Testability and Validity of WCAG 2.0: The Expertise Effect

Giorgio Brajnik
Dept. of Computer Science
University of Udine
Udine, Italy
brajnik@uniud.it

Yeliz Yesilada
Middle East Technical
University, Northern Cyprus
Campus, Mersin 10, Turkey
yyeliz@metu.edu.tr

Simon Harper
School of Computer Science
University of Manchester
Manchester, UK
sharper@cs.man.ac.uk

ABSTRACT

Web Content Accessibility Guidelines 2.0 (WCAG 2.0) require that success criteria be tested by human inspection. Further, testability of WCAG 2.0 criteria is achieved if 80% of knowledgeable inspectors agree that the criteria has been met or not. In this paper we investigate the very core WCAG 2.0, being their ability to determine web content accessibility conformance. We conducted an empirical study to ascertain the testability of WCAG 2.0 success criteria when experts and non-experts evaluated four relatively complex web pages; and the differences between the two. Further, we discuss the validity of the evaluations generated by these inspectors and look at the differences in validity due to expertise.

In summary, our study, comprising 22 experts and 27 non-experts, shows that approximately 50% of success criteria fail to meet the 80% agreement threshold; experts produce 20% false positives and miss 32% of the true problems. We also compared the performance of experts against that of non-experts and found that agreement for the non-experts dropped by 6%, false positives reach 42% and false negatives 49%. This suggests that in many cases WCAG 2.0 conformance cannot be tested by human inspection to a level where it is believed that at least 80% of knowledgeable human evaluators would agree on the conclusion. Why experts fail to meet the 80% threshold and what can be done to help achieve this level are the subjects of further investigation.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*; 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, expertise

ASSETS'10, October 25–27, 2010, Orlando, Florida, USA.

1. INTRODUCTION

Accessibility (the property of a website such that "people with some impairment can use it with the same effectiveness as non-disabled people" [23]) deals not only with technicalities of a user interface, but also with the way people perceive, interpret and act on the user interface. Whenever people are involved, subjectivity is introduced. Indeed, many factors influence the user experience such as: technical aspects, like the specific user agent, the assistive technology, the accessibility architecture of the user platform, the coding of the user interface; personal aspects, such as the kind and degree of impairment, the experience in using the web, in using the devices and software, in using the specific website, the motivation to do something; and finally contextual ones, like the actual posture, location, lighting, noise, simultaneous interaction with other people. All these factors affect the accessibility of a website and unless one is very careful in isolating them, the results of an evaluation will be difficult to repeat.

Rather than focusing on accessibility, in conformance reviews the compliance of a website to success criteria/checkpoints specified by accessibility guidelines is what many evaluators try to determine. There are several reasons why one could assume that evaluating conformance is mostly deterministic: (i) the fact that guidelines and success criteria are stated in a very carefully chosen language; (ii) the fact that (as is for WCAG 2.0 [5]) there are sufficient techniques on one side and known failures on the other side, that provide separate sufficient and necessary conditions for satisfaction of individual success criteria; (iii) the fact that these techniques and counter-examples refer to specific code fragments in HTML/CSS or other specific technologies; (iv) the fact that in the formulation of success criteria, emphasis is given to their testability, either through automatic tools or by competent evaluators.

However, when dealing with analytic evaluation methods based on the opinion of evaluators that try to estimate how people would perceive, interpret and act on a user interface, the subjective influences mentioned above are introduced, adding noise to the decisions of the evaluators. For instance, the following email was sent to the W3C-WAI (Web Accessibility Initiative) interest group mailing list:

> We have come across a scenario lately, where 2 different accessibility audits have produced different results. As a company we are legally obliged to provide AA compliant web applications, however this is very subjective. How, as a company, do we protect ourselves legally?

To determine if this is the case for WCAG 2.0, two experiments were carried out in the past by two independent research teams, both involving novice accessibility evaluators. The first such study

[26] suggested that for the majority of the success criteria, the number of evaluators that would agree on the outcome is well below 80%, a threshold used to discriminate between testable and non testable techniques. The second study [2], although yielding higher values for agreement, concluded likewise that testability of WCAG 2.0 is not to be taken for granted.

In this paper we aim to complete one or more piece of the puzzle by showing what happens when the evaluators are accessibility experts. We performed an experiment with 22 experts and 27 non-experts, asking them to rate all 61 WCAG 2.0 success criteria on four webpages. Results show that for experts half of the success criteria fail to meet the 80% agreement threshold; experts produce also 20% false positives and miss 32% of the true problems. We also compared performance of experts against that of novice evaluators; we found that agreement for non-experts drops by 6%, false positives reach 42% and false negatives 49%.

2. WEB ACCESSIBILITY EVALUATION

There are a number of different evaluation methods that can be used to assess the accessibility of web pages [4]. These methods have different pros and cons, and different properties. For example, they differ in how they are carried out (*e.g.*, analytical or observational), in who is leading the generation of results (*e.g.*, evaluator or user), in the purpose of the study (*e.g.*, to identify consequences of the problems or to identify the problems), in what they generate as outputs (*e.g.*, qualitative or quantitative) and finally in terms of the cost. We can summarize these methods into five categories:

1. Inspection Methods, based on an evaluator inspecting a web page for its accessibility. The most widely used inspection method is *Conformance Review*, where the evaluator uses a set of accessibility guidelines that focus on possible accessibility problems and has to decide if a page or web site complies to those requirements [1, 24, 10, 8]. Here, our focus is on WCAG 2.0 conformance reviews.

2. Automated Testing, which involves an evaluator using an automated accessibility tool to check conformance of a web page against the accessibility principles encoded in that tool. There are many tools available[1], yielding different results with different levels of quality [1, 24, 3].

3. Screening Techniques, based on using a website in a way that some sensory, motor or cognitive capabilities of the user are artificially reduced[2] [10], and in such a way to simulate some of the conditions that are typical for people with disabilities.

4. Subjective Assessment, which is a process where an evaluator hires a panel of users who are asked to explore/use a website autonomously and send back their opinions; the evaluator then collects such feedback to determine accessibility of pages [10].

5. User Testing, which is a process where formal or informal experiments are set up with real users, who are individually asked to perform goal-free or goal-oriented navigation on a web site, and whose behavior is observed by evaluators [16, 22, 10, 7, 20].

2.1 The Evaluator Effect

Hornbaek et al. [14] define the evaluator effect as "the observation that usability evaluators in similar conditions identify substantially different sets of usability problems". This has been studied across a variety of usability evaluation techniques such as cognitive walkthrough, heuristic evaluation and think-aloud studies [13, 12, 11, 15]. However, for accessibility this phenomenon has not been extensively studied. One of the few studies touching on this issue was performed by [21], which showed that while participants and evaluators 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.

As mentioned in the Introduction, two recent experimental studies focused indirectly on this effect for conformance reviews based on WCAG 2.0 [26, 2] by looking at how testable those success criteria are. Testability is related to the evaluator effect; in fact the W3C-WAI defines *Reliably Human Testable* as [25]:

> The technique can be tested by human inspection and it is believed that at least 80% of knowledgeable human evaluators would agree on the conclusion. The use of probabilistic machine algorithms may facilitate the human testing process but this does not make it machine testable.

In Lang's [17] review of accessibility evaluation methods, the skill requirements of conformance review is discussed, and the conclusion was that "[These methods] require evaluators to have a greater skill level to review, understand guidelines and recommend solutions". Thus, an additional factor affecting the evaluator effect is expertise.

2.2 The Expertise Effect

Subjectivity may be due, in part, to different expertise levels of different evaluators. We believe it is important to characterize the effect that expertise plays in accessibility evaluations for a number of reasons.

The widespread use of analytical methods when performing accessibility evaluations exacerbates the evaluator effect, since the output of these methods are opinions of evaluators. The mere existence of guidelines may induce inexperienced evaluators into wrong results. Combined with the existence of testing tools, the possibility to perform accessibility evaluations is given to anybody, and it is expected that most developers assess the accessibility level of their creations. Hence, accessibility evaluations are more and more often performed by people whose experience in accessibility is limited, leading to more variability of results. Because of their complexity, WCAG 2.0 require high levels of expertise in order to be properly used. This is on top of what accessibility evaluation requires: a good understanding of how disabled people access the web, what kind of assistive technologies [9] they use, how these assistive technologies work, what the limitations of these technologies are, how they inter-operate with other technologies.

We can characterize evaluator expertise in terms of:

1. the practice in using a specific method, which could also involve knowledge of a set of guidelines;
2. knowledge, practice, skill in accessibility in general (which can be characterized as experience on assistive technologies, typical accessibility problems, typical user behaviors or user preferences) and in the underlying web technologies;
3. the experience in evaluating websites for accessibility.

3. EMPIRICAL STUDY

An ideal web accessibility evaluation method should always yield accurate predictions of all and only the *real* accessibility problems that may occur in a web page when used by *real* users. In this paper, we focus on investigating the effectiveness of WCAG 2.0 by measuring validity and reliability of the results obtained through conformance reviews. For measuring validity, which is the extent to which all and only the true problems are found, we focus on accuracy, correctness, sensitivity and F-measure. For measuring

[1] http://www.w3.org/WAI/eval/selectingtools
[2] http://www.w3.org/WAI/eval/preliminary.html

reliability, which is related to the extent to which independent evaluations (for example, because performed by different evaluators, or at different times, or in different situations) produce the same results, we focus on max-agreement; testability of a success criterion can be defined then as whether max-agreement reaches 80%.

3.1 Research Questions

We investigate the following research questions:

1. *How testable are the WCAG 2.0 success criteria when expert evaluators are involved?* The previous preliminary studies [26, 2] with non-expert evaluators found that the levels of testability for WCAG 2.0 is very low. However, those studies had limitations because the evaluators that participated in the study had no experience with WCAG 2.0.
2. *What is the difference in testability when involving non-expert evaluators?* This is important because the audience of the WCAG 2.0 is not only people who by profession or interest have become experts, but also the multitude of developers and quality assurance people that now and then perform accessibility assessments.
3. *How valid are the WCAG 2.0 evaluations generated by expert evaluators?* Specifically, what is the accuracy, correctness, sensitivity and f-measure that can be achieved by expert evaluators?
4. *What is the difference in validity of WCAG 2.0 due to differences in expertise?* If validity of the results is affected by the expertise then this means that to obtain good results one needs to use trained evaluators. This will also show the importance of training in web accessibility evaluation with WCAG 2.0 conformance testing.

3.2 Materials

We used the following pages: (i) "I love God Father movie" Facebook group; (ii) "The Godfather at IMDB"; (iii) "Bloomberg .com: WorldWide"; (iv) "Biotechnology News, Articles, and Information from Scientific American". They were chosen because they differ in layout, complexity and also in terms of accessibility support. Before we did this study, we used an automated tool and confirmed that these pages had significant violations which was important for our data collection. Furthermore, two of these pages (IMDB and Facebook group) were used in our previous study [26], therefore for comparison purposes we decided to include those two in this study. Each judge evaluated one page and the pages assigned to judges were randomized. Each judge was given a sheet with a randomized list of WCAG 2.0 guidelines and success criteria to counterbalance order effects.

3.3 Procedure

Expert participants were invited to take part in this study via email, and to those who completed the study a voucher of an online store was given. Non-expert participants were students of one of the authors, and were told that their performance (in terms of validity) would be considered during the exam. When participants accepted our invitation, they were allocated a judge's number and asked to follow the instructions on the experiment web page[3]. This web page first provided a brief summary of this study and provided some information about our related previous work. Participants completed the study in their own time and working environment; students were given 2 weeks time.

In brief, the instructions included six steps. In *step 1*, participants were asked to read an information sheet which was also presented as a web page[4]. This page detailed the purpose of the study, and provided answers to questions such as "can I take part in this study?", "will my data be anonymous?", etc. In *step 2*, participants then were invited to fill in a screening questionnaire, that included questions about age, gender, expertise in accessibility and WCAG 2.0, etc. In *step 3*, participants were requested to download the corresponding worksheet and to get the corresponding web page for the given judge number. This was to ensure that each judge would get a randomized success criteria sheet, and a specific web page, since pages were also assigned randomly. In *step 4*, participants were instructed on how to use this worksheet which included the list of WCAG 2.0 guidelines and success criteria (SC). Participants were asked to rate each SC, indicate the difficulty of rating based on 5 point Likert scale and, based on their own experience, how often they observed this SC failing on other pages they evaluate. For rating each SC they were asked to use "not-applicable" if the SC did not apply to the page, "pass" or "fail" depending on whether the given page failed to conform to the SC. Participants were allowed to use any evaluation tool, browser extension and technique they liked. The instruction page contained a link to the WCAG 2.0 official document; students were told about the WCAG "Understanding ..." and "Techniques ..." documents during the lectures preceding the experiment. In *step 5*, after participants completed their evaluation, they were asked to fill in a post-evaluation questionnaire which captured how long it took to complete the study, the tools and techniques used, familiarity with the page and the participants' subjective rating of the level of effort, productivity required, and their confidence both in the evaluation and WCAG 2.0. Finally, participants were asked to email the worksheet, screening and post-evaluation data to us for data collection.

3.4 Participants

In our study there were 23 expert and 27 non-expert participants. Non-experts were students (aged between 21 and 29, $M = 23$, $\mathrm{sd} = 1.9$; 4 females) who were attending a 3^{rd} year university course on web usability and accessibility, and at that time (Dec. 2009) had attended about 14 hours of lectures on web accessibility (assistive technology, typical barriers, the Italian technical accessibility regulations, and WCAG 2.0). They all spoke Italian as their mother tongue.

Even though 23 experts participated in our study, we had to exclude the data from one expert due to incorrect usage of the spreadsheet. The remaining 22 were aged between 28 and 60 ($M = 40, \mathrm{sd} = 9.7$), 6 were females, 13 spoke English as mother tongue, 3 Italian, 2 Japanese, 1 Portuguese and 1 Bulgarian. None of the authors participated as evaluators. The experts were invited because either they had publications on web accessibility or they are currently working as professional consultants. In fact, we contacted more than 23 experts, but some were not able to participate due to time constraints, some did not respond to our invitation, some did not complete the study on time, some did not manage to use the spreadsheets, and finally some claimed that some of the WCAG 2.0 checkpoints are too subjective and they refused to take part in this study (we further discuss this in Section 5). We believe that the reason why these people did not participate is not related to experimental conditions; in other words, we do not think these refusals introduced bias in the results.

4. RESULTS

In this section we investigate research questions detailed in Sec-

[3] http://hcw.cs.manchester.ac.uk/research/riam/experiments/wcag/exp.php

[4] http://hcw.cs.manchester.ac.uk/research/riam/experiments/wcag/wcag-exp-info.php

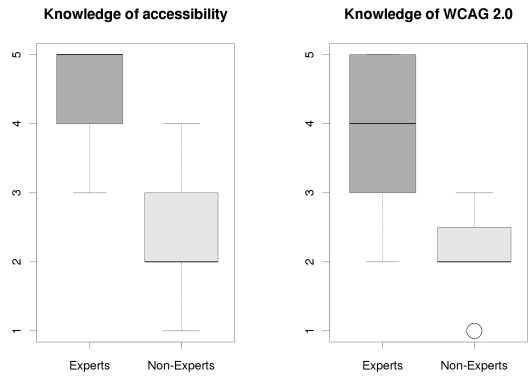

Figure 1: Boxplot of self-rated knowledge of accessibility (left) and of WCAG 2.0 (right). The thick horizontal line depicts the median.

tion 3.1 and present data collected about our participants. Regarding the ratings, overall we collected 3050 individual ratings, 1403 by experts and 1647 by non-experts; 138 of them were invalid and so had to be excluded; 976 concerned the Bloomberg page, 793 Facebook, 610 IMDB and 671 Scientific American.

4.1 Screening and Post Evaluation Data

As part of the screening data, we collected demographic information about our participants. One is the self-rated knowledge of accessibility (KA), using a 5-points Likert question (1:lowest; 5:highest): Table 1 and Figure 1 provide the distribution of the data. We can see that there is a difference between experts and non-experts; in fact a non parametric Wilcoxon rank sum test shows a significant difference in KA due to expertise ($W = 585, p < 0.0001$).

	Know. of accessibility		**Know. of WCAG 2.0**	
	expert	non-expert	expert	non-expert
min	3.00	1.00	2.00	1.00
max	5.00	4.00	5.00	3.00
median	5.00	2.00	4.00	2.00
mean	4.59	2.30	3.86	2.04
sd	0.59	0.72	0.94	0.71

Table 1: Knowledge of accessibility and WCAG 2.0.

Another data we collected was a 5-points Likert question on knowledge in WCAG 2.0 (KW); results are shown in Table 1 and Figure 1. Also in this case there is a significant difference due to expertise ($W = 548.5, p < 0.0001$). There is a significant and quite strong correlation between KA and KW (Pearson's $\rho = 0.84, t = 10.4(47), p < 0.0001$) indicating that the two types of knowledge are part of the same expertise.

We also asked each participant how many sites did he or she evaluate for accessibility in the previous six months; experts evaluated a median of 3.5 sites (ranging from 0 to 100+; one expert had not evaluated any site); only 4 non-experts had evaluated 1 site; the other ones none. There is a significant difference in the number due to expertise ($W = 567.5, p < 0.0001$). KA is significantly but weakly correlated with the number of tested websites ($\rho = 0.37, t = 2.75(47), p < 0.0001$).

Regarding the time it took to complete each evaluation (in minutes): for experts $M = 115, \mathrm{sd} = 72, \mathrm{range} = [40, 360]$; for non-experts $M = 370, \mathrm{sd} = 169, \mathrm{range} = [180, 840]$. After transforming time using logarithms (the distribution of time is positively skewed with a peak at 300 min), a t-test shows a significant difference due to expertise ($t = 8.83(36.5), p < 0.0001$); the 95% confidence interval of the *ratio* between the time for experts and that for non-experts is $[0.22, 0.38]$ (*i.e.*, experts took between .22 and .38 times the length of time needed by non-experts).

Regarding the self-rated effort required to perform the evaluation (5-points Likert question, 1:lowest, 5:highest), 43% of experts and 59% of non-experts rated it as 4 or 5 (we found no significant association between effort and expertise). Similarly for perceived productivity (33% of experts and 37% of non-experts rated it as 4 or 5). Regarding the confidence in applying the WCAG 2.0, there is a significant but weak association between those who rated it 4 or 5 with expertise: 57% of experts and 22% of non-experts rated it 4 or 5 ($\chi^2(1) = 4.75, p <= 0.0294, \mathrm{Cramer's}\ \phi = 0.31$).

All this data suggests that our a-priori classification of participants according to expertise is consistent with the data: experts rated themselves as more knowledgeable in accessibility, in WCAG 2.0, evaluated more sites in the previous 6 months, took about 31% less time to evaluate pages, were more confident in applying the WCAG 2.0. We believe therefore that our classification of participants as experts or non-experts is valid, and that using self-rated knowledge in accessibility and in WCAG, the number of evaluated sites, whether one works as a consultant as classification criteria is sound.

4.2 Testability

Reliability is related to repeatability of results, being therefore directly associated to *testability*. For us *reliability* is the extent to which independent evaluations (for example, because performed by different evaluators, or at different times, or in different situations) produce the same results. We measure reliability in terms of *maximum agreement* (MA), given the set of ratings provided by different judges on a success criterion with respect to a page. Max-agreement is defined as the relative frequency of the mode, *i.e.* the percentage of occurrence of the most frequent value of the set of ratings, which in our case is $\{\texttt{fail}, \texttt{pass}, \texttt{na}\}$[5]. After grouping all the data by page and by success criterion we obtained a set of $61 \cdot 4 = 244$ data-points; the overall mean of MA is 0.70 ($\mathrm{sd} = 0.17$); for normalized MA $M = 0.55, \mathrm{sd} = 0.25$. As illustrated by Figure 2 (left), there are 9 success criteria with a mean MA greater or equal to 80%; just one is consistently greater or equal to 80% all four pages; twelve success criteria never reach that threshold.

When splitting ratings given by experts from those of non-experts, we discovered that while non-experts' ratings were almost equally distributed across pages (between 6 and 8 non-experts evaluated each page), due to participants withdrawing from the study we had 10 experts evaluating Bloomberg, 6 Facebook, 4 IMDB and 3 Scientific American. Since 3 and 4 judges are too few to provide a reasonable basis, we decided to exclude all the ratings that judges gave on the latter two pages.

On the remaining subset of data (given by experts on the former two pages), we computed the max-agreement as above. Table 2

[5]Because the minimum value of max-agreement is determined by the resolution scale of the ratings (for example, in our case the minimum value for max-agreement is 0.33), for the purpose of comparing max-agreement reported in other studies we also compute a linear adjustment to normalize it within $[0, 1]$, so that 0 corresponds to the minimum value and 1 to 1; we call this *normalized max-agreement*.

Success criterion	Mean	Min	Max
3.3.6	0.40	0.40	0.40
2.4.5	0.45	0.40	0.50
3.1.5	0.50	0.40	0.60
2.4.8	0.55	0.40	0.70
1.3.2	0.60	0.60	0.60
2.4.7	0.60	0.60	0.60
3.3.3	0.60	0.60	0.60
3.3.5	0.60	0.60	0.60
1.4.1	0.65	0.60	0.70
2.4.6	0.65	0.60	0.70
1.4.9	0.65	0.40	***0.90**
2.2.1	0.65	0.50	***0.80**
3.3.2	0.65	0.50	***0.80**
3.3.1	0.68	0.60	0.75
1.4.3	0.68	0.56	***0.80**
2.1.3	0.68	0.56	***0.80**
2.4.4	0.70	0.60	***0.80**
3.1.4	0.70	0.50	***0.90**
3.2.2	0.70	0.60	***0.80**
1.4.6	0.71	0.62	***0.80**
1.4.8	0.71	0.62	***0.80**
1.4.7	0.74	0.60	***0.88**
4.1.2	0.74	0.60	***0.89**
1.3.1	0.75	0.70	***0.80**
2.4.1	0.75	0.60	***0.90**
2.4.3	0.75	0.60	***0.90**
3.1.6	0.75	0.70	***0.80**
3.2.3	0.75	0.60	***0.90**
3.2.4	0.75	0.60	***0.90**
2.2.2	0.78	0.56	***1.00**
1.3.3	***0.80**	0.60	***1.00**

Success criterion	Mean	Min	Max
1.4.5	***0.80**	0.60	***1.00**
2.1.1	***0.80**	***0.80**	***0.80**
2.2.3	***0.80**	0.60	***1.00**
3.1.1	***0.80**	***0.80**	***0.80**
3.1.2	***0.80**	0.60	***1.00**
3.2.5	***0.83**	0.67	***1.00**
1.2.2	***0.85**	0.70	***1.00**
1.2.6	***0.85**	0.70	***1.00**
2.1.2	***0.85**	0.70	***1.00**
2.2.4	***0.85**	0.70	***1.00**
4.1.1	***0.85**	0.70	***1.00**
2.3.2	***0.85**	***0.80**	***0.90**
2.4.10	***0.85**	***0.80**	***0.90**
3.1.3	***0.85**	***0.80**	***0.90**
2.2.5	***0.89**	0.78	***1.00**
2.4.9	***0.89**	0.78	***1.00**
1.2.1	***0.90**	***0.80**	***1.00**
1.2.3	***0.90**	***0.80**	***1.00**
1.2.4	***0.90**	***0.80**	***1.00**
1.2.5	***0.90**	***0.80**	***1.00**
1.2.7	***0.90**	***0.80**	***1.00**
1.2.8	***0.90**	***0.80**	***1.00**
1.2.9	***0.90**	***0.80**	***1.00**
1.4.4	***0.90**	***0.80**	***1.00**
2.3.1	***0.95**	***0.90**	***1.00**
3.2.1	***0.95**	***0.90**	***1.00**
3.3.4	***0.95**	***0.90**	***1.00**
1.1.1	***1.00**	***1.00**	***1.00**
1.4.2	***1.00**	***1.00**	***1.00**
2.4.2	***1.00**	***1.00**	***1.00**

Table 2: Means and range of MA over all success criteria by experts (sorted by increasing means and * is $\geq 80\%$).

shows the mean and range (over the two pages) of MA for each of the 61 success criteria. There are 31 success criteria with a mean MA that is greater or equal to 80%; 19 are consistently greater or equal to 80% over both pages; and 11 success criteria never reach that threshold. We can also notice that several success criteria have MA values that are relatively low; the first 13 ones show a mean MA that is less than $2/3$, meaning that about one out of three experts disagreed. Figure 2 (center) shows that data graphically.

Similarly, MA can be computed on data produced by non-experts (Figure 2 right). Fewer success criteria (12) have a mean MA that is at or above the threshold; just 2 success criteria are consistently above the threshold; 9 never make it. It is noticeable also the wider range of MA (across pages) in the case of non-experts.

Our data can be aggregated by page, and when only the two pages evaluated both by experts and non-experts are considered, we find a significant difference in MA between experts and non-experts: a paired two-tailed t-test on the mean MA (over all SCs) gives $T = 3.28(60), p < 0.00173$, with a 95% confidence interval of the difference in MA equal to $[0.024, 0.098]$. Similarly for a paired two-tailed t-test on the minimum MA (over all SCs): $T = 6.12(60), p < 0.0001$, confidence interval $[0.086, 0.170]$.

MA for experts is $M = 0.77(\mathrm{sd} = 0.18)$, whereas for non-experts it is $M = 0.71(\mathrm{sd} = 0.18)$. Normalized MA for experts is $M = 0.66(\mathrm{sd} = 0.27)$, whereas for non-experts it is $M = 0.57(\mathrm{sd} = 0.27)$.

4.3 Validity

We investigate validity in terms of correctness, sensitivity and f-measure, all based on correctness of judges' responses. We defined *correct rating* as those ratings where the majority of experts agreed; in other terms the values (taken from $\{\texttt{fail}, \texttt{pass}, \texttt{na}\}$) that, for each combination of page and success criterion, constitute the mode. In case of ties we considered as correct all the modes. Pooling the responses of participants as a way to identify correct answers was used by other researchers investigating validity of usability evaluation methods (*e.g.*, [19]); by adopting a majority rule to characterize correct answers, we were also able to investigate how many false positives were returned. In our case, an alternative method to identify true problems could have been running a user testing experiment and considering as true problems only those that would be elicited from such a process; however, the problem of ensuring that most of the success criteria should be covered would have made such a procedure very complex and costly, while not ruling subjectivity out because of the evaluator effect. We believe our method for characterizing correct answers, relying on what the majority of experts said, is sound.

After restricting to a given page, we define the *true violations* (TV) for that page the set of success criteria that are correctly rated as "fail"; the set of *found violations* (FV), given a judge and a page, is the set of ratings equal to "fail". These sets can be used to define three indexes:

Correctness $C = \frac{|TV \cap FV|}{|FV|}$ is the proportion of found success criteria violations that are also correct.

Sensitivity $S = \frac{|TV \cap FV|}{|TV|}$ is the proportion of all the true success criteria violations that were found.

F-measure $F = \frac{2C \cdot S}{C+S}$ is the harmonic mean of C and S, a balanced combination of C and S summarizing validity of an evaluation. Neither correctness nor sensitivity alone can characterize validity, they have to be considered jointly: f-measure is a convenient way to provide an overall index of validity. Notice that a given change of x% in f-measure is equivalent to an x% change of correctness *and* a simultaneous x% change of sensitivity.

Table 3 shows the means and standard deviations of correctness, sensitivity and f-measure, including the 95% confidence interval of the difference due to expertise. For each variable, the difference is significant when tested with a 2-sample two-tailed t-test: for correctness $T = 6.33(46.96), p < 0.0001$, for sensitivity $T =$

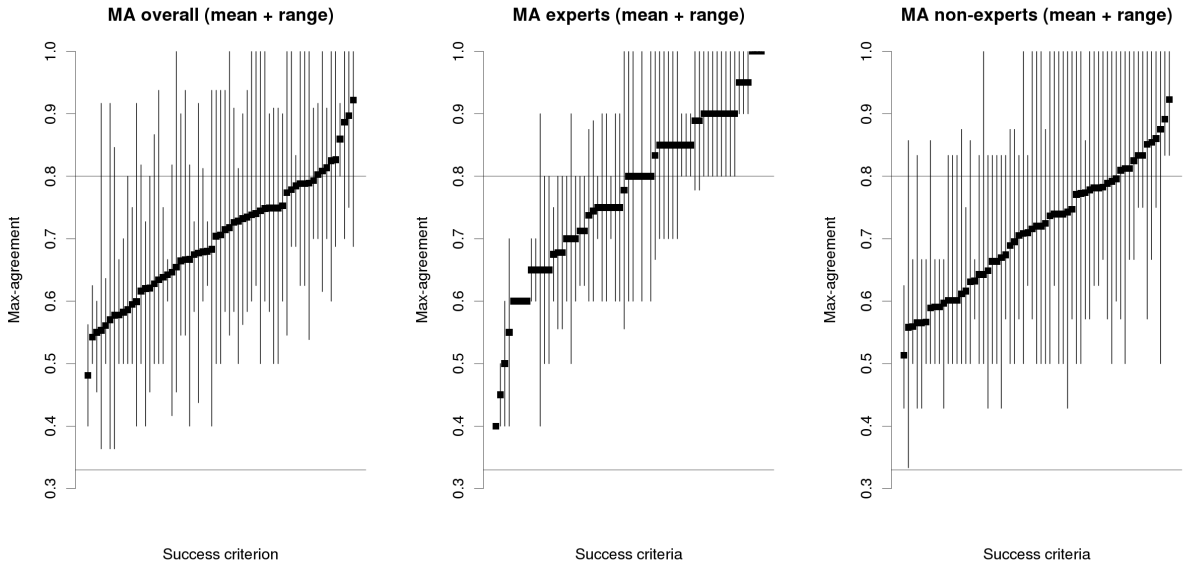

Figure 2: Mean and range of MA by success criterion, overall (left), on experts' (center) and non-experts' ratings (right); horizontal lines show the 80% threshold and the 33% theoretical minimum value.

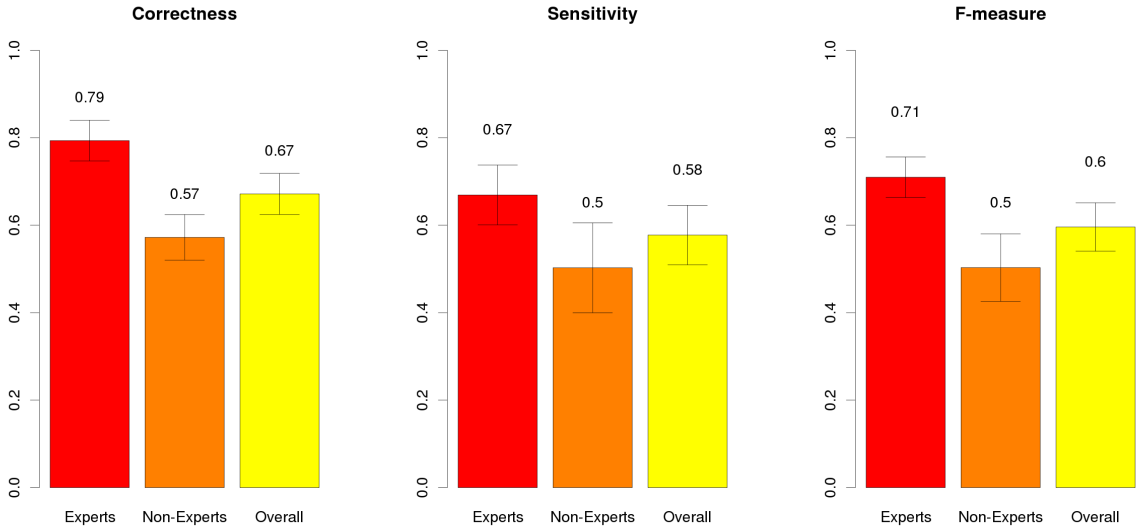

Figure 3: Correctness of judges.

$2.82(42.56), p < 0.0074$, for f-measure $T = 4.65(39.78), p < 0.0001$.

5. DISCUSSION

As can be seen from our results, our a-priori classification of participants according to expertise level is valid when compared to the data acquired through the screening and post-evaluation questionnaires. In particular, we believe our expert participants fit the "knowledgeable human evaluators" category required by the "Testability" definition.

Although required not to do it, students might have collaborated. However, the low reliability shown by the data suggests this is not

	Correctness	Sensitivity	F-measure
Overall	0.68 (0.16)	0.59 (0.24)	0.61 (0.20)
Experts	0.80 (0.11)	0.68 (0.16)	0.72 (0.10)
Non-Experts	0.58 (0.14)	0.51 (0.27)	0.51 (0.21)
Difference	$[0.15, 0.29]$	$[0.05, 0.30]$	$[0.12, 0.31]$

Table 3: Means (and standard deviations) for the three validity measures, and the 95% confidence intervals of the difference due to expertise.

the case. And if they did, the reliability we found would be overestimated (when the number of independent raters increases, reliability usually decreases). Thus collaboration among participants is not a concern. Students handed in the results toward the end of the allotted time period; we believe however that a two weeks time was more than enough to refresh one's knowledge of WCAG and to evaluate a page.

Our data from $10 + 6$ experts on two pages show that on average 31 of 61 success criteria reach the 80% threshold on max-agreement required by testability, and that only 19 of them (31%) reach it on both pages. Thus, many success criteria are not testable (at least on the 2 pages considered by experts), even though the mean max-agreement for expert is 77% (*i.e.*, on average 77% of judges agree on the rating of success criteria), fairly close to the 80%. However, in our previous studies [26], we investigated max-agreement of experts using the Barrier Walkthrough (BW) method to identify accessibility barriers. In that study we found a (normalized) MA equal to 81%, compared to 66% normalized MA found in this study for WCAG 2.0.

It is difficult to identify why this rate is so low. A brief examination of the success criteria at the extremities of Table 2 shows that even level A success criteria such as "1.4.1 Use of color to convey information" or "2.2.1 Timing adjustable" are not testable. On the other side of the table we find many testable success criteria that are based on features that are easy to identify within pages, such as "3.1.1 Language of page" or "2.1.1 Keyboard operability" or "1.2.6 Prerecorded significant language". Consider also that the interpretation of "testable" as defined in [25] might refer to the set of ratings $\{\texttt{fail}, \texttt{pass}\}$, excluding therefore the value `notapplicable`. In such a case our data would be higher, and a larger fraction of success criteria would be deemed testable.

In this study, experts produced, on average, 20% of false positives (*i.e.*, identified the wrong problems), missed 32% of the correct problems, with a combined validity (f-measure) of 72%. However, we have to consider that in our experiment, "correct rating" was defined in terms of what the majority of experts said. Another characterization of correct answers, for example based on results obtained from a large scale user testing procedure (involving many users, belonging to many disability groups, showing different levels of expertise in using the web and in using assistive technologies, having users use many different platforms) could have provided a much sounder definition of what were the correct violations. Notice however that the cost would have been prohibitive and that, in any case, the evaluator effect would have been introduced [14] thereby reducing the goodness of such data.

In our previous study with the BW method [26], the correctness of experts was 62% compared to 80% we found now with WCAG 2.0; sensitivity (*i.e.*, false negatives) in the previous study was 71% *vs* 68%; finally, f-measure was 63% *vs* 72%.

In this study, we also compared evaluation results between experts and non-experts. Compared to experts, non-experts show a relatively small decrease of max-agreement (between 2.4 and 10%), with a mean of 71%. According to non-experts' data, only 2 success criteria are consistently over the 80% threshold (on all four pages). This decrease in max-agreement is consistent with what we found in our previous study with the BW method [26].

In [26] a similar investigation on testability was carried out. In that study students not exposed to lectures on WCAG 2.0 evaluated pages by using WCAG 1.0 and WCAG 2.0 success criteria: max-agreement was found to be 68% for WCAG 1.0 and 61% for WCAG 2.0. In the other more recent study on testability of WCAG 2.0, max agreement for a subset of success criteria appears to be 77% (we manually computed the means of the values that can be gleaned from Figure 4 in [2]).

In this study, non-experts produced, on average, 42% false positives (between 15 and 29% more than experts), missed 49% of the correct problems (between 5 and 30% more than experts), and their average validity (f-measure) is 51% (between 12 and 31% less than experts). Notice however that the notion of "correct rating" is defined on the basis of what the majority of *experts* said; thus it is expected that non-experts would score more poorly than experts in this respect.

In our previous BW study [26], we also compared experts and non-experts data and we found that correctness of non-experts to be between 6 and 18% less than for experts and thus a smaller difference than what we found for WCAG 2.0 here; for sensitivity the difference was between 17 and 26%, a larger difference than for WCAG 2.0; for f-measure the difference was between 14 and 36%, slightly larger than what we found for WCAG 2.0. This may be interpreted as saying that when using BW, validity of results produced by students compared to experts is slightly worse than when using WCAG 2.0.

In this study we did not collect information regarding which WCAG techniques did the participants use when assessing each SC; the qualitative answers we collected do not mention them. In retrospect, it would have been interesting to know how consistent the participants were in adopting them.

Finally, we noted earlier that some experts that we invited refused to take part in this study. In one particular case, the expert indicated that the problem was conformance requirement 4 of WCAG 2.0 which reads as follows "*Only Accessibility-Supported Ways of Using Technologies: Only accessibility-supported ways of using technologies are relied upon to satisfy the success criteria. Any information or functionality that is provided in a way that is not accessibility supported is also available in a way that is accessibility supported.*" He claimed that without a document that includes a list of accessibility-supported technology uses for English web content, we would not be able to reliably evaluate pages for WCAG 2.0. However, to our knowledge no such document exists as a supporting document for WCAG 2.0. Therefore, whoever wants to evaluate pages against WCAG 2.0, they need to do the evaluation without such a document. Since we wanted to keep the study as close as possible to real world evaluations, we did not generate such a document.

6. CONCLUSION & FUTURE WORK

In this paper, we have presented a study which was performed with 22 experts and 27 non-experts, asking them to rate all 61 WCAG 2.0 success criteria on four different web pages. In summary, results show that for experts, half of the success criteria fail to meet the 80% agreement threshold; experts produce also 26% incorrect ratings, produce 20% false positives and miss 32% of the true problems. We also compared performance of experts against that of non-experts evaluators; we found that agreement for non-experts drops by 10%, incorrect ratings reach 32%, false positives 42% and false negatives 49%.

We believe that training is important to reduce the expertise gap on WCAG 2.0, even though in the paper we did not focus on the reasons behind such differences. When we look at the literature some researchers investigated the factors involved in the evaluator effect: in [12] authors argue that a vague evaluation procedure may make different evaluators focus on different things during the evaluation; [18] shows that there is a clear effect of the evaluator's cognitive style on heuristic evaluation and [6] demonstrates that the individual judgments of severity were highly personal. Similar to

these, [12] concludes that "the principal cause for the evaluator effect is that usability evaluation is a cognitive activity which requires that the evaluators exercise judgment". [14] shows that evaluators occasionally fail to observe the evidence of a particular problem. In our future work, we would like to investigate these further which will give us a better idea about how to improve testability and validity of WCAG 2.0 success criteria. Another interesting research is to compare effectiveness of other accessibility guidelines, such as Section 508.

Acknowledgments and Experimental Data

We would like to thank all the experts that we contacted, and especially those who contributed to this study. Many thanks also to the students of the course "Progettazione di siti web 2009–2010" who spent a lot of time performing the evaluation. Data of this study can be found at the Human Centred Web (HCW) Laboratories' data repository, `http://hcw-eprints.cs.man.ac.uk/132/`.
The URLs of the pages used in this study are: `http://www.new.facebook.com/group.php?gid=2416052053`, `http://www.imdb.com/title/tt0068646`, `http://www.bloomberg.com/news/worldwide`, `http://www.scientificamerican.com/biotechnology`.

7. REFERENCES

[1] S. Abou-Zahra. Web accessibility evaluation. In S. Harper and Y. Yesilada, editors, *Web Accessibility: A Foundation for Research*, Human-Computer Interaction Series, chapter 7, pages 79–106. Springer, London, first edition, Sept. 2008.

[2] F. Alonso, J.L. Fuertes, A.L. González, and L. Martínez. On the testability of wcag 2.0 for beginners. In *Web for All - W4A 2010*, Raleigh, USA, April 2010. ACM.

[3] G. Brajnik. Comparing accessibility evaluation tools: a method for tool effectiveness. *Int. Journal on Universal Access in the Information Society*, 3(3–4):252–263, Oct. 2004.

[4] G. Brajnik. Beyond conformance: the role of accessibility evaluation methods. In S. Hartmann, X. Zhou, and M. Kirchberg, editors, *WISE 2008: 9th Int. Conference on Web Information Systems Engineering – 2nd International Workshop on Web Usability and Accessibility IWWUA08*, LNCS 5176, pages 63–80, Auckland, New Zealand, Sept. 2008. Springer-Verlag. Keynote speech.

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

[6] M. Catani and D. Biers. Usability evaluation and prototype fidelity: users and usability professionals. In *Proc. of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 1998.

[7] K.P. 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.

[8] DRC. The web: Access and inclusion for disabled people. Technical Report, Disability Rights Commission (DRC), UK, 2004.

[9] A. D. N. Edwards. Assistive technologies. In S. Harper and Y. Yesilada, editors, *Web Accessibility: A Foundation for Research*, Human-Computer Interaction Series, chapter 10, pages 142–162. Springer, London, first edition, Sept. 2008.

[10] S.L. 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.E. Jacobsen. The evaluator effect during first-time use of the cognitive walkthrough technique. In *Proc. of HCI International on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I*, pages 1063–1067, 1999.

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

[13] M. Hertzum, N.E. 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.

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

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

[16] C. Jay, D. Lunn, and E. Michailidou. End user evaluations. In S. Harper and Y. Yesilada, editors, *Web Accessibility: A Foundation for Research*, Human-Computer Interaction Series, chapter 8, pages 107–126. Springer, London, first edition, September 2008.

[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. Journal of Human-Computer Studies*, 67(4):382–393, 2009.

[19] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proc. of CHI 1992*, pages 373–380, Monterey, CA, USA, May 1992. ACM.

[20] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, 1993.

[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] J. Rubin and D. Chisnell. *Handbook of Usability Testing*. Wiley, second edition, 2008.

[23] J. Slatin and S. Rush. *Maximum Accessibility: Making Your Web Site More Usable for Everyone*. Addison-Wesley, 2003.

[24] J. Thatcher, M. Burks, C. Heilmann, S. Henry, A. Kirkpatrick, P. Lauke, B. Lawson, B. Regan, R. Rutter, M. Urban, and C. Waddell. *Web Accessibility: Web Standards and Regulatory Compliance*. FriendsofED, 2006.

[25] W3C/WAI. Requirements for WCAG 2.0 Checklists and Techniques. `http://www.w3.org/TR/2003/WD-wcag2-tech-req-20030207`, 2003.

[26] Y. Yesilada, G. Brajnik, and S. Harper. How Much Does Expertise Matter? A Barrier Walkthrough Study with Experts and Non-Experts. In *Proc. of 11th Int. ACM SIGACCESS Conference on Computers and Accessibility – ASSETS 2009*, pages 203–210, Pittsburgh, PA, Oct. 2009.