is defined as $\max\{0, 1 - \frac{\text{sd}}{M}\}$, where sd is the standard deviation and M is the mean of the number of correctly identified barriers/requirements (over a set of evaluations).

Efficiency simply represented by the elapsed time in minutes during an evaluation.

5. FINDINGS

Results of four subjects were excluded due to the fact that they completed only one of the assigned tasks. These four subjects did not include those that worked at home.

The remaining 12 subjects identified 92 checkpoint violations using CR on website A and 69 on B, and 102 barriers on A and 124 on B when using BW.

This is a remarkable difference between the methods, probably due to the higher "resolution" of the BW method, that suggests potential barriers that are more finely tuned and less general than generic checkpoints.

Table 1 shows the contingency table of the differences in ratings of checkpoints and barriers by subjects and by the judge. Diagonal cells in both tables represent the number of times there was agreement; cells above the diagonal represent the number of times subjects over-rated severity, and viceversa for cells below the diagonal. Table 2 presents the same information as proportions.

For CR the number of over-rated and under-rated checkpoint violations are roughly the same and close to 73, over a total of 526 judgments, 381 of which were correctly rated; for BW the number of over/under-rated barriers is larger: 790 barriers were correctly rated, 118 were over-rated and 124 were under-rated, over a total of 1032.

	sev	
j.sev	0	1
0	294	74
1	71	87

	sev			
j.sev	0	1	2	3
0	705	19	22	21
1	45	35	12	12
2	56	23	50	32

Table 1: Number of checkpoints (left) and barriers (right) rated by the judge (j.sev) and by subjects (sev). There were no barriers with severity 3 (according to the judge).

These differences can be better appreciated if we consider that the BW method requires an additional step after the evaluator has identified a barrier, *i.e.* severity assignment. Such a step is known to be highly subjective for usability evaluation methods [11]. Even though the BW method is well structured in terms of severity assignment procedure, the actual severity of a barrier is itself dependent a lot on the experience of the user: with assistive technology, with the browser, with the WWW, with the website domain and with the actual website. As a consequence, unless the scenario adopted for the evaluation is extremely specific, the margin for subjectivity in assigning severity scores remains quite high.

	sev	
j.sev	0	1
0	0.80	0.20
1	0.45	0.55

	sev			
j.sev	0	1	2	3
0	0.92	0.02	0.03	0.03
1	0.43	0.34	0.12	0.12
2	0.35	0.14	0.31	0.20

Table 2: Same data as before expressed as row proportions.

		N	C	S	F	Time
CR	mean	7.33	0.54	0.56	0.54	44:30
	sd	2.27	0.09	0.16	0.11	18:12
BW	mean	13.67	0.72	0.62	0.66	42:10
	sd	5.21	0.19	0.23	0.21	19:54

Table 3: Means and standard deviations for number of correct ratings N, correctness C, sensitivity S, F-measure F and time (min:sec).

Figure 2 shows the actual data for correctness, sensitivity, F-measure and time, while Figure 3 shows how the data is distributed, and Table 3 gives mean values and standard deviation of the number of correctly rated barriers/checkpoints, correctness, sensitivity, F-measure and time.

BW leads to a slight improvement in correctness, sensitivity and (consequently) F-measure: the number of evaluations where BW scores better than CR tends to be higher than the other way around. Even though the mean completion time for CR is higher than that for BW, CR evaluations appear to be faster than BW in more cases. The boxplots and Table 3 show that BW leads to a higher variability of all indexes, and with the exception of time, an improvement in the median and mean. However, if we take a closer look to the F-measure (taken as a representative index of the validity of each evaluation), we can see that for BW the two quartiles are 0.50 and 0.79, whereas for CR they are much closer each other, at 0.44 and 0.59. Therefore, the increase of variance is significant.

Figure 4 shows the data used to compute reliability, *i.e.* number of correctly identified barriers and violated checkpoints. These values are computed by considering those cases where both the judge and the subject rated the barrier/checkpoint with severity 0 (not an accessibility problem) and where both rated it with severity greater than 0, not necessarily the same. In other words, we did not consider disagreement of severity when it was greater than 0. Although BW shows a higher number of correctly identified barriers/checkpoints ($M = 13.7$ compared to 7.3), it is readily seen that this number changes a lot evaluation by evaluation. Therefore, according to the definition we gave earlier, reliability of BW is 0.62, which is smaller than reliability of CR, at 0.69. A reliability of 0.62 means that the

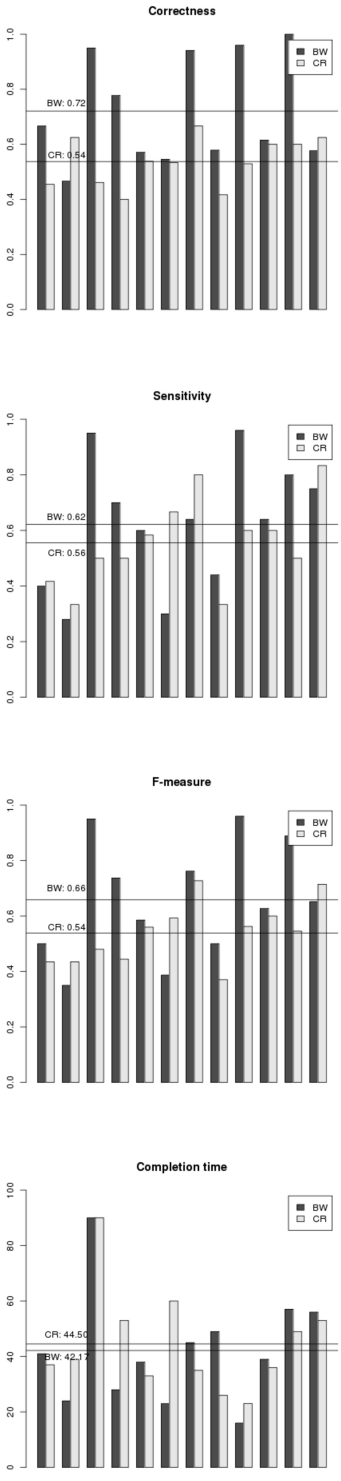

Figure 2: Barplots of correctness, sensitivity, F-measure and completion time.

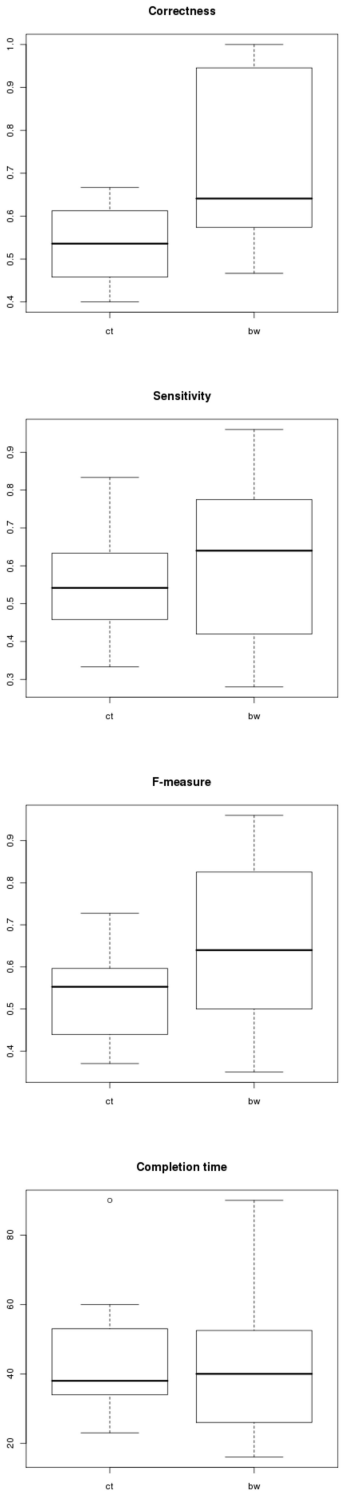

Figure 3: Boxplots of correctness, sensitivity, F-measure and completion time.

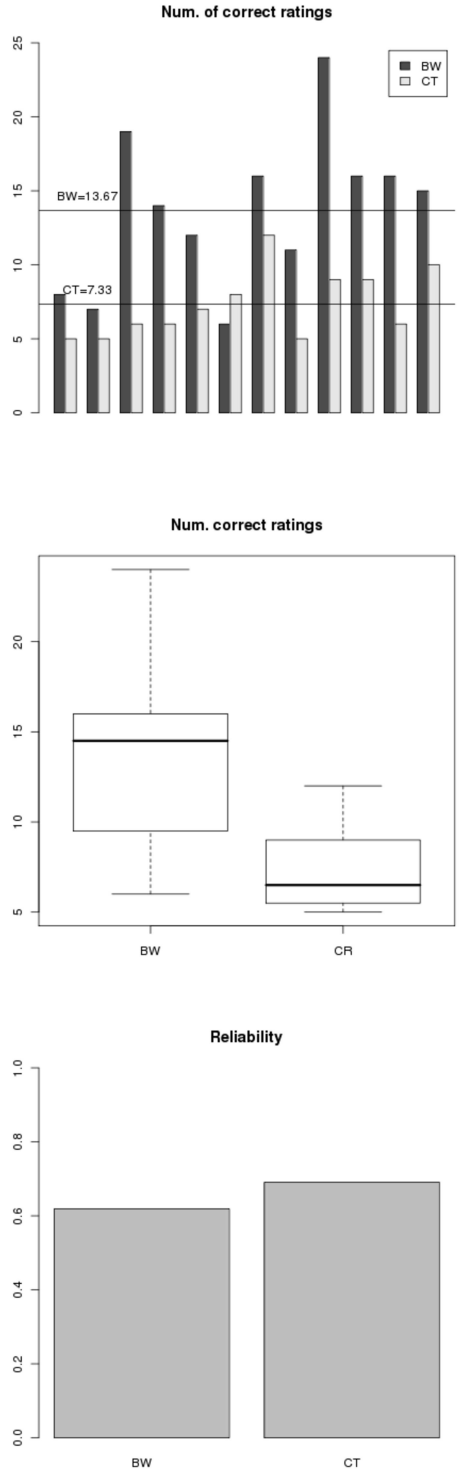

Figure 4: Number of correctly identified barriers/checkpoints (with mean values), and reliability of the two methods, respectively 0.62 for BW and 0.69 for CR.

standard deviation of the number of correct ratings is 0.38 times its mean value.

This is not an unexpected result, again due to the fact that BW requires additional interpretation (*i.e.* subjective) steps to be carried out by the evaluators, namely severity assignment. The positive finding is that the worsening of reliability due to such an additional step is relatively small (reliability has worsened by 10%).

In terms of number of correctly identified violated checkpoints and barriers, analysis of variance (repeated measures ANOVA) shows a significant effect of method used ($df = 1$, $F = 20.56$, $p < 0.0019$) and no other effect nor interactions; see Figure 4. A corresponding planned comparison with a paired t-test highlights a significant difference ($M_{CR} = 7.33, M_{BW} = 13.7, t = 4.65, df = 11, p < 0.001$, two tailed), with a considerable effect size of the difference $d = 0.99$; the 95% confidence interval around the difference in means is $[3.34, 9.32]$, which corresponds to an increase between 45% and 127%. Although such a difference depends on the number of *potential* checkpoints and barriers, it nevertheless implies that when using BW evaluators can identify a larger number of correctly identified problems, and therefore produce more useful reports.

Analysis of variance showed no significant effect (with $\alpha = 0.05$) on correctness, sensitivity and F-measure of task order and of website. The only other significant effects found were of the method being used with respect to correctness ($p = 0.039$) and task order with respect to completion time ($p = 0.011$); see Table 5 and 6.

A paired two-tailed t-test on correctness yields $t = 2.98, df = 11, p < 0.0125$, $M_{CR} = 0.53$, $M_{BW} = 0.72$, an effect size $d = 0.77$ and a confidence interval around the difference of the means of $[0.05, 0.31]$. This corresponds to an increase of correctness between 9% and 60% when using BW. The non-parametric Wilcoxon test yields $V = 69, p = 0.016$, confirming the significance of this result. Sensitivity, as predicted by the ANOVA, showed no significant difference. As a consequence, no difference was found for F-measure as well.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
site:met	1	13.89	13.89	0.63	0.4498
site:order	1	22.86	22.86	1.04	0.3378
met:order	1	20.30	20.30	0.92	0.3649
Residuals	8	175.96	21.99		
site	1	2.67	2.67	0.23	0.6447
met	1	238.93	238.93	20.56	0.0019
order	1	5.38	5.38	0.46	0.5154
site:met:order	1	23.06	23.06	1.98	0.1966
Residuals	8	92.96	11.62		

Table 4: RM ANOVA results of number of correct ratings with respect to website, method and task order; all factors are within-subjects. First 4 rows represent within-subjects variance.

No significant correlation was found between completion time and correctness nor sensitivity: we tried with data transformations (like considering $log(time)$), Pearson linear and Spearman rank correlation. The only significant and moderate correlation that we found was between correctness and sensitivity when using BW: the Pearson's correlation test gave $t = 3.863, df = 10, p = 0.0031, r = 0.77$. When

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
site:met	1	0.00	0.00	0.08	0.7901
site:order	1	0.02	0.02	0.90	0.3694
met:order	1	0.01	0.01	0.47	0.5120
Residuals	8	0.21	0.03		
site	1	0.02	0.02	0.77	0.4046
met	1	0.18	0.18	6.08	0.0390
order	1	0.00	0.00	0.01	0.9123
site:met:order	1	0.00	0.00	0.00	0.9919
Residuals	8	0.24	0.03		

Table 5: RM ANOVA results of correctness with respect to website, method and task order; all factors are within-subjects. First 4 rows represent within-subjects variance.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
met:site	1	1760.30	1760.30	4.28	0.0725
met:order	1	1459.38	1459.38	3.55	0.0965
site:order	1	36.48	36.48	0.09	0.7735
Residuals	8	3293.17	411.65		
met	1	32.67	32.67	0.45	0.5211
site	1	112.93	112.93	1.56	0.2474
order	1	784.03	784.03	10.81	0.0111
met:site:order	1	0.04	0.04	0.00	0.9820
Residuals	8	580.33	72.54		

Table 6: RM ANOVA results of completion time with respect to method, website and task order; factors are within-subjects. First 4 rows represent within-subjects variance.

considering data for CR the same test gave results close to significance and with a smaller correlation: $t = 2.069, df = 10, p = 0.0654, r = 0.55$. Figure 5 shows the scatter plots of correctness against sensitivity.

If such a difference in correlation were confirmed by additional experiments, then the consequence could be that when using BW increases of correctness (which is easier to verify) are accompanied by increases of sensitivity (which is much more difficult to verify).

6. CONCLUSIONS

The paper shows how accessibility evaluation methods can be compared within a framework based on sound and measurable criteria.

Stronger results could have been achieved by ameliorating the experimental plan so that a few disturbance effects would be eliminated:

1. The drop in attention that subjects payed to the second evaluation task, probably due to the longer-than-predicted time required to perform each evaluation.

2. The fact that the tools used by the subjects were particularly well geared towards CR, and no tool was/is available to support BW at the same level of usability.

3. While the detailed description of barriers was in English, checkpoint formulation and explanation was in Italian, the mother tongue of all the subjects.

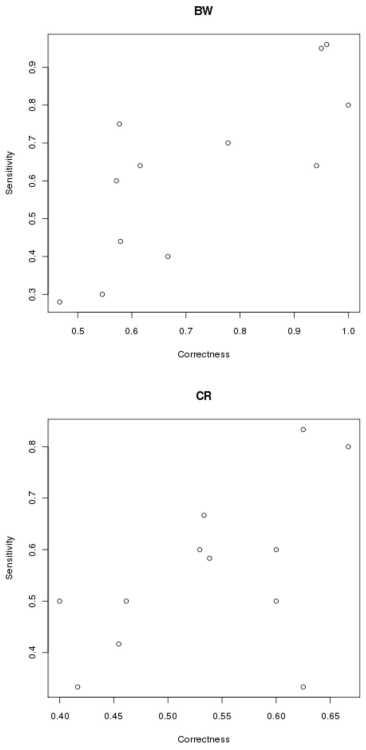

Figure 5: Scatter plot, separately for the two methods, of correctness vs. sensitivity values.

4. A quick preliminary review of barriers and checkpoints could have improved the students ability to correctly identify them.

Regarding the comparison between CR and BW, the main conclusions that can be derived from the experimental data are that, with respect to novice evaluators:

- In terms of reliability, BW scores 10% worse than CR, most probably due to the additional interpretation steps that are required by BW, that contribute to the higher standard deviation in the number of correctly rated problems. However, BW achieved a reliability of 0.62, which is reasonably high.

- Some indexes, namely sensitivity, F-measure, completion time, do not show any generalizable effect. However it is likely that with a larger sample of evaluators significant results could be achieved, since values for BW are systematically higher than CR.

- When using BW, subjects were able to identify a larger number of correctly judged barriers (an increase between 45% and 127%). This difference can be safely generalized beyond the particular pool of subjects we used in the study. Therefore BW leads to more useful results.

- Usage of BW improves correctness by at least 0.05 and as much as 0.31 (corresponding to an increase between 9% and 60% over CR), which can be quite a marked effect.

With more experienced evaluators it is likely that reliability, sensitivity, correctness and F-measure improve when using BW, since more experience would probably lead to reduced variance.

7. REFERENCES

[1] G. Brajnik. Web Accessibility Testing: When the Method is the Culprit. In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *ICCHP 2006, 10th International Conference on Computers Helping People with Special Needs*, Lecture Notes in Computer Science 4061, Linz, Austria, July 2006. Springer Verlag.

[2] G. Brajnik. Web accessibility testing with barriers walkthrough. `www.dimi.uniud.it/giorgio/projects/bw`, March 2006. Visited May 2008.

[3] G. Brajnik. Beyond conformance: the role of accessibility evaluation methods. In *2nd International Workshop on Web Usability and Accessibility IWWUA08*, Auckland, New Zealand, Sept. 2008. Keynote speech.

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

[5] K. Coyne and J. Nielsen. How to conduct usability evaluations for accessibility: methodology guidelines for testing websites and intranets with users who use assistive technology. `http://www.nngroup.com/reports/accessibility/testing`, Nielsen Norman Group, Oct. 2001.

[6] A. Dey. Accessibility evaluation practices - survey results. `http://deyalexander.com/publications/accessibility-evaluation-practices.html`, 2004. Visited May 2008.

[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] H. R. Hartson, T. S. Andre, and R. C. Williges. Criteria for evaluating usability evaluation methods. *Int. Journal of Human-Computer Interaction*, 15(1):145–181, 2003.

[10] S. Henry and M. Grossnickle. *Just Ask: Accessibility in the User-Centered Design Process*. Georgia Tech Research Corporation, Atlanta, Georgia, USA, 2004. On-line book: `www.UIAccess.com/AccessUCD`.

[11] M. Hertzum and N. Jacobsen. The evaluator effect: a chilling fact about usability evaluation methods. *Int. Journal of Human-Computer Interaction*, 1(4):421–443, 2001.

[12] M. Hertzum, N. Jacobsen, and R. Molich. Usability inspections by groups of specialists: Perceived agreement in spite of disparate observations. In *CHI 2002 Extended Abstracts*, pages 662–663. ACM, ACM Press, 2002.

[13] Italian Government. Requisiti tecnici e i diversi livelli per l'accessibilità agli strumenti informatici. `www.pubbliaccesso.it/normative/DM080705.htm`, July 2005. G. U. n. 183 8/8/2005.

[14] B. Kelly, D. Sloan, S. Brown, J. Seale, H. Petrie, P. Lauke, and S. Ball. Accessibility 2.0: people, policies and processes. In *W4A '07: Proc. of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pages 138–147, New York, NY, USA, 2007. ACM.

[15] 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.

[16] 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 2005: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 41–50, New York, NY, USA, 2005. ACM.

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

[18] 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.

[19] A. Sears. Heuristic walkthroughs: finding the problems without the noise. *Int. Journal of Human-Computer Interaction*, 9(3):213–234, 1997.

[20] W3C/WAI. How people with disabilities use the web. World Wide Web Consortium — Web Accessibility Initiative, `w3.org/WAI/EO/Drafts/PWD-Use-Web/20040302.html`, March 2004. Visited May 2008.

[21] W3C/WAI. Conformance evaluation of web sites for accessibility. `www.w3.org/WAI/eval/conformance.html`, 2008. Visited May 2008.

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