On the testability of WCAG 2.0 for beginners

Fernando Alonso
Dept. LSIIS
Technical University of Madrid
Campus de Montegancedo
28660-Boadilla del Monte
Madrid. Spain
+34 91 352 25 46
falonso@fi.upm.es

José Luis Fuertes
Dept. LSIIS
Technical University of Madrid
Campus de Montegancedo
28660-Boadilla del Monte
Madrid. Spain
+34 91 352 25 46
jfuertes@fi.upm.es

Ángel Lucas González
Dept. LSIIS
Technical University of Madrid
Campus de Montegancedo
28660-Boadilla del Monte
Madrid. Spain
+34 91 352 25 46
agonzalez@fi.upm.es

Loïc Martínez
Dept. LSIIS
Technical University of Madrid
Campus de Montegancedo
28660-Boadilla del Monte
Madrid. Spain
+34 91 352 25 46
loic@fi.upm.es

ABSTRACT

Web accessibility for people with disabilities is a highly visible area of research in the field of ICT accessibility, including many policy activities across many countries. The commonly accepted guidelines for web accessibility (WCAG 1.0) were published in 1999 and have been extensively used by designers, evaluators and legislators. W3C-WAI published a new version of these guidelines (WCAG 2.0) in December 2008. One of the main goals of WCAG 2.0 was testability, that is, WCAG 2.0 should be either machine testable or reliably human testable. In this paper we present an educational experiment performed during an intensive web accessibility course. The goal of the experiment was to assess the testability of the 25 level-A success criteria of WCAG 2.0 by beginners. To do this, the students had to manually evaluate the accessibility of the same web page. The result was that only eight success criteria could be considered to be reliably human testable when evaluators were beginners. We also compare our experiment with a similar study published recently. Our work is not a conclusive experiment, but it does suggest some parts of WCAG 2.0 to which special attention should be paid when training accessibility evaluators.

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

Keywords

Web accessibility, Web accessibility evaluation, Teaching Web accessibility

W4A2010 - Technical, April 26-27, 2010, Raleigh, USA. Co-Located with the 19th International World Wide Web Conference.

INTRODUCTION

One of the key aspects of the social integration of people with disabilities today is information and communication technologies (ICT) accessibility. ICTs are the core of the information society. This is recognized in article 9 – Accessibility of the United Nations Convention on the Rights of Persons with Disabilities [10]:

> *"1. To enable persons with disabilities to live independently and participate fully in all aspects of life, States Parties shall take appropriate measures to ensure to persons with disabilities access, on an equal basis with others, to the physical environment, to transportation, to information and communications, including information and communications technologies and systems, and to other facilities and services open or provided to the public, both in urban and in rural areas. These measures, which shall include the identification and elimination of obstacles and barriers to accessibility, shall apply to, inter alia:*
>
> *1. Buildings, roads, transportation and other indoor and outdoor facilities, including schools, housing, medical facilities and workplaces;*
>
> *2. Information, communications and other services, including electronic services and emergency services"*

The web is an essential component of the information society and, as such, has been lent particular attention by the promoters of accessibility for people with disabilities.

The commonly accepted guidelines for web accessibility were the Web Content Accessibility Guidelines (WCAG) 1.0, published in 1999 by the World Wide Web Consortium (W3C) [5]. These guidelines have been in use for several years, and there is an arguably large consensus among practitioners about how to interpret and evaluate them.

A new version of WCAG has been under development for several years, and was published in December 2008 as WCAG 2.0 [4]. This new version had two main goals. Firstly, it aimed to be technology-independent, so it was to be applicable to current and future web technologies, either from the W3C or from other sources. Secondly, it was to be testable, that is, practitioners should agree about how to evaluate the conformance of a web site with WCAG 2.0. This paper focuses on this second WCAG 2.0 goal: testability.

WCAG 2.0 has a different language, a different structure and a different rationale to WCAG 1.0. All of these influence how conformance with WCAG 2.0 is to be evaluated in the future, either manually or with the support of evaluation tools, as listed by the W3C [12].

We have run an experiment with students of an intensive course on web accessibility. They all had to evaluate the accessibility of the same web page, according to the 25 level-A success criteria of WCAG 2.0. Here we present an analysis of the results and a comparison with a recent similar experiment.

The content of this paper is structured as follows. Section 2 will provide an overview of WCAG 2.0. Section 3 will deal with the key concept of testability. Section 4 will describe the approach used for the educational experiment and Section 5 will analyze the results. Section 6 will compare our results with a recent experiment by Brajnik [3] and will draw some concluding remarks.

WCAG 2.0 OVERVIEW

WCAG 2.0 is a W3C Recommendation published in December 2008 [4]. This document contains three layers of guidance: principles, guidelines and success criteria:

- The principles provide the foundation for web accessibility. There are four principles:
    1. Perceivable: Users must be able to perceive the information being presented (it must not be undetectable by all of their senses).
    2. Operable: Users must be able to operate the interface (the interface cannot require interaction that a user cannot perform).
    3. Understandable: Users must be able to understand the information as well as the operation of the user interface (the content or operation cannot be beyond their understanding).
    4. Robust: Users must be able to access the content as technologies advance (the content should remain accessible as technologies and user agents evolve).
- The 12 guidelines provide the basic goals that web designers should work toward in order to make content more accessible to users with different disabilities. The guidelines are not testable, but provide the framework and overall objectives to help web designers understand the success criteria and better implement the techniques.
- For each guideline, testable success criteria are provided so that WCAG 2.0 can be used wherever requirements and conformance testing are necessary, such as in design specification, purchasing, regulation and contractual agreements. In order to meet the needs of different groups and different situations, three levels of conformance are defined: A (lowest), AA, and AAA (highest).

There are additional layers of guidance provided by an external document that supplements WCAG 2.0: “Techniques for WCAG 2.0” [9]. This document is “informative” and provides three additional layers, referred to as sufficient techniques, advisory techniques and common failures:

- Sufficient techniques provide guidance and examples for meeting the guidelines using specific technologies. The sufficient techniques are considered sufficient by the W3C to meet the success criteria. However, it is not necessary to use these particular techniques. If techniques other than those listed by the W3C are used, then some other method for establishing the technique's ability to meet the success criteria would be needed. Most success criteria list multiple sufficient techniques. Any of the listed sufficient techniques can be used to meet the success criterion. There may be other techniques not documented by the W3C that could also meet the success criteria. As new sufficient techniques are identified, they can be added to the listing.
- Advisory techniques. There are a number of advisory techniques that can enhance accessibility, but did not qualify as sufficient techniques because they are not sufficient to meet the full requirements of the success criteria, they are not testable, and/or are good and effective techniques in some circumstances but not effective or helpful in others. Web designers are encouraged to use these techniques wherever appropriate to increase the accessibility of their web pages.
- Common failures. These are examples of bad practices that cause web pages to fail to meet the success criteria. Failures during evaluation are interpreted differently than for techniques: if a common failure is found in a web page, then that web page fails the respective success criterion.

In addition to these layers of guidance, there is another important part of WCAG 2.0: the conformance section. This section lists five requirements for conformance to WCAG 2.0: (1) one conformance level is met in full; (2) conformance is for full web pages; (3) all web pages in a process conform to the same level; (4) only accessibility-supported ways of using the technologies are relied upon to satisfy the success criteria; and (5) technologies that are used in a way that is not accessibility supported do not interfere with the accessibility of the page.

The conformance section also gives information about how to make optional conformance claims. Finally, it describes what “accessibility supported” means, since only accessibility-supported ways of using technologies can be relied upon for conformance [11].

TESTABILITY IN WCAG 2.0

One key concept behind the development of WCAG 2.0 is testability. According to W3C, the success criteria are written as testable sentences, that is, each criterion is written as a statement that will be either true or false when specific Web content is tested against it. The goal is to objectively determine if content satisfies the success criteria. While some of the testing can be automated using software evaluation programs, human testers are required for part of or the entire test in other cases.

This is recognized by the W3C when providing a definition of testability for techniques [6]: a technique is testable if it is either machine testable or reliably human testable. It is machine testable if there is a known algorithm (regardless of whether that algorithm is known to be implemented in tools) that will determine absolutely reliably whether or not the technique has been implemented. It is reliably human testable if the technique can be tested by human inspection, and it is believed that at least 80% of knowledgeable human evaluators would agree on the finding.

There was some debate about the implications of testability during the development of WCAG 2.0 [7, 8], but now that WCAG 2.0 is complete the real challenge is to find out whether the success criteria and the techniques are actually testable.

The testability of WCAG 2.0 depends on several factors. The first one is the objectivity of the language used to write the success criteria. In fact, the language of many success criteria is more precise and objective in WCAG 2.0.

Let us look at an example to see this shift in the language from WCAG 1.0 to WCAG 2.0. The requirement regarding color contrast in WCAG 1.0 is checkpoint 2.2, which reads [5]:

> *"2.2 Ensure that foreground and background color combinations provide sufficient contrast when viewed by someone having color deficits or when viewed on a black and white screen [Priority 2 for images, Priority 3 for text]."*

In this checkpoint it is not clear what "sufficient contrast" is (no formula is given and no thresholds are provided). In addition, this sufficient contrast depends on unspecified needs of people with color blindness or people using black-and-white displays.

Several formulas for computing this color contrast and different thresholds were devised over the years that WCAG 1.0 was in use. This made it difficult for the evaluators to agree on whether a web page conformed with this checkpoint.

The corresponding success criterion in WCAG 2.0 is 1.4.3 [4]:

> *"1.4.3 Contrast (Minimum): The visual presentation of text and images of text has a contrast ratio of at least 4.5:1, except for the following: (Level AA)*
>
> *Large Text: Large-scale text and images of large-scale text have a contrast ratio of at least 3:1;*
>
> *Incidental: Text or images of text that are part of an inactive user interface component, that are pure decoration, that are not visible to anyone, or that are part of a picture that contains significant other visual content, have no contrast requirement.*
>
> *Logotypes: Text that is part of a logo or brand name has no minimum contrast requirement."*

There are several evident changes. First, we have a given threshold (4.5:1) of a concept (contrast ratio) that is fully defined in WCAG 2.0. Second, there are some explicit exceptions concerning large text (with a less restrictive threshold) and incidental content and logotypes (with no contrast requirement).

So, in this particular case, it is clear that success criterion 1.4.3 is testable, whereas checkpoint 2.2 is not. But it is not clear that this is true for all the success criteria.

A second factor affecting the testability of WCAG 2.0 is the clarity of the language used. Not only do the success criteria need to be objective, but they also have to be clear enough to avoid confusion. The editors of WCAG 2.0 have put a lot of effort into clarity, by providing well-defined terminology and using it consistently throughout the recommendation, as the above example illustrates. However, some of the terminology is new, and it will take evaluators some time to get used to the new terms when evaluating pages against WCAG 2.0.

A third factor for the testability of WCAG 2.0 is the openness of the techniques and failures. The techniques belong to a non-normative document [9] that is intended to be a living document that will change as new techniques are defined, either inside or outside the W3C.

This sets a challenge for testability. If, for a given web page element and a given success criterion, none of the documented techniques apply and none of the common failures apply, then the evaluator has no information to decide whether or not the success criterion has been met. There could be a technique for that particular element in that particular case that makes the content accessible. However, if the W3C has not yet documented the technique, it will be difficult to provide a reliable result.

Of course, this is not an issue with the most basic web technologies (such as standard HTML 4.01 elements), because, given all the experience gained since WCAG 1.0 was written, it is well known how they can be made accessible.

Given all of these factors that may influence the testability of WCAG 2.0, we decided to perform an experiment to analyze what happens when the students attending an intensive course on web accessibility apply WCAG 2.0 to evaluate the accessibility of a single web page.

THE EXPERIMENT

As part of our academic activity at our university, we have been teaching intensive courses within the framework of the ATHENS Program for the last few years [2]. This program involves the twice yearly (March and November) organization of exchange courses: students from the program partners (24 European universities) travel to other institutions to attend one-week intensive courses that are accompanied by cultural exchange activities.

In our case, we have been teaching a web accessibility course as part of this program since 2005. During our most recent course, held in March 2009, we used WCAG 2.0 for the first time. The main contents of the course are as follows:

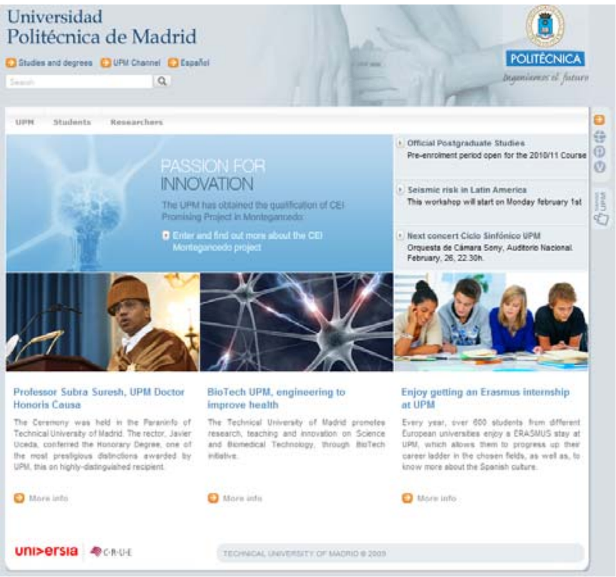
1. Introduction: disabilities, independent living, design for all, standards, legislation
2. The Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C)
3. Web Content Accessibility Guidelines (WCAG): principles, guidelines, success criteria, techniques
4. Evaluation of Web Accessibility.

The core of the course is the teaching of WCAG 2.0, which accounts for more than 60% of the workload. In the March 2009 course we used collaborative learning techniques, in particular jigsaw-based sessions [1]: the 17 students that attended the course were given fragments of contents that they had to work with, discuss with the others and, finally, present to all. We had three jigsaw sessions:

- Principles, guidelines and success criteria.
- Techniques (sufficient techniques for level-A success criteria only).
- Failures (for level-A success criteria only).

The students attending the course were assessed based on their participation in the collaborative learning sessions and how well they completed an exercise. The exercise consisted of evaluating the same web page according to the WCAG 2.0 success criteria.

The page to be analyzed was the English version of our university's homepage, which contains recent news about our university and access to the main content areas of the site. Figure 1 shows that web page with content on the date of writing. The content of the web page evaluated in March, 2009 was different, but the structure and layout was the same.



PASSION FOR INNOVATION

The UPM has obtained the qualification of CEI Promising Project in Montegancedo

Enter and find out more about the CEI Montegancedo project

Official Postgraduate Studies
Pre-enrolment period open for the 2010/11 Course

Seismic risk in Latin America
This workshop will start on Monday february 1st

Next concert Ciclo Sinfónico UPM
Orquesta de Cámara Sony, Auditorio Nacional. February, 26, 22.30h.

Professor Subra Suresh, UPM Doctor Honoris Causa

The Ceremony was held in the Paraninfo of Technical University of Madrid. The rector, Javier Uceda, conferred the Honorary Degree, one of the most prestigious distinctions awarded by UPM, this on highly-distinguished recipient.

BioTech UPM, engineering to improve health

The Technical University of Madrid promotes research, teaching and innovation on Science and Biomedical Technology, through BioTech initiative.

Enjoy getting an Erasmus internship at UPM

Every year, over 600 students from different European universities enjoy a ERASMUS stay at UPM, which allows them to progress up their career ladder in the chosen fields, as well as, to know more about the Spanish culture.

TECHNICAL UNIVERSITY OF MADRID © 2009

Figure 1. The web page evaluated in the experiment (with updated content)

Given that there were no automated tools providing support for WCAG 2.0 at the time, all the evaluations were done manually, and the students had to fill in a spreadsheet (which we provided) with the detailed results of their accessibility assessment.

The values that the students could assign to each success criterion were:

- Pass: the content conforms with the success criterion
- Fail: the content does not conform with the success criterion
- Partial: the content almost conforms with the success criterion. Only easy-to-solve minor issues prevent the content from fully conforming.
- Not applicable (NA): the success criterion is not applicable to the existing content.
- Unknown: the student is unable to decide on success criterion conformance.

Students were set the exercise on the second day of the course. They were given time slots to perform the exercise, and they were asked to submit the results by midday on the last day of the course. A discussion session was then held to enable students to defend their evaluation results, and all the participants were entitled to comment on these results.

This exercise is the starting point of our WCAG 2.0 testability experiment. Our experimental data are the ratings provided by the 17 students for the 25 level-A success criteria. These were the success criteria that almost all the students were able to complete in the set amount of time.

We also have the ratings provided by two course instructors, who are experts in web accessibility evaluation according to WCAG 1.0. The two instructors evaluated the web page separately and then discussed the results and agreed on what they could consider to be the "correct value" for each success criterion. It has to be noted that, initially, the two instructors only agreed on the values for 13 of the 25 success criterion, although this point is not relevant for evaluating WCAG 2.0 testability, given the low number of expert evaluators. When they met and looked at their respective evaluation results, they were able to agree on the final value. The two experts found it quite easy to reach agreement in most cases: the typical issue was that one of the experts discovered a relatively hidden failure that the other overlooked. Note also that at that time (March 2009) the experts had limited practical knowledge of WCAG 2.0. In addition, we can say that, now, a year later, we have greater levels of agreement when evaluating websites.

In the experiment data, we merged the "partial" and "fail" values into a single, more generic "fail" category. This provides a stricter interpretation of conformity assessment for WCAG 2.0: the content is considered to fail even if there are only minor outstanding issues. This way we do away with the subjectiveness of deciding on the severity of the content accessibility problems. Concerning the "unknown" value, only two students assigned this value to one of the success criteria, and thus we have overlooked this value. We will only use the other 15 values for that particular success criterion (4.1.2).

Concerning students' previous knowledge, none of them had received any training in accessibility and particularly web accessibility before attending our course. In addition, they all actively participated in the collaborative learning sessions so we can infer that they have a similar degree of knowledge of web accessibility and WCAG 2.0.

The goal of our experiment was to find out whether the 25 level-A success criteria can be considered testable by beginners according to the definition of "reliably human testable" given by the W3C: 80% of evaluators should agree on the result of each success criterion. More formally:

- Hypothesis: all the 25 level-A success criteria of WCAG 2.0 can be considered to be reliably human testable by beginners. This means that for each success criterion 80% of the students should agree on one value, and this value should be the correct one as provided by the expert evaluators.
- Null hypothesis: there is one or more level-A success criteria that cannot be considered to be reliably human testable.

RESULTS OF THE EXPERIMENT

We examined our students' and our own results, and we compared the results for each success criterion. Figure 2 shows, for each success criterion, the number of times that each possible value (pass, fail, not applicable) appeared in the results provided by our students. It also shows what was considered to be the right result, that is, the result generated by the instructors, preceded by a check mark (✓) and written in bold style.

The first noteworthy point is that the number of success criteria where the majority of students provided the correct result is relatively low: 16 success criteria. For the remaining cases (9), the majority of students gave an incorrect response.

The second key issue is the percentage of agreement on the results provided by the students:

- 80% or more of students (that is, 14 or more students out of a total of 17) agreed on 11 success criteria (44%)
- 75% or more of the students (that is, 13) agreed on 13 success criteria (52%).
- 70% or more of the students (that is, 12) agreed on 14 success criteria (56%).

- 64% or more of the students (that is, 11) agreed on 17 success criteria (68%).

There were four success criteria for which the level of agreement was just above half of students (52.94%, that is, 9 out of 17): 1.3.2 (meaningful sequence), 2.2.1 (timing adjustable), 3.1.1 (language of page) and 3.3.1 (error identification). And there was one success criterion on which fewer than half of the students agreed: 3.3.2 (labels or instructions) with 41.18% (that is, 7 out of 17).

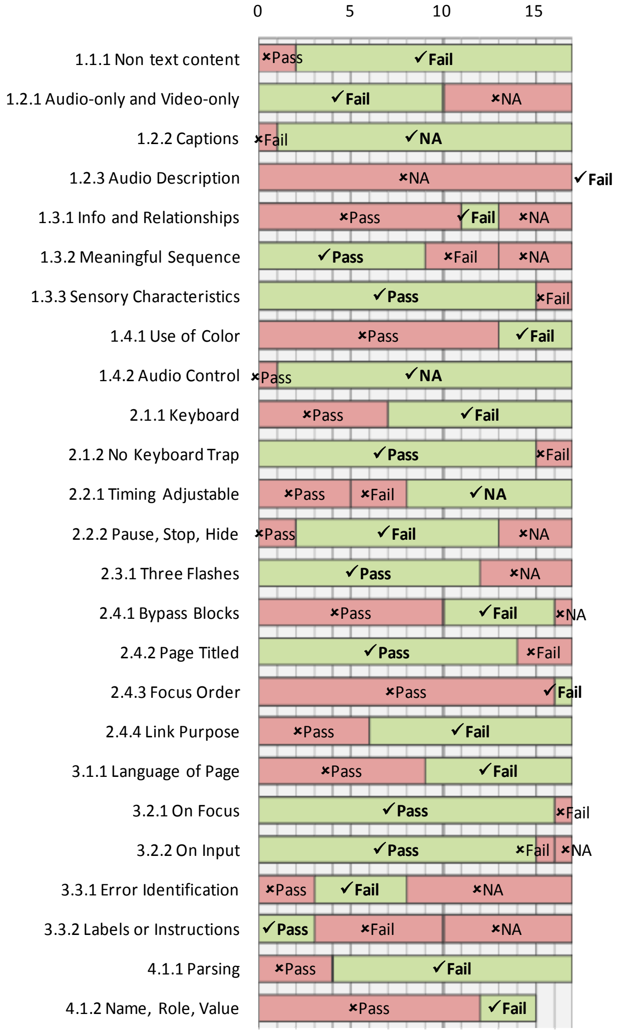


Figure 2. Results of evaluating Level-A success criteria during a web accessibility course

Combining these two observations we can analyze the reliability of the evaluation of the success criteria. In our case we consider that a success criterion has been reliably evaluated if the majority of the students provided the correct result and if there is a big enough majority. Given our results, we can use four percentage thresholds (80, 75, 70 and 64) to analyze reliability. The reliability data are summarized in Figure 3. Note that the value for the 4.1.2 success criterion has been scaled up to the 0-17 range that applies to all the other success criteria.

Figure 3 shows that only 8 success criteria (32%) are reliably evaluated at an 80% threshold of agreement. If we relax the agreement threshold, we get 9 (36%), 10 (40%) and 12 (48%) for thresholds of 75%, 70% and 64%, respectively.

This means that even with a relatively low threshold of 64% of the students (that is, 11 out of the 17), there are still 13 success criteria that are not reliably evaluated, that is, that cannot be considered to be testable by beginners in our experiment.

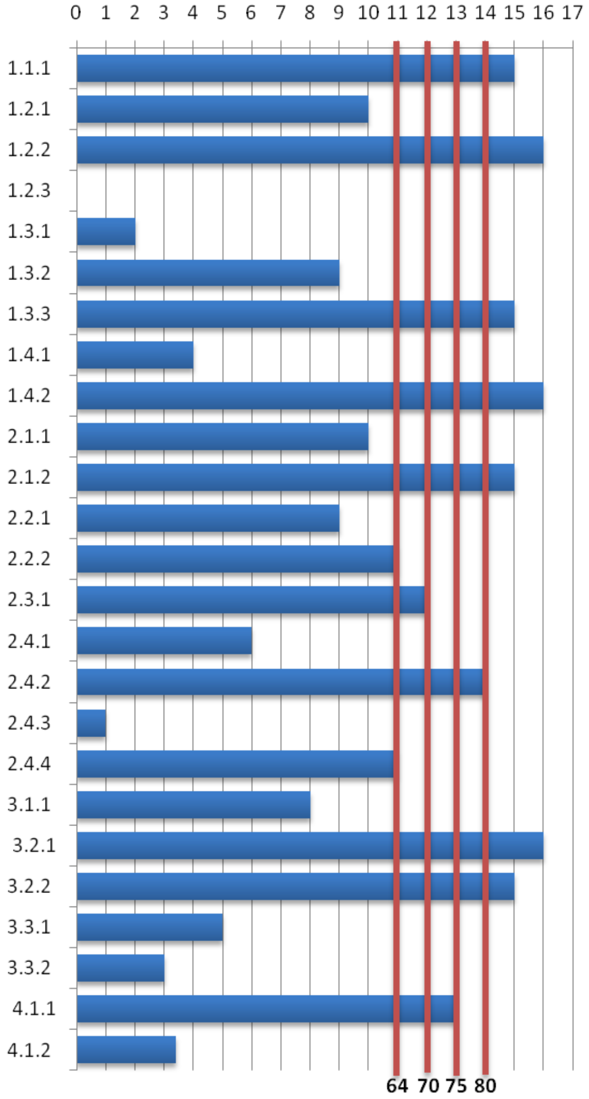


Figure 3. Reliability of the evaluation of the success criteria. Bars indicate de number of correct values provided for each success criteria

There are two categories of non-testable success criteria:

- First, we have some cases in which the majority of the students agreed on the correct value, but this majority was insufficient to consider the success criteria to be reliably human testable. The criteria that belong to this category are 1.2.1 (with 10 out of 17 good results), 1.3.2 (9), 2.1.1 (10) and 2.2.1 (9).
- Second, we have the cases where the majority of the students provided incorrect results. The criteria that belong to this

category are 1.2.3, 1.3.1, 1.4.1, 2.4,1, 2.4.3, 3.1.1, 3.3.1, 3.3.2 and 4.1.2.

We will discuss the details of these success criteria in these two categories.

Success criterion 1.2.1 deals with alternatives to audio-only and video-only content. For this criterion the correct result is “fail” and the responses given by students were: 10 fail and 7 not applicable. The problem with this criterion was the concept of “video-only” content. The analyzed web page contained a non-interactive Adobe Flash animation displaying some text about news of the organization and the user could click on any place of the animation to go to the news page. What happened is that 7 students failed to identify this content as “video-only” and thus decided that the success criterion was not applicable.

Success criterion 1.3.2 deals with the sequence for reading content. For this criterion the correct result is “pass”, and the responses given by the students were: 9 pass, 4 fail and 4 not applicable. The problem with this criterion is twofold:

- First, some of the students correctly identified the fact that the sequence in which the content was presented on the page affected its meaning but thought that the correct reading order could not be programmatically determined, although there were tabindex attributes that could provide that sequence.
- Second, some of the students were wrong in thinking that the web page did not have any content whose presentation sequence affected its meaning. The problem was that the definition of “sequence affecting the meaning of the content” was not clear for the students.

Success criterion 2.1.1 deals with keyboard access. For this criterion the correct result is “fail” and the responses given by students were: 7 pass and 10 fail. The problem is that several students failed to discover that the Flash animation could not be used without a pointing device. In this case the success criterion is clear, and it just happened that some students did not interact enough with the web page to discover interactions that were unreachable with the keyboard.

Success criterion 2.2.1 deals with user-adjustable timing. For this criterion the correct result is “not applicable” and the responses given by students were: 5 pass, 3 fail and 9 not applicable. The key element of this criterion is to identify whether the content sets time limits. In fact the page did not contain any and this was the reason for assigning a “not applicable” value. Some students reasoned in a more positive way and decided that if there were no time limits then the criterion was passed. This results in a different value than expected, but one that could be considered correct. Finally, only a few students (3) considered that animation implied time limits for reading and then marked this criterion as a failure, because they did not find any user control for time limits.

This brings to an end the discussion on the success criteria that the majority of students rated as correct, but where the majority was insufficient to be considered reliable. Now we will discuss the success criteria that the majority of students rated incorrectly.

Success criterion 1.2.3 deals with the provision of audio descriptions for video content. The correct value for this criterion is “fail”, but it was rated as “not applicable” by all of the students. The problem is very similar to 1.2.1: the students failed to recognize that the Flash animation provided visual-only content that has to be described to blind users. Probably this has to do with the language used in the success criterion, which is difficult for beginners to understand (what is the meaning of the expression “prerecorded video content of a synchronized media”?).

Success criterion 1.3.1 deals with information, structure and relationships. The correct value for this criterion is “fail”, and the responses given by students were: 11 pass, 2 fail and 4 not applicable. First of all, 4 students failed to realize that the page had content with information, structure and relationships. This was probably due to some gap in their knowledge of this criterion. On the other hand, 11 of the students thought that the content conformed to this criterion, when, in fact, it was wrong due to a small mistake: the header structure of the content was incorrect because there were no h1 elements.

Success criterion 1.4.1 deals with color use. The correct value for this criterion is “fail”, and the responses given by students were: 13 pass and 4 fail. In this case, the main mistake made by the students was that they did not notice that there were a few text links that were not visually evident without color vision. In this case the success criterion was easy to understand, but more difficult to evaluate. Students had to pay attention to subtle uses of color for representing links.

Success criterion 2.4.1 deals with bypassing blocks of content. For this criterion, the correct value was “fail”, and the responses given by students were: 10 pass, 6 fail and just 1 not applicable. The main issue with this point was that, although there was a link to skip navigation and go to the main content, there were several problems with the link: it was not visible using the keyboard or the mouse (there was a blank link and its text was only visible if the style sheets were disabled), the text was in Spanish (and the whole page is in English), and the link was wrong (it pointed to the wrong place). Most of the students who said that this criterion was correct thought that the fact that there was a link was enough (most of them did not notice the problems with the link or thought that the problems would be covered by other criteria).

Success criterion 2.4.3 deals with the navigation focus order. The correct result for this criterion is “fail”, but student responses were as follows: 16 pass and only 1 fail. The criterion fails because the page used the tabindex attribute to create a tab order that did not preserve meaning and page operability. The problem was confined to just a couple of links in the top right-hand corner of the page. These two links were in the last position of the tab order, but, when reading the page from left to right, from top to bottom, they were near the beginning of the page. Most of the students did not notice this point. Others thought that this was the correct order, as these two links represented access to web utilities (the intranet and the University’s on-line shop) and it was a good idea for them to be at the end. On the other hand, a few students found one or more sufficient techniques that were correct and they did not check for any failures.

Success criterion 3.1.1 deals with the natural language of the page. The correct value for this criterion was “fail” and the values provided by the students were: 9 pass and 8 fail. Analyzing the XHTML source code of the web page, we found that the language was indeed specified in a syntactically correct way. But the problem is that the language identified was Spanish whereas the whole web page was written in English. More than half of the students just checked that a language was identified, but they did not bother about checking whether or not this identification was semantically correct.

Success criterion 3.3.1 deals with error identification. The correct value for this criterion is “fail”. The students provided the following results: 3 pass, 5 fail and 9 not applicable. The web page had a search field with a default text inside the edit box (“Search”). The question is that there is no client-side or server-side validation of the search text field and it is possible for the user to launch the search for the default value of the text field (which is “Search”). The students that said that the criterion passed explained that if you input incorrect data in the search box, the search engine will not find anything and will indicate that no match is found (and they considered that this was a way to identify the error). The students that said that this criterion was not applicable thought that it was not mandatory to type information in the field or that there were no possible errors in a search field, as the search engine will search anything, and it is impossible to programmatically identify an input mistake.

Success criterion 3.3.2 deals with labels or instructions for input assistance. For this criterion the correct result is “pass”, and the responses given by students were: 3 pass, 7 fail and 7 not applicable. In the analyzed page there is a label (although invisible), and there is a “button” (an image of a magnifying glass with “search” as alternative text) for the search field, so these two elements are enough to tell the user what to do. Students thought that there were no instructions (they did not consider the button alternative text was enough) or no label (they did not notice that the label existed and appeared when disabling style sheets). The main problem was due to a strict interpretation of the success criterion: they thought that both labels and instructions should be provided when, in fact, one of these was enough.

And finally, success criterion 4.1.2 deals with names, roles and values. The correct result for this criterion is “fail”, but the students provided the following results: 12 pass and 3 fail (and two students failed to provide a value). The criterion fails because there was information missing from the form elements: the search text field has no title and its name is just “q”, plus the search button did not contain any information. The students did not notice this defective name, role or value information in the form elements.

After this detailed analysis, we have found that there are three main reasons for the students to provide incorrect values:

- **Comprehension** (C): Students found the language used in the success criterion or the concepts behind it hard to understand. In these cases we believe that the problem lies in the language used in WCAG 2.0. This made it difficult for our students to fully comprehend the meaning of and thus evaluate the success criteria, techniques and failures.
- **Knowledge** (K): Students lacked some knowledge that was required to correctly evaluate the success criterion. In these cases we believe that the problem lies in students’ training (which was limited, as it was an intensive course) and previous knowledge of web technologies, coming as they did from different fields.
- **Effort** (E): Students did not spend enough time and/or effort on evaluating the success criteria. In these cases we believe that the results could have been better if the students had put more effort into the evaluation.

Table 1 shows the reasons related to each success criterion, also showing the category of problems: either insufficient majority or wrong majority.

Table 1. Reasons for the lack of reliability of success criteria

SC	Category	C	K	E
1.2.1	Insufficient	X		
1.2.3	Wrong	X		
1.3.1	Wrong		X	X
1.3.2	Insufficient	X	X	
1.4.1	Wrong			X
2.1.1	Insufficient			X
2.2.1	Insufficient	X		
2.4.1	Wrong	X		X
2.4.3	Wrong		X	X
3.1.1	Wrong			X
3.3.1	Wrong	X	X	
3.3.2	Wrong	X		
4.1.2	Wrong			X

In addition, we can devise a global trend: the instructors performed a stricter evaluation than the students. That is, the instructors found more accessibility failures than the students. If we look at the values for the majority of students, we find that:

- Where the instructors find a failure (14 in total), students give a wide range of responses: fail (6), pass (6) or not applicable (2).
- Where the instructors determine that the success criterion is not applicable (3 cases), then the students agree (also 3 cases).
- Where the instructors find that the content conforms to the success criteria (8 cases), then the majority of students will agree in almost all cases (7 of 8).

DISCUSSION AND CONCLUSIONS

We have found a similar experiment conducted recently that we will use to discuss our results. Brajnik [3] published the result of an experiment on the effectiveness of both WCAG 1.0 and WCAG 2.0. His experiment is also based on using non-expert evaluators, although it has some differences with respect to ours:

- There are more evaluators (35 instead of 17).
- They evaluated two web pages instead of only one.
- They had to evaluate subsets of the checkpoints from WCAG 1.0 and the success criteria from WCAG 2.0.
- The non-expert evaluators had to estimate how difficult it was to find out whether the success criterion or checkpoint applied to the web page and how difficult it was to evaluate.

Although the scope of our experiment is more limited (and thus we cannot apply the same statistical rigor as Brajnik did), we can compare some of the results.

The subset of WCAG 2.0 success criteria that were evaluated in Brajnik’s experiment is: 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, and 4.1.2. Looking at our study, we have a smaller subset of 15 common success criteria.

We can analyze the maximum percentage of agreement for each success criterion between the evaluators (irrespective of whether that result was the good one), as shown in Figure 4. In the case of the Brajnik's experiment we calculated the arithmetic mean of the agreement values obtained for the two different web pages.

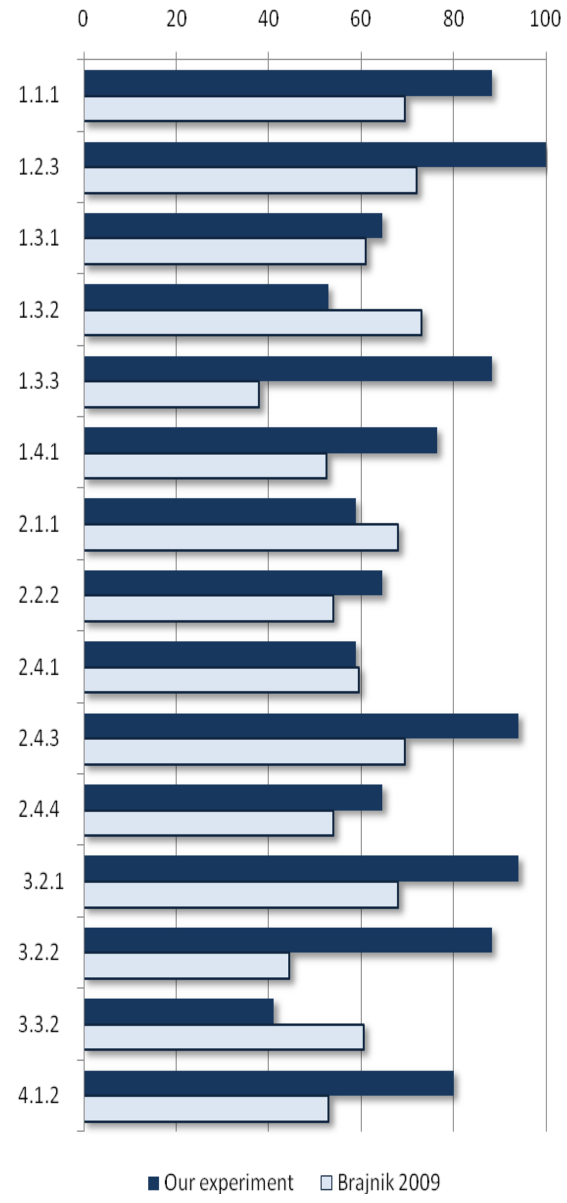


Figure 4. Comparison of maximum agreement values obtained for each success criterion between our and Brajnik's study

We can see that almost half of the 15 common success criteria have significant differences in maximum agreement (a difference of over 20%). Given that these experiments were performed by different people, for different web pages and in different educational contexts, it is difficult to interpret what the reasons for these differences are.

What we cannot do here is compare the accuracy for the judges in both experiments, because we and Brajnik measured accuracy differently. We can confirm, though, that the difference between the values of maximum agreement is independent of whether or not we have considered that a success criterion is reliable in our study.

But we can say that some success criteria show up as being particularly weak in both studies: 2.1.1 (timing adjustable) 2.2.2 (pause, stop, hide) 2.4.1 (bypass blocks) 2.4.4 (link purpose in context) and 3.3.2 (labels or instructions). It would be interesting to see whether this trend keeps up in further experiments run to complement Brajnik's and our studies.

Looking at our results and the comparison with Brajnik's experiment, we can draw a number of conclusions.

The first conclusion that we can reach from this experiment is that, in our case, WCAG 2.0 is far from testable for beginners. There are 13 success criteria on which either our students did not agree (fewer than 64%) or which the majority rated incorrectly. Brajnik came up with a similar overall result for WCAG 2.0 testability, although the details may be quite different.

The detailed results that we have obtained led us to identify three sources of unreliability, as shown in Table 1: comprehension, knowledge and effort.

The first cause, that is, the difficulties in interpreting the success criteria and their associated techniques and failures is directly attributable to WCAG 2.0 testability and, after more experiments are performed, could lead to proposals for rewriting those success criteria. This applied to 7 success criteria in our experiment. Due to the limitations of our study, we are not in a position to propose a better wording for any of these success criteria. We are running further experiments on both extensive and intensive courses to get a better understanding of the comprehension issues.

The second cause, that is, the lack of knowledge, will have an impact on the way we teach web accessibility and how we explain the corresponding success criteria (and their techniques and failures). In our experiment, this applied to 4 success criteria. We are using this information to prepare forthcoming courses and we will measure if we are able to improve the students' accuracy.

Finally, the third cause is related to the students, not to WCAG 2.0 testability. In our experiment this applied to 7 success criteria. It is reasonable to think that with better training and motivation the students can easily improve their results for the respective success criteria.

Of course, our experiment is limited by the number of students (17), the short student training time (one-week intensive course) and the lack of diversity of the web pages under evaluation. In fact we have just repeated the experiment with a different group of students. These students are attending a module that is part of the computer science degree taught at our university. This module also focuses on web accessibility. The main difference is that it is not an intensive module: it is taught from October 2009 to February 2010 and we still need to analyze the results. We also plan to involve expert evaluators in assessing the web pages in order to compare their results and confirm whether there is a real difference in WCAG 2.0 testability between beginners and experts, as has been shown for a different evaluation method, the Barrier Walkthrough [13].

Finally, we can conclude that although our experiment is not, by any means, conclusive, it does point out that the testability of the WCAG 2.0 success criteria is not to be taken for granted and that support material and tools will be needed to help evaluators to provide consistent results in the future.

REFERENCES

[1] Aronson, E., and Patnoe, S. The Jigsaw Classroom: Building Cooperation in the Classroom, Longman, New York, NY, 1997.

[2] The ATHENS Programme. http://www.athensprogramme.com/

[3] Brajnik, G. “Validity and Reliability of Web Accessibility Guidelines” Proceedings of ASSETS’09 (October 25-28, Pittsburg, Pennsylvania, USA). ACM Press. 2009. 131-138.

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

[5] Chisholm, W., Vanderheiden, G., and Jacobs, I. (eds.). Web Content Accessibility Guidelines 1.0. W3C Recommendation (1999). http://www.w3.org/TR/WCAG10/

[6] Requirements for WCAG 2.0 Checklists and Techniques. http://www.w3.org/TR/wcag2-tech-req/

[7] Sampson-Wild, G. Testability Costs Too Much. A List Apart (2007). http://www.alistapart.com/articles/testability/

[8] Smith, J. Testability in WCAG 2.0. WebAIM Blog (2007). http://webaim.org/blog/wcag-2-testability/

[9] Techniques for WCAG 2.0. http://www.w3.org/TR/WCAG20-TECHS/

[10] United Nations. Convention on the Rights of Persons with Disabilities. Adopted on 13 December 2006 at the UN Headquarters in New York. Opened for signature on 30 March 2007. Entered into force on 3 May 2008. http://www.un.org/disabilities/default.asp?navid=12&pid=150

[11] Understanding Conformance. http://www.w3.org/TR/UNDERSTANDING-WCAG20/conformance.html.

[12] Web Accessibility Evaluation Tools: Overview. http://www.w3.org/WAI/ER/tools/

[13] Yesilada, Y., Brajnik, G., and Harper, S. “How Much Does Expertise Matter?”. Proceedings of ASSETS’09 (October 25-28, Pittsburg, Pennsylvania, USA). ACM Press. 2009. 203-210.