Validity and Reliability of Web Accessibility Guidelines

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

ABSTRACT

Although widely used, Web Content Accessibility Guidelines (WCAG) have not been studied from the viewpoint of their validity and reliability. WCAG 2.0 explicitly claim that they are based on "testable" criteria, but no scientific evidence exists that this is actually the case. Validity (how well all and only the true problems can be identified) and reliability (the extent to which different evaluations of the same page lead to same results) are key factors for quality of accessibility evaluation methods. They need to be well studied and understood for methods, and guidelines, that are expected to have a major impact.

This paper presents an experiment aimed at finding out what is the validity and reliability of different checkpoints taken from WCAG 1.0 and WCAG 2.0. The experiment employed 35 young web developers with some knowledge on web accessibility. Although this is a small-scale experiment, unlikely to provide definite and general answers, results unequivocally show that with respect to the kind of evaluators chosen in the experiment, checkpoints in general fare very low in terms of reliability, and that from this perspective WCAG 2.0 are not an improvement over WCAG 1.0.

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 guidelines, evaluation

ASSETS'09, October 25-28, 2009, Pittsburgh, Pennsylvania, USA.

1. INTRODUCTION

The maturity level of a discipline reflects on the body of shared knowledge concerning the methodologies, methods, techniques and tools that practitioners and scientists use; their pros and cons, strengths and weaknesses, their applicability (pre)conditions should be known and agreed upon. Unfortunately this appears not to be the case for web accessibility.

To evaluate accessibility of websites several methods can be employed, including conformance reviews, user testing, subjective assessments and screening techniques [7, 6, 22]. It is reasonable to expect that these methods differ in terms of their validity, reliability, efficiency and usefulness. See [24, 2, 4] for a conceptual framework on quality of accessibility evaluation methods (AEMs), and [12, 13, 14, 15, 3] for actual analyses of AEMs.

One problem with accessibility evaluation methods is that many unwanted factors can creep in and bias the results; for example, the evaluator effect (*i.e.* the observation that different evaluators in similar conditions identify substantially different sets of usability problems, [8, 9]) and expertise levels (*i.e.* differences in evaluators expertise influence identified accessibility problems, [24]), the difficulty in reliably rating severity of problems (*i.e.* user ratings differ from evaluators ratings and both differ from checkpoint priorities, [15]), validity of the guidelines and complexity of the WAI conformance model [11], bias caused by the sampling method adopted to select pages to evaluate [5], bias caused by error rates of the testing tools used [1, 19]. The fact is that which accessibility problems are identified and how their severity is rated are two aspects of accessibility investigations that lack substantial standardization, potentially leading to low reproducibility of results.

The recent release of Web Content Accessibility Guidelines 2.0, and the widespread influence that they have on private and public organizations, including regulatory bodies, increases the importance to analyse them to understand their strengths and weaknesses, and take appropriate corrective actions if needed. However, no study has been published regarding quality of accessibility guidelines (not even for WCAG 1.0, released in May 1999).

In this paper I present an experimental evaluation of effectiveness of WCAG. The experiment was conducted with 35 junior evaluators who were asked to evaluate conformance of web pages with respect to WCAG 1.0 and WCAG 2.0. Results are presented in terms of several metrics related to effectiveness, namely: accuracy rates, proportions of false positives and true negatives, agreement rates. Although this

is a small-scale experiment, unlikely to provide definite and general answers, results show unequivocally that with respect to the kind of evaluators chosen in the experiment, guidelines in general fare very low in terms of reliability, and that WCAG 2.0 are not an improvement over WCAG 1.0.

2. QUALITY OF ACCESSIBILITY EVALUATION METHODS

Accessibility can be defined in various ways; see [2] for nine different ones. Depending on its definition, in the best of cases, accessibility is a property that can be observed empirically, after setting up formal or informal experiments where certain user performance indexes are measured. For example, when accessibility is meant as "if disabled people can use a website with the same effectiveness as non-disabled people" [20, 17], then indexes like success levels, success rates, number of errors can be used to draw appropriate conclusions on accessibility. However, when employing analytical methods no such indexes can be used, and only indirect inferences can be made starting from subjective opinions of evaluators. The analytic evaluation method becomes a conceptual tool for *observing* the desired property, more or less like a radio-telescope can be used to "observe" distant phenomena. For guidelines-based methods, checkpoints are assertions that need to be demonstrated (through direct experience) to be either true or false when applied to a website. An assertion is called *testable* when it can be demonstrated to be true if it is true, or false if it is false.

The important properties for an accessibility evaluation method are *effectiveness* (the ability to systematically and accurately predict all and only the real problems that will occur when real users use the site), *efficiency* (the effort, time, cost, infrastructure needed to apply the method), *usefulness* (related to usability of the results produced by the method) and *usability* (how easily can the method be learned, remembered, applied). In this paper I focus on effectiveness, which is probably the most important quality criterion. To be effective, a method should be *valid* (*i.e.* the extent to which it yields all and only the true accessibility problems) and *reliable* (*i.e.* the extent to which application of the method on the same pages in different contexts, like different evaluators, leads to the same results).

2.1 The testability problem for WCAG

WCAG 2.0 are claimed to be based on testable success criteria: "For each guideline, testable success criteria are provided to allow WCAG 2.0 to be used where requirements and conformance testing are necessary such as in design specification, purchasing, regulation, and contractual agreements" [23].

"Testable" is defined in another document [21]. "*Reliably Human Testable:* 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".

The experiment described in the rest of the paper aims at understanding how testable are WCAG 1.0 and WCAG 2.0 checkpoints and if there is a relationship between testability and validity.

3. RESEARCH QUESTIONS AND PROCEDURE

In the context of conformance reviews based on either WCAG 1.0 or WCAG 2.0, the following research questions are addressed:

1. Which checkpoints[1] are more or less reliable? How are the reliability figures related to the 80% threshold used in the definition of *reliably human testable*? Knowing this would help in isolating the problem and estimate the effects of known reliability levels onto conformance claims.
2. What are some of the causes for such differences? Is the difficulty in understanding whether the checkpoint applies to the current page (*applicability*) a cause? Is the difficulty of actually evaluating the checkpoint (*evaluability*) a cause? Knowing this would help isolating the problem and identifying the reasons, leading perhaps to additional investigations and reformulation of checkpoint assertions.
3. Which checkpoints achieve higher or lower accuracy levels? Is there any relationship between accuracy and reliability? If there were such a positive relationship, then increasing reliability could likely lead to an increase in accuracy levels, yielding a more effective and more reliable evaluation method.
4. Are there differences between the guidelines sets?

3.1 Procedure

The experiment[2] was conducted with university students of a course on user-centered web development; when involved in the study, students had already attended about 16 hours of lectures on web accessibility; they were also told that their outcome, ranked in terms of F-measure (see below), was to be used in the final rating for the exam. They were exposed mainly to the Italian technical accessibility requirements, which are a mixture of WCAG 1.0 and Section 508 requirements [10]. Students age was in the 21-46 range ($M = 24, \text{sd} = 4$). Their mean subjective rating of knowledge in web accessibility is 2.3 (sd=0.9), on a scale 1 to 5, 5 being the highest (as gathered through a questionnaire with Likert scale questions).

They were instructed to read instructions pages (located in the course website) that asked them:

1. to read/review the official pages for WCAG 1.0 and WCAG 2.0
2. to read the "Understanding" documents for the WCAG 2.0 21 success criteria we preselected;
3. to fill-in a demographic questionnaire;
4. to download four spreadsheets, 2 for each of the two webpages we selected[3]; for each page, one spreadsheet was for WCAG 1.0 and the other for WCAG 2.0. Each spreadsheet contained the list of checkpoints, with links to the appropriate W3C resource, a drop-down menu for the rating {`fail`, `na`, `pass`}, a drop-down menu for applicability difficulty (0 to 4), another for testability difficulty (0 to 4), and a notes field to be filled in with

[1] In the following, "checkpoint" and "success criterion" are used as synonyms.

[2] Data collected for this experiment is available for inspection and further analyses; see `http://www.dimi.uniud.it/giorgio/projects/wcag`.

[3] `www.hotel.it` and `www.tg1.rai.it`

WCAG 1.0	1.1, 1.4, 3.4, 3.5, 3.6, 5.1, 5.3, 6.1, 6.4, 7.3, 8.1, 9.2, 10.1, 12.3, 12.4, 13.1, 13.3, 13.4, 13.6, 14.1
WCAG 2.0	1.1.1, 1.2.3, 1.3.1, 1.3.2, 1.3.3, 1.4.1, 1.4.3, 1.4.4, 2.1.1, 2.2.2, 2.4.1, 2.4.3, 2.4.4, 2.4.5, 2.4.10, 3.1.5, 3.2.1, 3.2.2, 3.3.2, 4.1.2

Table 1: Preselected checkpoints (21 for each set of guidelines).

a short explanation of the chosen rating. The list of checkpoints, the pair of pages, and the order of the guidelines sets were randomized, to counterbalance order effects. Participants had also to write the time required to do the evaluation.

5. To download a final questionnaire with comments on the execution;
6. to email to the course instructor all the files within a 5-day period.

The pages and the subsets of checkpoints that we used were chosen in order to: (a) provide a sufficiently large number of potential violations for many of the available checkpoints; and, at the same time (b) be viable as test pages in a test that was planned to require no more than 4 hours of work. For this reason we restricted the set of checkpoints to be considered; see Table 1. The two sets of checkpoints are equivalent in terms of the accessibility problems they cover; the only WCAG 2.0 checkpoints that are not included in the WAI comparison table are 1.3.3 and 3.3.2. We also excluded checkpoints that were obviously not applicable to selected pages (for example, 1.2.1: "Prerecorded audio and video").

3.2 Dependent variables

The experiment was based on two within-subjects factors, checkpoint and page. Judge ratings for a checkpoint on a page are coded as `fail`, when the checkpoint is not satisfied; `pass`, when it is satisfied; and `na` when it is not applicable.

Reliability, which is the extent to which independent evaluations of the same page with respect to the same criteria yield same results, is measured in terms of *max-agreement*: given a checkpoint and a page, it is the relative frequency of the modal score (the rating that occurs more frequently, i.e. the one on which the highest proportion of evaluators agreed on). Thus, for example, if 3 evaluators said `fail`, 4 said `na`, and 3 said `pass`, max-agreement would be $\frac{4}{10} = 0.4$.

When evaluating a page with respect to a checkpoint, evaluators were asked to rate also the difficulty they faced when determining if the checkpoint applied to the current page (*applicability*) and when determining the outcome of the checkpoint, provided it was deemed applicable (*evaluability*). Both these opinions were scored in a scale from 0 to 4.

Accuracy, one of the metrics used to represent *validity*, is measured on the basis of the proportion of correct ratings that each evaluator gave for each page and each group of checkpoints. Correct ratings were given by an independent, more experienced, judge ("GD") and further reviewed by the author.

The other metric used to measure validity is F-measure. Given a page and a set of guidelines, the set of *true violations* (TV) is the set of checkpoints that were rated as `fail` by "GD" on that page; given a judge, a page, and a set of guidelines, the set of *reported violations* (RV) is the set of checkpoints that the judge rated as `fail`. We can then define *correctness* as the fraction of reported violations that are also true ones $C = \frac{|RV \cap TV|}{|RV|}$; *sensitivity* as the fraction of true violations that were reported $S = \frac{|RV \cap TV|}{|TV|}$; and finally F-measure as $F = \frac{2C \cdot S}{C+S}$, the harmonic mean of C and S.

Notice that effectiveness used in this experiment refers to conformance rather than accessibility, since we did not consider true accessibility problems, in the sense defined in Section 2.

4. RESULTS

Thirty five judges produced 140 evaluations (2 pages by 2 sets of 21 checkpoints each), for a total of 2635 individual valid ratings (Table 2 provides the detailed numbers). Each checkpoint was evaluated by 28 to 33 judges, with a mean of 31.4. These numbers include also judge "GD".

	fail	na	pass	sum
wcag10	594	154	571	1319
wcag20	535	184	597	1316
sum	1129	338	1168	2635

Table 2: Number of ratings.

It is noticeable that roughly the same number of fail/pass ratings were found, globally and across the two guidelines, which suggests that the two sets of checkpoints are balanced.

In terms of time required to complete each evaluation (for each of four combinations of guidelines and pages), the overall mean time is 81 minutes (sd=48); for WCAG 1.0, $M = 79, \text{sd} = 56$; for WCAG 2.0, $M = 83, \text{sd} = 37$. The difference is significant ($T(2307) = 2.23, p = 0.026$) but negligible in magnitude (a difference of 4 minutes out of 79, about 5%), especially when compared to the high standard deviation (the effect size is in fact a mere $\mathbf{d} = 0.088$).

Reliability.

Tables 3 and 4 provide the values for max-agreement for all checkpoints of both sets of guidelines and on both pages (which are not shown). One table gives those that are significantly below the 80% threshold, whereas the other one gives those that overlap the threshold. Each table provides checkpoint number, the guidelines it belongs to, a short textual description, the value for the max-agreement, the p-value resulting from testing the hypothesis "max-agreement is close to 80%" (computed by the proportion test, two-tailed), the modal score and its absolute frequency; entries are sorted by decreasing values of max-agreement. Listed p-values are the probability of error if the alternative hypothesis is accepted, *i.e.* that "max-agreement differs from 80%".

The tables show that max-agreement ranges from 0.36 to 0.94, that 12 checkpoints of WCAG 1.0 and 16 of WCAG 2.0 are definitely below the 80% threshold (i.e. have a p-value < 0.05), and that 16 from WCAG 1.0 and 10 from WCAG 2.0 overlap with the threshold (i.e. have a p-value ≥ 0.05). The 10 worst checkpoints in terms of max-agreement are "1.3.3 Sensory Characteristics", "9.2 Device Independent Interfaces", "3.2.2 On Input", "2.4.1 Bypass Blocks", "10.1 No Pop-Ups", "7.3 No Movement", "1.4.1 Use of Color", "1.4.4 Resize Text", "2.4.4 Link Purpose (in Context)", "4.1.2 Name, Role, Value".

On the other hand, the 10 best checkpoints are "13.3 Lay-

out Info", "14.1 Clear Language", "2.4.5 Multiple Ways", "6.1 Without Style Sheets", "5.1 Identify Row/Col Headers", "1.1 Text Equivalents", "1.4 Synchronize", "13.4 Consistent Navigation", "3.4 Relative Units", "1.4.3 Contrast (Minimum)".

It is remarkable that checkpoints believed to be ambiguous (for example, "14.1 Clear Language") turn out to be quite reliable; and that no checkpoint has a max-agreement that is significantly above the threshold.

There is a main effect, on max-agreement, by guidelines set (ANOVA $F(1,41) = 9.93, p < 0.0030$), by checkpoint ($F(40,41) = 2.2, p < 0.0072$), and by page ($F(1,41) = 4.18, p < 0.0473$); no interactions occur. If we discard the checkpoint factor, then only guidelines have a main effect ($F(1,81) = 6.26, p < 0.0144$). For WCAG 1.0, $M = 0.68, \text{sd} = 0.14$; for WCAG 2.0, $M = 0.61, \text{sd} = 0.12$; a two-sample t-test shows that there is a significant difference due to guidelines: $T(80.9) = 2.47, p < 0.0154$, effect size $\mathbf{d} = 0.54$, 95% confidence interval for the difference $[0.014, 0.127]$.

Figure 1 illustrates the 95% confidence intervals for max-agreement for each pair $\langle \text{checkpoint}, \text{page} \rangle$. It can be seen that the width of the intervals is similar for the all the checkpoints, indicating that the amount of uncertainty is the same across checkpoints and guidelines; also the range across the two guidelines is quite similar. Although none of the checkpoints have an interval that is above the threshold, WCAG 2.0 has more checkpoints that are definitely below it, which is confirmed also by the result of the t-test.

Applicability and evaluability difficulties.

Table 5 shows the mean values for applicability and evaluability; in terms of applicability, the most difficult outcome is `na`, as expected; in terms of evaluability there is no difference between `fail` and `pass`.

	fail	na	pass	overall
applicability	0.79	0.94	0.75	0.79
evaluability	1.09	0.41	1.09	1.00

Table 5: Means for applicability and evaluability difficulties, by outcome and overall (on a scale 0:4, 4=highest difficulty).

Applicability is not normally distributed (it is negatively skewed with a peak on 0); a systematic pairwise Wilcoxon test with Holm's p-value adjustment shows that there is no significant difference due to the outcome of the rating $\{\texttt{fail}, \texttt{na}, \texttt{pass}\}$; similarly, there is no significant difference due to the page being evaluated; there is however a significant difference due to guidelines: for WCAG 1.0, $M = 0.66, \text{sd} = 0.92, \text{median} = 0$; for WCAG 2.0, $M = 0.92, \text{sd} = 1.06, \text{median} = 1$; $W > 6 \cdot 10^6, p < 0.0001$.

Evaluability is also not normally distributed. A pairwise Wilcoxon comparison on rating outcome confirms that `na` differs significantly from the other two values, which are similar to each other. Also in this case there is a significant difference due to guidelines: for WCAG 1.0, $M = 0.95, \text{sd} = 1.07, \text{median} = 1$; for WCAG 2.0, $M = 1.06, \text{sd} = 1.11, \text{median} = 1$; $W > 7 \cdot 10^6, p = 0.0142$. There is no significant difference due to the page being evaluated.

There is a moderate significant correlation between applicability and evaluability (Pearson's $r = 0.74, p < 0.0001$), and a weak negative significant correlation of applicability with max-agreement ($r = -0.28, p = 0.0087$); no significant correlation is present between max-agreement and evaluability.

Table 6 and 7 show the best and worst checkpoints with respect to applicability and evaluability difficulties.

checkpoint	name	applicability
	BEST	
5.1	Identify Row/Col Headers	0.17
6.1	Without Style Sheets	0.34
3.4	Relative Units	0.39
14.1	Clear Language	0.40
13.1	Identify Link Target	0.42
1.1	Text Equivalent	0.45
12.4	Form Labels	0.45
7.3	No Movement	0.48
1.4.4	Resize Text	0.50
	WORST	
1.2.3	Audio Description	1.09
1.3.2	Meaningful Sequence	1.11
3.1.5	Reading Level	1.22
6.4	Device Independent Events	1.24
3.2.2	On Input	1.26
1.3.3	Sensory Characteristics	1.31
8.1	Scripts and Applets	1.45
1.3.1	Info and Relationships	1.48
4.1.2	Name, Role, Value	1.83

Table 6: Best and worst checkpoints for applicability (0 to 4, 4=most difficult).

checkpoint	name	evaluability
	BEST	
5.1	Identify Row/Col Headers	0.17
5.3	No Layout Table	0.52
1.2.3	Audio Description	0.56
7.3	No Movement	0.48
5.1	Identify Row/Col Headers	0.63
1.4.1	Use of Color	0.88
2.2.2	Pause, Stop, Hide	0.59
1.4	Synchronize	0.57
6.1	Without Style Sheets	0.34
1.1	Text Equivalent	0.45
	WORST	
9.4	Logical Tab Order	0.65
13.4	Consistent Navigation	0.84
2.1.1	Keyboard	1.06
7.3	No Movement	0.93
3.6	List Items	0.76
14.1	Clear Language	0.62
6.4	Device Independent Events	1.24
4.1.2	Name, Role, Value	1.54
1.3.1	Info and Relationships	1.48
8.1	Scripts and Applets	1.45

Table 7: Best and worst checkpoints for evaluability (0 to 4, 4=most difficult).

Accuracy.

Correct ratings given by the judges (excluding "GD") are 1502 over 2551 ($M = 0.59$, 95% confidence interval $[0.57, 0.61]$). For WCAG 1.0, correct ratings are 843 over 1277 ($M = 0.66$, 95% confidence interval $[0.63, 0.69]$); for WCAG 2.0, they are 659 over 1274 ($M = 0.52$, 95% confidence interval $[0.49, 0.55]$). The difference is significant ($\chi^2(1) = 53.18, p < 0.0001$, 95% confidence interval of the difference is $[0.10, 0.18]$). Figure 2 shows accuracy scores for individual checkpoints: several WCAG 2.0 checkpoints are below all WCAG 1.0 checkpoints, and no WCAG 2.0 checkpoint is above all WCAG 1.0.

checkpoint	wcag	short	p.value	max.agr	mode	freq
1.3.3	wcag20	Sensory Characteristics	0.00	0.36	fail	12
1.3.3	wcag20	Sensory Characteristics	0.00	0.40	fail	12
9.2	wcag10	Device Independent Interfaces	0.00	0.42	fail	13
3.2.2	wcag20	On Input	0.00	0.42	na	14
2.4.1	wcag20	Bypass Blocks	0.00	0.43	pass	13
10.1	wcag10	No Pop-Ups	0.00	0.45	pass	14
3.2.2	wcag20	On Input	0.00	0.47	fail	14
7.3	wcag10	No Movement	0.00	0.48	na	15
1.4.1	wcag20	Use of Color	0.00	0.48	pass	16
1.4.4	wcag20	Resize Text	0.00	0.50	fail	15
2.4.4	wcag20	Link Purpose (In Context)	0.00	0.50	fail	15
4.1.2	wcag20	Name, Role, Value	0.00	0.52	fail	16
12.4	wcag10	Form Labels	0.00	0.53	pass	17
8.1	wcag10	Scripts and Applets	0.00	0.53	fail	17
2.2.2	wcag20	Pause, Stop, Hide	0.00	0.53	na	16
2.4.10	wcag20	Section Headings	0.00	0.53	pass	16
4.1.2	wcag20	Name, Role, Value	0.00	0.54	fail	15
2.2.2	wcag20	Pause, Stop, Hide	0.00	0.55	fail	18
12.4	wcag10	Form Labels	0.00	0.55	pass	17
13.1	wcag10	Identify Link Target	0.00	0.55	fail	17
12.3	wcag10	Divide Blocks of Info	0.00	0.56	pass	18
9.4	wcag10	Logical Tab Order	0.00	0.56	pass	18
1.4.1	wcag20	Use of Color	0.00	0.57	pass	17
2.4.10	wcag20	Section Headings	0.00	0.58	fail	19
2.4.4	wcag20	Link Purpose (In Context)	0.00	0.58	pass	19
2.4.5	wcag20	Multiple Ways	0.00	0.58	pass	19
9.2	wcag10	Device Independent Interfaces	0.01	0.58	fail	18
9.4	wcag10	Logical Tab Order	0.01	0.58	pass	18
1.1.1	wcag20	Non-Text Content	0.01	0.60	fail	18
1.3.1	wcag20	Info and Relationships	0.01	0.60	pass	18
2.1.2	wcag20	No Keyboard Trap	0.01	0.60	pass	18
3.2.1	wcag20	On Focus	0.01	0.60	pass	18
3.3.2	wcag20	Labels or Instructions	0.01	0.60	pass	18
1.4.4	wcag20	Resize Text	0.01	0.61	fail	20
3.1.5	wcag20	Reading Level	0.01	0.61	pass	20
3.3.2	wcag20	Labels or Instructions	0.01	0.61	fail	20
6.1	wcag10	Without Style Sheets	0.01	0.61	pass	20
5.3	wcag10	No Layout Table	0.02	0.61	pass	19
13.1	wcag10	Identify Link Target	0.02	0.62	fail	20
1.3.1	wcag20	Info and Relationships	0.02	0.62	pass	20
1.4	wcag10	Synchronize	0.02	0.62	na	20
3.6	wcag10	List Items	0.02	0.62	pass	20
3.1.5	wcag20	Reading Level	0.04	0.63	pass	19
7.3	wcag10	No Movement	0.04	0.63	fail	19

Table 3: Checkpoints and their max-agreement on both pages; p-values for the null hypothesis "Max-agreement overlaps with the 80% threshold". These checkpoints are below the threshold.

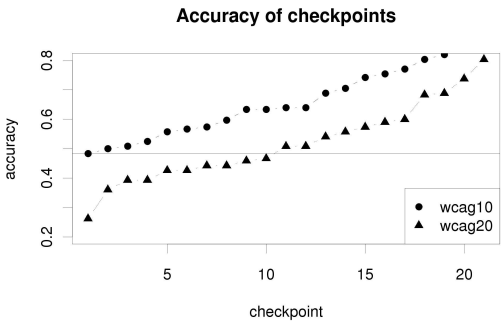

Figure 2: Accuracy of individual checkpoints.

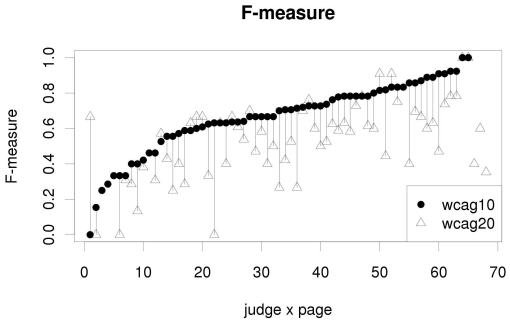

Figure 3: For each pair of $\langle \text{judge}, page \rangle$, F-measure for WCAG 1.0 (circles) and difference with F-measure for WCAG 2.0 (segments).

checkpoint	wcag	short	p.value	max.agr	mode	freq.
13.4	wcag10	Consistent Navigation	0.05	0.65	pass	20
3.5	wcag10	Use Header	0.05	0.65	fail	20
6.4	wcag10	Device Independent Events	0.05	0.65	fail	20
6.4	wcag10	Device Independent Events	0.05	0.65	fail	20
1.1	wcag10	Text Equivalent	0.07	0.66	fail	21
12.3	wcag10	Divide Blocks of Info	0.07	0.66	pass	21
13.6	wcag10	Group Links	0.07	0.66	fail	21
2.1.1	wcag20	Keyboard	0.07	0.66	fail	21
3.5	wcag10	Use Header	0.07	0.66	fail	21
1.2.3	wcag20	Audio Description	0.09	0.67	fail	22
2.1.2	wcag20	No Keyboard Trap	0.09	0.67	pass	22
5.3	wcag10	No Layout Table	0.14	0.68	pass	21
2.4.3	wcag20	Focus Order	0.21	0.69	pass	20
2.4.3	wcag20	Focus Order	0.21	0.70	pass	23
1.3.2	wcag20	Meaningful Sequence	0.25	0.70	pass	21
2.1.1	wcag20	Keyboard	0.25	0.70	fail	21
10.1	wcag10	No Pop-Ups	0.30	0.71	fail	22
5.1	wcag10	Identify Row/Col Headers	0.35	0.72	na	23
1.3.2	wcag20	Meaningful Sequence	0.69	0.76	pass	25
2.4.1	wcag20	Bypass Blocks	0.69	0.76	fail	25
3.2.1	wcag20	On Focus	0.69	0.76	pass	25
1.2.3	wcag20	Audio Description	0.82	0.77	na	23
13.3	wcag10	Layout Info	0.89	0.77	fail	24
13.6	wcag10	Group Links	0.89	0.77	fail	24
3.6	wcag10	List Items	0.89	0.77	pass	24
8.1	wcag10	Scripts and Applets	0.89	0.77	fail	24
14.1	wcag10	Clear Language	0.96	0.78	pass	25
3.4	wcag10	Relative Units	0.96	0.78	fail	25
1.1.1	wcag20	Non-Text Content	1.00	0.79	fail	26
1.4.3	wcag20	Contrast (Minimum)	1.00	0.79	fail	26
1.4.3	wcag20	Contrast (Minimum)	0.82	0.83	fail	25
3.4	wcag10	Relative Units	0.45	0.87	fail	27
13.4	wcag10	Consistent Navigation	0.40	0.88	pass	28
1.4	wcag10	Synchronize	0.25	0.90	fail	27
1.1	wcag10	Text Equivalent	0.23	0.90	fail	28
5.1	wcag10	Identify Row/Col Headers	0.23	0.90	na	28
6.1	wcag10	Without Style Sheets	0.23	0.90	pass	28
2.4.5	wcag20	Multiple Ways	0.11	0.93	pass	28
14.1	wcag10	Clear Language	0.10	0.94	pass	29
13.3	wcag10	Layout Info	0.07	0.94	pass	31

Table 4: Checkpoints and their max-agreement on both pages; p-values for the null hypothesis "Max-agreement overlaps with the 80% threshold". There is not sufficient evidence to claim that these checkpoints differ from the threshold.

Accuracy is moderately correlated with max-agreement (Pearson's $r = 0.65, p < 0.0001$), weakly and negatively with applicability ($r = -0.39, p < 0.0003$), but not with evaluability.

F-measure.

The other metric used to measure validity is F-measure, computed over each combination of ⟨judge, page, guidelines set⟩. The overall mean for F-measure is 0.59, sd=0.21, and the 95% confidence interval is $[0.55, 0.62]$. For WCAG 1.0, we get $M = 0.65, \text{sd} = 0.20, \text{c.i.} = [0.61, 0.71]$; for WCAG 2.0, $M = 0.52, \text{sd} = 0.21, \text{c.i.} = [0.47, 0.57]$. Such a difference is significant ($T(121.3) = 3.64, p < 0.0004, \mathbf{d} = 0.65$, 95% confidence interval for the difference $[0.06, 0.20]$).

Beside a main effect of guidelines, ANOVA reveals also a main effect of page ($F(1) = 27.31, p < 0.0001$) and no interactions. As can be seen from Figure 3, most of the differences in F-measure across all pairs of ⟨judge, page⟩ are in favour of WCAG 1.0; the differences are also quite large in magnitude, as confirmed by the effect size and the confidence interval (reported in the previous paragraph).

5. DISCUSSION

In terms of reliability, WCAG 1.0 are superior to WCAG 2.0, by a differences that range from 1% to 13%. Both guidelines have checkpoints that get close to the minimum (36%) and the maximum (94%). Consider that the minimum max-agreement that we could ever get on a 3-point scale is 33%. It is also unfortunate that no checkpoint is significantly more reliable than 80%. As seen, reliability does not depend on pages as much as it does on the guidelines used. Low agreement values could be used as cues for assigning low confidence levels to results produced by those checkpoints.

We expected that certain checkpoints had particularly low reliability (*e.g.*, "1.3.1 Info and Relationships"), but they are not on the top; on the other hand, apparently unambiguous checkpoints like "1.4.4 Resize Text", "2.4.4 Link Purpose", "1.4.1 Use of Color" are worse. Conversely, checkpoints that we thought were unreliable, like "14.1 Clear Language", turned out to reach at least 78% of reliability.

The perception that evaluators had about reliability, expressed in terms of applicability and evaluability difficulties, does not match actual reliability results (weak negative correlation of max-agreement and applicability, no correlation of max-agreement and evaluability). In fact, there is little overlap of checkpoints between the best/worst ones in terms of max-agreement, and those for applicability and evaluability. One way to interpret this is that applicability and/or evaluability difficulties cannot be used to explain differences in max-agreement. In any case, the difference of WCAG 1.0

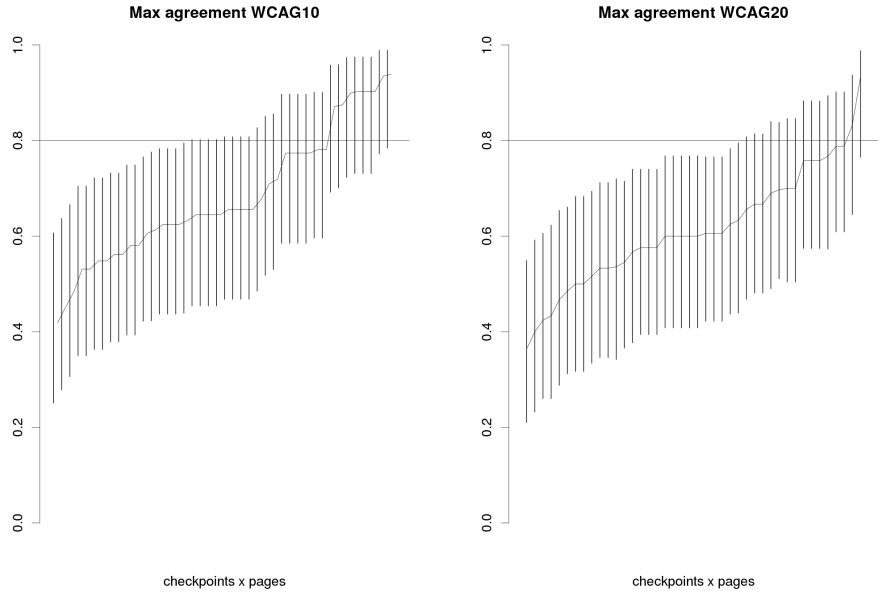

Figure 1: Confidence intervals for max-agreement; solid line is the actual value for the sample, on both sites.

vs WCAG 2.0 in terms of applicability is 0.26 over 4 (or 6.5%); for evaluability it is 0.11 (2.75%). Although being significant, both are small in magnitude; we believe these differences should be neglected.

The only explanation we have for why applicability and evaluability show a high correlation is that it may be difficult to distinguish one from the other while figuring out what rating should be given to a checkpoint.

A global accuracy rate ranging between 57% to 61% is low, compared to what we found in other similar studies, with accuracy rates of up to 79% for similar types of judges [24]. A difference in accuracy of 10% to 18% due to guidelines set is a large one; it means that simply by switching the guidelines to be used will increase or reduce the number of checkpoints that are correctly assessed by 18% in the worst case. The moderate correlation of accuracy with max-agreement means that even if we used a majority rule to decide what a correct rating would have been, we would get similar results in terms of accuracy. Further analysis could confirm that we could expect increases of reliability to lead to increases of validity, and viceversa.

In terms of F-measure, a global value of 59% is slightly higher than we found in similar studies (where it reached 49% for inexperienced evaluators and 59% only for more experienced ones). A difference between guidelines from 6% to 20% is substantial; in other words, it means that both correctness and sensitivity may have changed by that amount, and this in turn means that as many as 20% more false positives could be returned *and* as many as 20% fewer true problems may have been reported.

Thus, in terms of validity, WCAG 1.0 are better than WCAG 2.0.

6. CONCLUSIONS

This experiment leads to the conclusion that there are large differences in effectiveness for the different checkpoints, and also between guidelines sets. While WCAG 1.0 score better than WCAG 2.0 on reliability and validity, none of the guidelines sets have checkpoints whose reliability is definitely higher than 80%. The experiment failed in explaining the causes for differences in reliability.

There are also differences in validity between checkpoints and the guidelines sets, which we cannot explain either.

Although there are several checkpoints requiring interpretation when being evaluated, in many cases this subjective process does not lead to low reliability. This is the case for the "image equivalence" and "clear language" checkpoints.

These conclusions, however, should be scoped by the type of evaluators that we choose: young web developers with limited practical experience in usability and accessibility, and no practical experience with either set of guidelines (except for a classroom exposure to the Italian technical requirements, which are closer to WCAG 1.0 than WCAG 2.0). Although this sample is representative of an important segment of the relevant audience for accessibility guidelines, both in age and possible experience with WCAG 1.0 and WCAG 2.0, similar experiments conducted with different types of evaluators could lead to different conclusions.

Another important limit of this study is due to using only two pages. It is entirely possible that on different pages results would come out different, even if employing the same type of evaluators.

For these reasons I hope that other similar experiments are carried out, to confirm or contradict these results, and possibly to cover other conditions (like more pages, different evaluators, perhaps more experienced ones).

Other future research openings include a qualitative analysis of notes and comments produced by participants to find out possible explanations for the differences in validity across checkpoints, in such a way to provide suggestions to standards developers. Another interesting analysis would be to understand which are the checkpoints that were given the same rating by any 2, 3, 4, ... evaluators.

Acknowledgments

Many thanks to my students who participated to this study, and to Giulio Darrigo who managed to set up instruction pages and spreadsheets, and helped me collecting the data.

7. REFERENCES

[1] 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.

[2] 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.

[3] G. Brajnik. A comparative test of web accessibility evaluation methods. In S. Harper, editor, *10th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Halifax, Canada, Oct. 2008. ACM Press.

[4] G. Brajnik. Towards a sustainable accessibility. In *Accessible Design in the Digital World*, York, UK, September 22-24 2008. `http://www.addw08.org`York University.

[5] G. Brajnik, A. Mulas, and C. Pitton. Effects of sampling methods on web accessibility evaluations. In S. Trewin and E. Pontelli, editors, *9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Tempe, AZ, Oct. 2007. ACM Press.

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

[7] 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`.

[8] 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.

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

[10] Italian Government. Technical assessment and technical accessibility requirements of internet technology-based applications. `http://www.pubbliaccesso.it/normative/DM080705-A-en.htm`, July 2005.

[11] 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.

[12] 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.

[13] 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.

[14] H. Petrie, F. Hamilton, N. King, and P. Pavan. Remote usability evaluations with disabled people. In *CHI 2006: Proc. of the SIGCHI conference on Human factors in computing systems*, pages 1133–1141, New York, NY, USA, 2006. ACM.

[15] 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.

[16] G. Sampson-Wild. Testability costs too much. `http://www.alistapart.com/articles/testability`, June 2007.

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

[18] J. Smith. Testability in WCAG 2.0. `http://webaim.org/blog/wcag-2-testability`, June 2007.

[19] 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.

[20] U.S. Dept. of Justice. Section 508 of the Rehabilitation Act. `www.access-board.gov/sec508/guide/1194.22.htm`, 2001.

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

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

[23] W3C/WAI. Web content accessibility guidelines (wcag) 2.0. World Wide Web Consortium — Web Accessibility Initiative, `www.w3.org/TR/WCAG20`, December 2008.

[24] 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*, Pittsburgh, PA, Oct. 2009. ACM SIGACCESS.