Journal of Universal Computer Science, vol. 30, no. 1 (2024), 85-105
submitted: 9/2/2023, accepted: 18/9/2023, appeared: 28/1/2024

The use of WCAG and automatic tools by computer science students: a case study evaluating MOOC accessibility

Francisco Iniesto
(School of Computer Science, The National Distance Education University (UNED), Madrid
https://orcid.org/0000-0003-3946-3056, finiesto@lsi.uned.es)

Covadonga Rodrigo
(School of Computer Science, The National Distance Education University (UNED), Madrid
https://orcid.org/0000-0001-8135-3163, covadonga@lsi.uned.es)

Abstract: Web Content Accessibility Guidelines (WCAG) have been the de facto standard for Web accessibility evaluation for more than two decades and therefore have been introduced into legislation and university curriculum in Computer Science. At The National Distance Education University (UNED) in Spain, we have been teaching the guidelines for the last 15 years but learning how to apply WCAG criteria is complex. In this paper, we present the results of the analysis of students' performance in applying accessibility heuristic evaluation of an online resource (a Massive Open Online Course – MOOC) using WCAG. The experiment was carried out over two academic years to evaluate how accurate and easy it is to understand and use WCAG criteria by trained students as well as their perceptions of usefulness to evaluate accessibility barriers using automatic tools in combination with manual evaluation. Results from the study show that errors identified are aligned with accessibility evaluation literature: 65% of success criteria in WCAG do not reach 80% of agreement among raters which confirms the complexity of WCAG conformance. In total 62 (86%) criteria are marked as not being correctly addressed by automatic tools with an overlap of those showing false positives, and 25 criteria (34%) are indicated as difficult to evaluate manually. While all areas where raters disagree are potential opportunities for WCAG improvement, this research reinforces that WCAG evaluations are complex and difficult even with current automatic tools, and that the possible solutions for the way forward are: (1) a well-defined evaluation protocol including a combination of automatic tools and manual evaluations; (2) better training and professional development opportunities.

Keywords: Computer Science curriculum, accessibility teaching, accessibility evaluation, MOOC, WCAG
Categories: D.2.2, D.2.4, D.2.5, D.2.9, H.5.2, H.5.4, J.1
DOI: 10.3897/jucs.101704

1 Introduction

The Web Content Accessibility Guidelines (WCAG)[1], created by the World Wide Web Consortium (W3C)[2] as part of its Web Accessibility Initiative (WAI)[3], have been part

[1] WCAG, https://www.w3.org/WAI/standards-guidelines/wcag/
[2] W3C, https://www.w3.org/
[3] WAI, https://www.w3.org/WAI/

of the Web accessibility evaluation standards since 1999. WCAG guidelines are currently the most universally accepted set of Web accessibility guidelines, and the legislation to adopt these guidelines exist in several countries that seek to reduce disability discrimination [Seale et al., 2019]. The aim of the guidelines is to achieve an equally usable Internet for all, which is easy to understand and navigate, and interact with without barriers [Petrie et al., 2015]. WCAG guidelines are recurrently named in the most recent *"Inclusion and Education"* report [UNESCO, 2020] which assesses the progress towards Sustainable Development Goal 4 to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.

Web accessibility is nowadays included in the University curriculum as a part of 'Web design and development' in Computer Science courses [Baker et al., 2020; Waller et al., 2009]. Besides, Industry already considers accessibility as a key core component of their processes, and, hence, Computer Science students need to be prepared accordingly.

Research studies have reported on the usefulness of using WCAG to evaluate Web accessibility in educational contexts [Kumar et al. 2021]. However, application of WCAG requires expertise. In this paper, we present the results of a redesigned assignment in one undergraduate course devoted to learning the key concepts of Web accessibility and usability for human-computer interfaces as part of the Computer Engineering degree at The National Distance Education University (UNED), Spain. The empirical study was carried out over two academic years to evaluate how accurate and easy it is to understand and to use WCAG by students (who had already been trained) as well as their perceptions of the usefulness of automatic tools to evaluate accessibility barriers. A limitation of this research is that it is based on students who do not have the same experience as experts. All students evaluated the same resource: a Massive Open Online Course (MOOC).

2 Accessibility in computers science teaching

Web usability and accessibility are two interrelated concepts. Usability focuses on designing a website to meet users' expectations and adapting it to their needs efficiently and easily so that it enables optimal use by the target users [Weichbroth, 2020]. Accessibility and usability are related as less accessibility implies low usability; non-accessible content is not usable [Sauer et al., 2020]; however usable content may not necessarily be accessible [Petrie and Bevan, 2009]. Web accessibility is defined as [Petrie et al., 2015]:

> *All people, particularly disabled and older people, can use websites in a range of contexts of use, including mainstream and assistive technologies; to achieve this, websites need to be designed and developed to support usability across these contexts*.

2.1 Teaching accessibility

Computer scientists need to apply accessibility in terms of not only complying with legislation in their industry roles but should learn them as part of their curriculum when studying for their university degrees [Shinohara et al., 2018]. In that sense accessibility is part of most of the international educational curricula for Computer Science and Engineering degrees [Bohman, 2012]. But accessibility is not an easy topic to teach and

learn since it involves complex terminology and methodologies, as well, it is usually embedded in dedicated courses instead of including it across the whole curriculum [Baker et al., 2020; Lewthwaite et al., 2020].

Therefore, there exists a need to better integrate accessibility into the Computer Science curriculum [Gay et al., 2020]. Accessibility education in Computing Science presents challenging characteristics for those engaged in accessibility capacity building [Lewthwaite and Sloan, 2016]. That is related to the fact that not all Computer Science faculty know enough about accessibility; they may not have had the professional development to teach accessibility [Kawas et al., 2019]. Some of the best practices for teaching accessibility within the Computer Science curriculum include [Putnam et al., 2016]:

1. research projects that directly involve users with accessibility needs;
2. guest speakers who are experts in the area;
3. simulating disabilities to understand the reality of those with accessibility needs;
4. the use of videos or alternative formats that can be combined with the textbook; and
5. include other types of resources like research papers or online resources.

In that sense, the literature indicates that the use of a variety of methods is helpful such as traditional lectures, collaborative learning sessions, exercises on website evaluation and the development of accessible websites [Alonso, 2010].

2.2 Overview of WCAG

As detailed in the introduction one key aspect of Web accessibility is the WCAG set of guidelines. WCAG is an international standard for Web accessibility evaluation; legislation and policies across the world refer to WCAG compliance as their reference standard [Kumar et al., 2021].

There are several coexisting versions of WCAG: 1.0, 2.0 and 2.1, while version 2.2 is in a candidate recommendation status and 3.0 is being developed. WCAG 1.0 were critically important when first released in 1999 by the W3C. Versions 2.X have been designed to replace WCAG 1.0 because, with the rapid growth of Web technologies, they were obsolete. WCAG 2.X are differently organised, its first version 2.0 was available in 2008 and adopted as an international standard ISO 40500:2012[4]. They have four design principles (perceivable, operable, understandable, and robust - POUR). Principles contain guidelines, and each guideline has testable success criteria at levels A (lowest), AA (mid-range), or AAA (highest). WCAG 2.1, were released in June 2018, the updated guidelines include specific criteria for users with cognitive or learning disabilities and with low vision, and access from mobile devices is included. WCAG 2.1 guidelines have 13 guidelines and 78 success criteria which are written as testable statements that are not technology specific. The standard guides satisfying the success criteria in specific technologies, as well as general information about interpreting the success criteria. Other key differences between the latest versions 2.X with 1.0 are:

- The guidelines in 2.X are designed to be easier to test than those in 1.0.

[4] International standard ISO 40500:2012, https://www.iso.org/obp/ui/#iso:std:iso-iec:40500:ed-1:v1:en

- The 2.X versions reflect efforts to harmonise Web accessibility standards that are already in place.
- The usability is improved in versions 2.X. For example, they include specific instances to make them easier to follow.

While there is a lot of overlap between WCAG 2.X and WCAG 3.0, WCAG 3.0 will include additional tests and different scoring mechanisms. More information about WCAG 2.1 principles and guidelines are detailed in the Appendix.

To support the use of the guidelines, WCAG-EM5 evaluation methodology was implemented for experts to follow a common heuristic approach for evaluating the conformance of websites to WCAG. WCAG-EM has been designed with a heuristic evaluation approach in mind and based on previous methodologies such as the Unified Web Evaluation Methodology (UWEM)6. Heuristic evaluations using WCAG guidelines are difficult to apply without prior expertise and could produce false positives [Brajnik et al, 2010; Brajnik et al, 2012]. As the literature indicates, several factors influence the uncertainty in heuristic evaluations: (1) the vagueness of the evaluation process may cause evaluators to focus on aspects that are not necessarily related to the criterion to be evaluated; (2) the individual decision for success or error is personal; and (3) training is important to address the expertise gap related to WCAG.

Heuristic evaluations with WCAG require expertise for manual evaluation but are usually supported by automatic evaluation tools. When deciding to choose the tools to use, we need to consider the weaknesses of automated accessibility tools [Duran, 2017]. However, since automated accessibility tools are considered to have weaknesses, some researchers in the field suggest that evaluators should use a combination of tools [Vigo et al., 2017] since automatic tools only cover a few of the criteria, and need manual evaluation to be double-checked (e.g., the use of alternative text in a decorative image). That complexity occurs even with the most recent set of guidelines and documentation besides the efforts of introducing 2019 Accessibility Conformance Testing (ACT) rules7 [Alajarmeh, 2022].

Teaching and learning how to apply WCAG is complex, and the challenge is even greater if students do not have the technical skills expected for WCAG's applicability [Restrepo et al., 2012]. From the teaching perspective, the basis to determine conformance with WCAG 2.X success criteria is not straightforward. In fact, in WCAG 2.X, a single accessibility barrier can be covered by more than one success criterion at different levels. For example, colour contrast is covered by two success criteria. WCAG evaluations should consider the combination of automatic tools with heuristic approaches [Iniesto, 2020] and when possible, consider a collaborative and complementary approach for accuracy [Brajnik et al., 2016; Brajnik et al., 2011].

[5] WCAG-EM, https://www.w3.org/TR/WCAG-EM/
[6] UWEM, http://www.wabcluster.org/uwem1_2/
[7] ACT Overview, https://www.w3.org/WAI/standards-guidelines/act/

3 Methodology

The context of this study is the *"Usability and Accessibility"*[8] course which is part of the Computer Engineering degree at UNED. Third-year Computer Science undergraduates are introduced to the guidelines for designing and implementing accessible graphical user interfaces, developing accessible webpages and the use of automatic and manual tools in methodologies for assessing Web accessibility (i.e., the use of W3C standards). An adaptation of the WCAG-EM protocol with WCAG 2.X and using a combination of automatic tools was used in this study to evaluate an online resource and to understand students' experiences and perceptions while evaluating WCAG and using automatic tools.

3.1 Aims, research questions and sample

If students who have been trained to use the tools still find them difficult to use, then the methods and tools for teaching accessibility need to be updated. The objective of this research was to evaluate how accurate and easy it is to understand and use WCAG by students as well as their perceptions of usefulness to evaluate accessibility barriers using automatic tools. Having empirical evidence from these experiences will help to improve protocols and training by identifying which WCAG criteria are more complicated to evaluate, and which features of automatic tools can better support accessibility evaluation. The research questions addressed in this research were:

1. **RQ1**. What is the level of agreement of trained students in WCAG evaluation?
2. **RQ2**. What are the perceptions of trained students using WCAG automatic evaluation?

The course has two assignments: the first one is an assignment to understand the multiple accessibility barriers for users using the Web, and the second one is an in-depth study of WCAG guidelines including an accessibility evaluation. The learning objective of the second assignment is to evaluate the accessibility level of a website, both automatically and manually (i.e., applying heuristic evaluations) using WCAG. The complexity of teaching WCAG and the new appearance of evaluating methods and automatic tools increased the need for updating the assignment. The second assignment was fully redesigned incorporating the accessibility evaluation of a MOOC. The MOOC selected is named *"Accessible digital materials"*[9] and it is delivered by UNED Abierta platform. This MOOC is devoted to developing skills to produce accessible resources and the identification of accessibility barriers [Rodríguez-Ascaso and Letón-Molina, 2018], 2016]. Therefore, our students evaluated the accessibility of the MOOC and its platform while participating in an open educational experience with similar content to the course and having social interaction with other students in the forums

[8] Usabilidad y accesibilidad, http://portal.uned.es/portal/page?_pageid=93,69881628&_dad=portal&_schema=PORTAL&idAsignatura=71023105

[9] Accessible Digital Materials, https://iedra.uned.es/courses/course-v1:UNED-ONCE+MatDigAcc_005+2021/d6f6d8e81a624a9f8da4a4dd9c068ca2

[Rodrigo et al., 2020]. The new version of the assignment was first included in 2017-2018 at the time when WCAG 2.1 were released.

The sample in this study included students in academic courses 2017-2018 and 2018-2019 with 52 and 33 students enrolled in the courses (89% - 87% male and 92% - 93% Spanish, respectively). From those, 37 and 26 (n=63) completed the assignment. Students, at the time of registering, consented to the use of anonymised data from their educational interactions for research purposes.

3.2 Methods

Following a procedure for MOOC accessibility evaluation [Iniesto and Rodrigo, 2016], the evaluation on the MOOC platform and educational resources (as web-based) was sampled including:

1. The homepage of UNED Abierta (Open edX[10] based).
2. The registration and authentication page.
3. The MOOC homepage.
4. A page of the MOOC including a video.
5. A page of the MOOC including a test or a quiz (webform).

Students used an adapted checklist [Iniesto, 2020] with the following characteristics which apply to each WCAG criterion:

- **What to test for**: information about the criterion being evaluated.
- **Testing method**: information to help on how to test the criterion.
- **Comments**: space for the student to add open text.

For the evaluation of each criterion the following rating method was applied:

- **NA (Not achieved):** The feature to test is missing.
- **PA (Partially achieved):** The feature to test is available but not integrated.
- If the criterion is not applicable, ***"Not Applicable"*** is added to the comments.
- **LA (Largely achieved):** The feature to test is available and partially integrated.
- **FA (Fully achieved):** The feature to test is available and fully integrated.

The use of *"What to test"* for and "*Testing method"* is based on the accessibility evaluation template by the Inclusive Design Research Centre (IDRC) [Treviranus et al., 2019]. That checklist focuses on success criteria that are mandated by the Accessibility for Ontarians with Disabilities Act (AODA)[11] including WCAG levels A and AA (lowest and mid-range). The EU Web and Mobile Accessibility Directive[12] recommends a level AA of accomplishment for websites. The checklist has been adapted for MOOCs and extended to level AAA (highest), including Accessible Rich Internet Application

[10] OpenedX, https://openedx.org/
[11] AODA, https://accessontario.com/aoda/
[12] EU Web and Mobile Accessibility Directive, https://ec.europa.eu/digital-single-market/en/web-accessibility

(ARIA)[13] indications when possible. The four evaluation criteria are taken from the OpenUpEd quality label benchmark [Rosewell and Jansen, 2014]. Both aspects are summarised in Figure 1. (Note: Further information about WCAG guidelines and their success criteria are available in the Appendix).

Success Criterion	Level	What to test for	Testing Method	NA	PA	LA	FA
2.1.1: Keyboard	A	Are all controls operable with a keyboard?	Navigate to the entire page using only "TAB" and "Shift+TAB" keys to check that all interactive elements receive focus and can be operated with a keyboard	☐	☐	☐	☐
Comments:							
Success Criterion	**Level**	**What to test for**	**Testing Method**	**NA**	**PA**	**LA**	**FA**
3.1.1: Language Page	A	Is page language defined programmatically?	Inspect the HTML source code to see if the language is defined programmatically at the very top of the page at the beginning HTML tag (e.g. <html lang="en">)	☐	☐	☐	☐
Comments:							

Figure 1: Success criteria template example

Students first mark the values associated with WCAG which are collected automatically on the checklist sheet and then perform the required manual evaluation which allows them to reflect on the advantages and disadvantages of both types of evaluation and the points that are difficult to assess with manual or heuristic evaluations. For the automatic evaluation, two tools were used, TAW[14] and WAVE[15] (with the added benefit of having a plugin extension for Chrome or Firefox browsers). Complementary tools included NVDA[16], VoiceOver[17], ContrastChecker[18] and HTML validator[19] along with several plugins for Chrome and Firefox Web browsers.

WCAG exercise	
Task 1. (RQ1)	A single document will be filled out for the evaluation (checklist), for which labels will be used to determine on which page each type of error (problem) or warning. has been detected and not verified. The student must navigate through the MOOC sample to fill in as many WCAG criteria as possible.

[13] WAI ARIA, https://www.w3.org/TR/wai-aria-1.1/
[14] TAW, https://www.tawdis.net/?lang=en
[15] WAVE, https://wave.webaim.org/
[16] NVDA, https://www.nvaccess.org/
[17] VoiceOver, https://webaim.org/articles/voiceover/
[18] ContrastChecker, https://contrastchecker.com/
[19] HTML checker, https://validator.w3.org/nu/

WCAG exercise	
Task 2. (RQ2)	Answer the following questions: 1. Reflect on the advantages and disadvantages of using automatic tools. 2. Detect which criteria are not correctly evaluated with automatic tools. 3. Tell what false positives you have detected with the automatic tools. 4. Comment on which criteria are difficult to evaluate with manual evaluation.

Table 1: WCAG assignment instructions

A mixed-methods approach has been used for this research [Myers and Powers, 2017]. As summarised in Table 1, task 1 included the quantitative data from the checklist to answer RQ1. While task 2 involved the qualitative data of the questions included in the script to support RQ2. For the analysis of RQ1 inter-rater reliability using Fleiss' Kappa fixed-marginal multi-rater was used since students were assigned a set number of cases to each category [Landis and Koch, 1977]. While for RQ2 the analysis method was inductive thematic analysis guided by the research questions [Gavin, 2018]. Names from students have been anonymised using ST (from "student") and a number.

4 Results

Results are discussed for each of the RQs to understand the level of agreement using WCAG guidelines and student perception of the usefulness of automatic tools for heuristic evaluation.

4.1 The level of agreement

The results of students' evaluation have been divided by principles and criteria, in each of the following figures, the diverging stacked bar charts include not achieved, partially achieved, not applicable, largely achieved and fully achieved ratings (from left to right). Kappa (K) Interpretation is - 0.0-0.20 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement and 0.81-1.0 almost perfect agreement – [Landis and Koch, 1977]. Two Kappa values have been calculated, K1 includes the five rating options, while K2 is reduced to three options (disagreement, neutral and agreement). Moderate agreement values are presented with a * while substantial and perfect agreements are shown with a + to facilitate the visibility of the results.

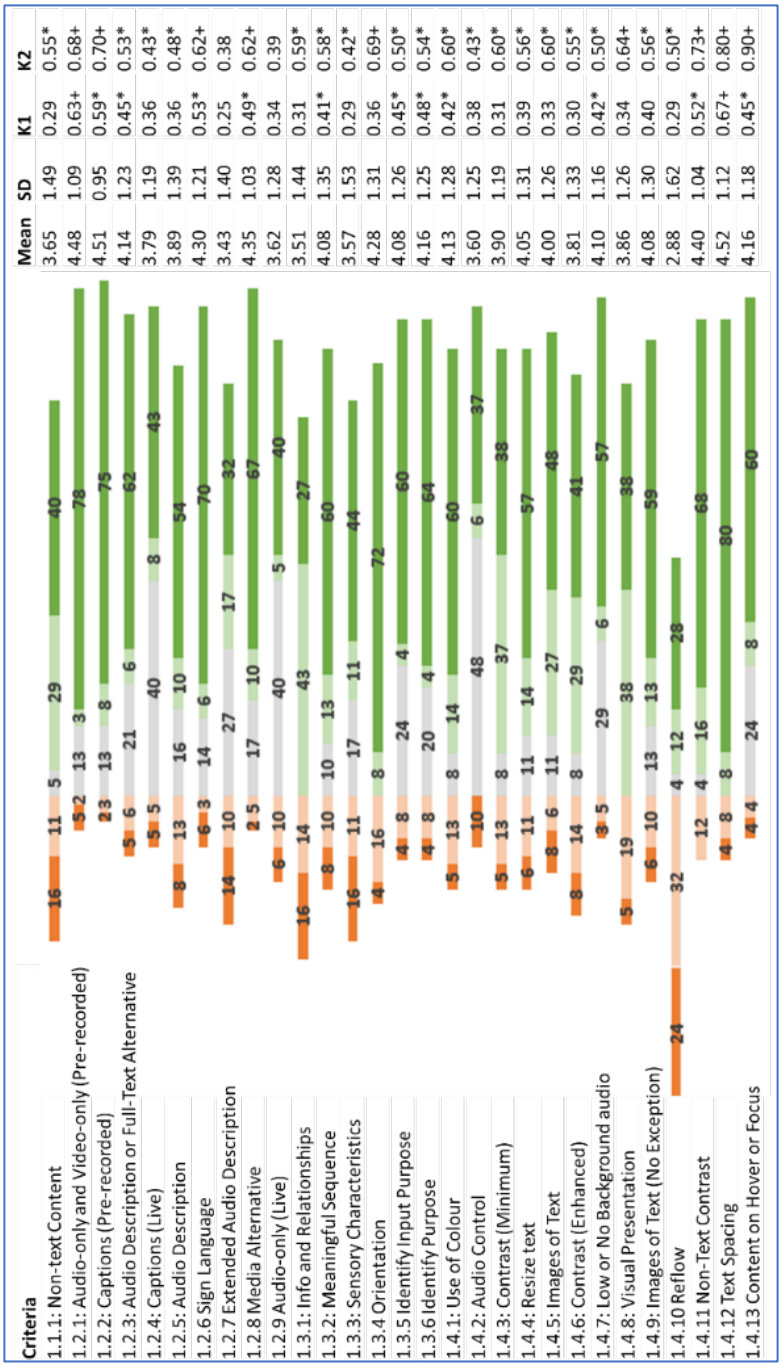

Figure 2: Perceivable principle - evaluation results

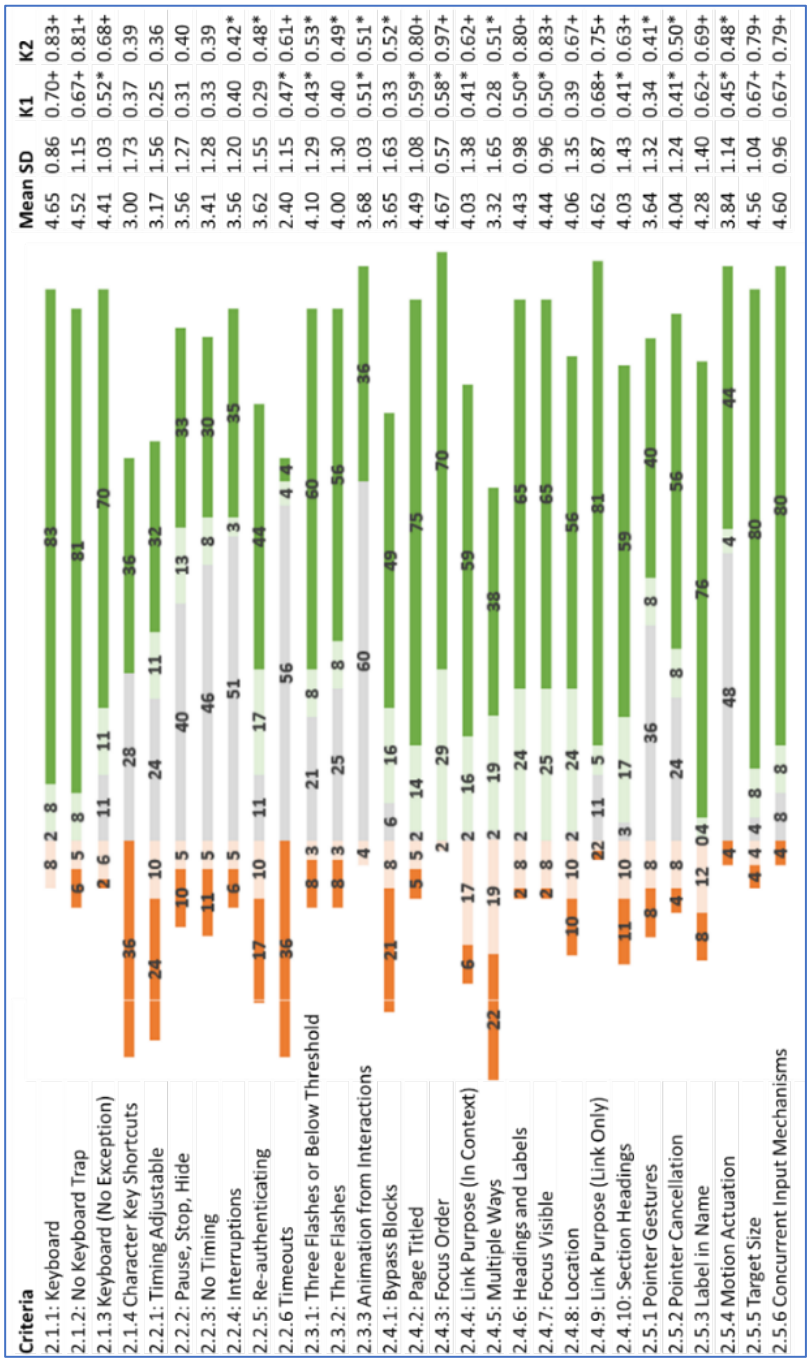

Figure 3: Operable principle - evaluation results

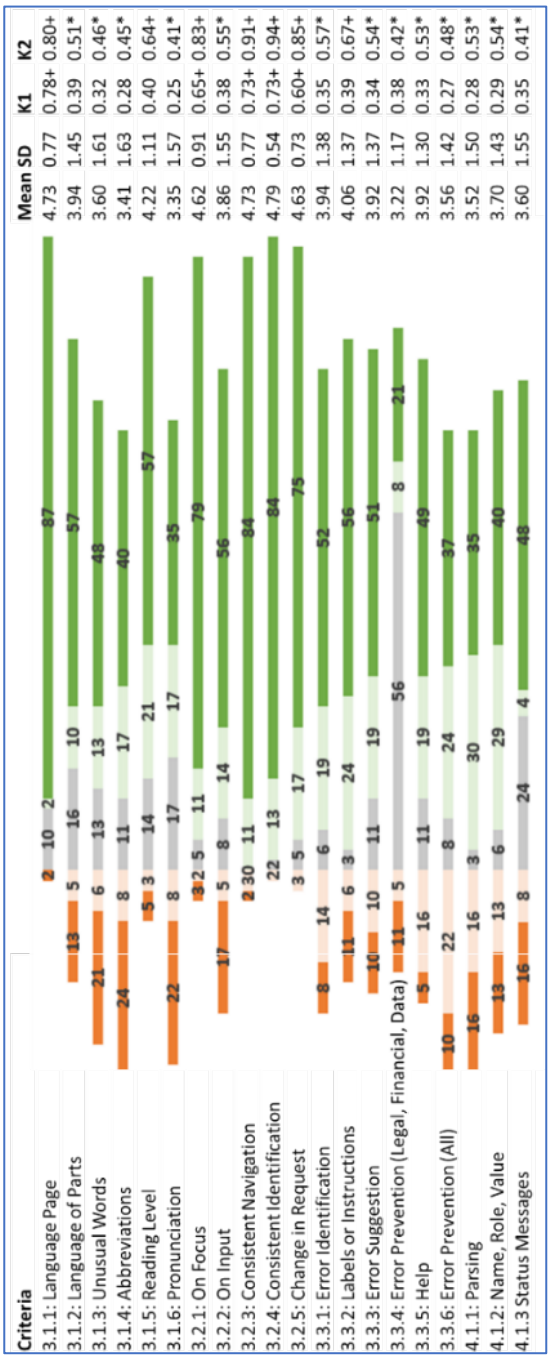

Figure 4: Understandable and robust principles - evaluation results

In the case of the perceivable principle (Figure 2), it can be observed that criteria with better evaluations are when video or podcast content is in an alternate format, such as (1) transcript (1.2.1), videos have accurate captions (1.2.2) and transcriptions (1.2.8), and (2) there exists sign language interpretation, which is not a common aspect in MOOCs (1.2.6). Other good evaluations include when the programmatic order of the content coincides with the visual order and remains oriented if we access from several devices (1.3.4), and there is sufficient text spacing for correct reading (1.4.12). While those with worse evaluations include unlabelled form controls (1.1.1 and 1.3.1) and where content is not responsive when the browser zoom is used to scale content (1.4.10).

Concerning the operable principle (Figure 3), the following are positively evaluated criteria: navigation with the use of the keyboard (2.1.1, 2.1.2 and 2.1.3), having a descriptive title (2.4.2), the focus order (2.4.3), the use of links (2.4.9), labels containing the displayed text (2.5.3), the suitable size of the areas that the pointer has to access, and content does not restrict the input functionalities (2.5.5 and 2.5.6). While alternatives for keyboard shortcuts (2.1.4), time adjustments (2.2.1 and 2.2.6) and lack of different ways to access the content (2.4.5) got the major number of negative evaluations.

Finally, regarding the understandable and robust principles (Figure 4) positive evaluations include: the pages have their language defined (3.3.1); focus works correctly (3.2.1); navigation and identification are consistent (3.2.3 and 3.2.4); and components that have the same functionality within the sample are consistently identified (3.2.5). On the other hand, most errors are in identifying a mechanism available to specific definitions of words, abbreviations, and pronunciations (3.1.3, 3.1.4 and 3.1.6), error prevention in tests and inconsistent use of markup language (3.3.6, 4.1.1).

To answer RQ1, Fleiss Kappa values were computed in both K1 using the five rating options in the questions, and K2 which reduced the evaluation to three options (disagreement, neutral and agreement). For the 72 criteria checkpoints using K1 scores were: 42 fair agreements; 23 moderate agreements; and 13 substantial and perfect agreements. In addition, using K2 scores were: six fair agreements; 41 moderate agreements; and 31 substantial and perfect agreements. These results indicate that while reasonable agreement for some items was achieved, for other items the responses were more variable. The lower levels of agreement can be interpreted as different interpretations of evaluation tasks (see Table 2).

Principles		Fair	Moderate	Substantial and perfect
Perceivable	K1	16	11	2
	K2	2	18	9
Operable	K1	11	12	6
	K2	4	10	15
Understandable	K1	12	0	5
	K2	0	10	7
Robust	K1	3	0	0
	K2	0	3	0

Table 2: Agreement across principles by students

In this research question, the focus is on interpreting the results from the perspective that variable ratings represent the variability of interpretation by raters. The relationship between disagreement evaluations and agreement statistics provides a potential prioritisation mechanism to address WCAG improvements. While prioritising the six K2 fair agreement criteria (see Table 3) would help to improve the overall evaluation instructions, it does not mean that it would solve the amount of disagreement in future evaluations. For example, even with fair agreement criterion 1.2.9 does not apply to the MOOC since there is no live audio included, which is correctly indicated by 40% of the students (Figure 3). The same applies to 2.2.2 and 2.2.3 criteria. Furthermore, while agreement is better within understandable and robust principles, there are fewer criteria with a substantial agreement (see Table 2).

Principles	Criteria
Perceivable	1.2.7 Extended Audio Description 1.2.9 Audio-only (Live)
Operable	2.1.4 Character Key Shortcuts 2.2.1: Timing Adjustable 2.2.2: Pause, Stop, Hide 2.2.3: No Timing

Table 3: Criteria with fair agreement at K2 by students

As can be seen in Table 3, with the strengths and limitations of using agreement statistics in mind, these results suggest focusing improvement of the instructions on six criteria where there was fair agreement using the K2 calculation.

4.2 Perceptions of automatic evaluation

Advantages and disadvantages

As students report, automatic evaluation tools have advantages such as the speed of operation and allowing criteria to be reviewed simultaneously and help to certify when criteria are not met. Automatic evaluation tools save time, and they inform the manual review by providing the steps to follow. Therefore, the automatic evaluation can be applied first, but it has certain shortcomings that must be checked and evaluated manually, according to accessibility standards:

> *Automatic tools present a faster evaluation, help to have a first impression of the accessibility of a Web page, objective and timely. They analyse the pages based on the accessibility guidelines. They can be carried out as many times as considered necessary, being able to carry out continuous monitoring if desired. It is possible to perform them on many pages or the entire site.* (ST4)

Regarding disadvantages, the interpretation of the results of the analysis can be complex, and many aspects of accessibility can only be verified through a complementary manual review because the use of automatic tools alone cannot assess that the WCAG criteria are being met. Further, automatic tools have certain limitations such as giving false positives, or not being able to detect some errors that the user must check manually:

> *They have several drawbacks, so they should be used as complimentary evaluations to the manual tests carried out by a Web accessibility expert, and these automatic tests should be used as a first step, not as the only one. It does not provide us with a definitive*

and reliable analysis since it cannot detect all errors as important errors or detect false positives. (ST53)

Automatic tools must be understood as an aid in the evaluation process and not to support a complete and definitive analysis. If they are used incorrectly without clarity about their function or usefulness, they can cause developers to relax into believing that they are creating accessible websites when in fact they are not. Therefore, the intervention of a (human) expert is necessary to complete the evaluation process alongside the application of automatic tools. Table 4 summarises the five key advantages and disadvantages elicited from students on automatic and manual evaluations.

Automatic tools	
Advantages	**Disadvantages**
1. They allow validation according to different standards. 2. They allow downloading the analysis or be sent via email. 3. They allow a more efficient evaluation in time and form. 4. They help evaluate page structures, headings, images, broken links, ARIA form elements and text contrast. 5. They can be used as many times as desired on the same pages.	1. It is not a substitute for review by a Web accessibility expert. 2. Many criteria cannot be evaluated and therefore must be analysed manually. 3. Verification of a set of pages does not ensure the accessibility of an entire website. 4. They can generate false positives. 5. They require complex interpretation and require knowing basic principles of accessibility.
Manual evaluation	
Advantages	**Disadvantages**
1. It is more difficult to generate false positives. 2. They can detect problems that automatic validation does not provide. 3. Manual reviews carried out by experts identify deviations and propose corrections following the accessibility and quality requirements. 4. Data collection techniques by users (surveys, user tests, etc.) that need to be performed or evaluated manually are varied. 5. Reviews provide greater detail (than automatic tools) and reliability regarding the problems detected.	1. Validation is much slower than automatic validation. 2. The reviews will have a greater periodicity than the automatic reviews. 3. Analysis may only be carried out on a limited set of pages thus being less exhaustive. 4. Require the final judgment of the reviewer (expertise is needed). 5. Some aspects are difficult to simulate, and some accessibility bugs may not be detected.

Table 4: Advantages and disadvantages of automatic and manual tools elicited from students

Evaluation failure

Observing the list of points that the students have considered as not being automatically evaluated (see Table 5), most of these require manual evaluation. In other cases, automatic evaluation seems possible if in the future the tool(s) will be able to understand certain semantics of the webpage:

In most automated accessibility tests, there is a tendency for the test to be limited to checking conformance to accessibility standards. Therefore, other factors directly linked to the interaction of the end-user with the content are lost, which in many cases limits a truly satisfactory and effective result. Navigating and operating the contents of a page with a screen reader allows you to detect accessibility problems that may be missed by other types of reviews. (ST43)

Automatic tools evaluate the accessibility criteria that are related to source code and its presentation, those evaluations are based on the evaluator's criteria or require another specific tool such as an HTML validator. Automatic tools do not evaluate criteria related to the information present in images, and in audio or video assets. The same is true if there are subtitles or transcripts present, or if there is background noise in the videos, and there are descriptive and flickering images. The tools cannot validate the order of a page without styles, or if the information is presented only in colour (i.e., most of the perceivable criteria). The tools are also not able to check keyboard navigation and time limits (2.1.X and 2.2.X).

Regarding false positives, non-textual content appears as a problem that must be corrected on all pages (1.1.1). The purpose of the links also appears as an error but analysing all the links on each page shows they are usually well described (3.2.4). Finally, section headers are also presented as an error; instead, pages are well structured with headers and titles (1.3.1). Table 5 lists the criteria that are not correctly evaluated by automatic tools, those which show false positives (in italics) and those identified that are not correctly evaluated and even show false positives (in bold).

Perceivable	Operable	Understandable and Robust
1.1.1: Non-text Content 1.2.1: Audio-only and Video-only (Pre-recorded) 1.2.2: Captions (Pre-recorded) 1.2.3: Audio Description or Full-Text Alternative 1.2.4: Captions (Live) 1.2.5: Audio Description 1.2.6 Sign Language 1.2.7 Extended Audio Description 1.2.8 Media Alternative 1.2.9 Audio-only (Live) **1.3.1: Info and Relationships** 1.3.2: Meaningful Sequence 1.3.3: Sensory Characteristics 1.3.4: Orientation 1.3.5: Identify Input Purpose 1.3.6: Identify Purpose 1.4.1: Use of Colour 1.4.2: Audio Control 1.4.3: Contrast (Minimum) **1.4.4: Resize text** **1.4.5: Images of Text** **1.4.6: Contrast (Enhanced)** **1.4.7: Low or No Background audio** **1.4.8: Visual Presentation** **1.4.9: Images of Text (No Exception)** 1.4.10: Reflow 1.4.13: Content on Hover or Focus	**2.1.1: Keyboard** **2.1.2: No Keyboard Trap** 2.1.3: Keyboard (No Exception) **2.2.1: Timing Adjustable** **2.2.2: Pause, Stop, Hide** **2.2.3: No Timing** 2.2.4: Interruptions **2.2.5: Re-authenticating** **2.3.1: Three Flashes or Below Threshold** **2.3.2: Three Flashes** **2.4.1: Bypass Blocks** **2.4.2: Page Titled** **2.4.3: Focus Order** **2.4.4: Link Purpose (In Context)** **2.4.5: Multiple Ways** **2.4.6: Headings and Labels** **2.4.7: Focus Visible** 2.4.8: Location 2.4.9: Link Purpose (Link Only) **2.4.10: Section Headings** 2.5.3: Label in Name	*3.1.1: Language Page* **3.1.2: Language of Parts** 3.1.3: Unusual Words 3.1.4: Abbreviations 3.1.5: Reading Level 3.1.6: Pronunciation **3.2.1: On Focus** **3.2.2: On Input** 3.2.3: Consistent Navigation **3.2.4: Consistent Identification** *3.2.5: Change in Request* *3.3.2: Labels or Instructions* *3.3.4: Error Prevention (Legal, Financial, Data)* **4.1.2: Name, Role, Value**

Table 5: Criteria not correctly evaluated by automatic tools

Complex criteria

For those criteria that are complicated to evaluate manually, students point out that reading level (3.1.5) is difficult to assess because it is subjective and different for each user. It is also difficult to assess whether the time limit is adequate (2.2.1 and 2.2.6), as it will depend on the end-user.

From my point of view, the most complicated points of manual evolution are those that are within the perceptible principle since they are the most tedious to analyse and perform the appropriate checks on the code to be inspected. (ST22)

Those criteria in which an exhaustive search of the elements of a page must be carried out, such as seeing if the contrast is correct, looking at all the parts of the page that have different colours in text and background (1.4.3 and 1.4.11) or checking if all the elements include correct labels and descriptions for both their use and state, are difficult to evaluate manually (1.3.1 and 4.1.2). Table 6 lists criteria identified as being difficult to evaluate manually by students.

Principles	Criteria
Perceivable	1.3.1: Info and Relationships 1.3.2: Meaningful Sequence 1.3.3: Sensory Characteristics 1.4.3: Contrast (Minimum) 1.4.8: Visual Presentation 1.4.11: Non-text Contrast 1.4.12: Text Spacing 1.4.13: Content on Hover or Focus
Operable	2.1.1: Keyboard 2.1.2: No Keyboard Trap 2.2.1: Timing Adjustable 2.2.2: Pause, Stop, Hide 2.2.6: Timeouts 2.4.1: Bypass Blocks 2.4.10: Section Headings 2.5.5: Target Size
Understandable	3.1.1: Language Page 3.1.3: Unusual Words 3.1.5: Reading Level 3.1.6: Pronunciation
Robust	4.1.1: Parsing 4.1.2: Name, Role, Value 4.1.3: Status Messages

Table 6: Criteria difficult to evaluate manually by students

To answer RQ2, as can be seen in Table 5, 27 out of 29 criteria in the perceivable principle (93%) are indicated as not being identified by automatic tools, the same applies to 21 out of 29 criteria in the operable principle and 13 out of 17 in the understandable principle (72%) and 1 out of 3 for the robust principle (33%). In total 62 (86%) criteria are marked as not being addressed by automatic tools. It is shown, as well, an overlap between those incorrectly evaluated and indicating false positives (i.e., 28 criteria). Finally, 25 criteria (34%) are tagged as difficult to evaluate manually. That includes criteria marked as not correctly evaluated with automatic tools, giving false positives and difficult to evaluate manually (i.e., 1.3.1, 2.1.1, 2.2.1 and 4.1.2).

5 Discussion and conclusions

WCAG evaluations are complex and difficult, even with current automatic tools, and better evaluation protocols and training opportunities are needed. This research is an effort to better integrate accessibility evaluation into the Computer Science curriculum by considering the pedagogy of redesigning an assignment to improve students' engagement [Lewthwaite and Sloan, 2016]. For that purpose, we have included a MOOC about accessibility which allows students to interact and try different educational resources than those included in the course [Putnam et al., 2016].

As shown in this research, despite the benefits of WCAG, these guidelines remain difficult for students to understand, learn, and apply. Some of the errors listed above relate to unlabelled form controls (1.1.1 and 1.3.1); time adjustments (2.2.1 and 2.2.6); the recognition of different ways to access the content in the course platforms (2.4.5); and error prevention (3.3.6) in tests and quizzes. Also, discrepancies arise in identifying specific definitions of words, abbreviations, pronunciations (3.1.3, 3.1.4 and 3.1.6), and in general an incorrect use of markup language (4.1.1). These results are aligned with research on MOOC accessibility evaluation [Ingavélez-Guerra et al., 2020] which declares errors that exist usually in the use of form-based content such as in tests and quizzes. Fortunately, in the case of UNED Abierta MOOC, we report positive results in the design of video recordings and the variety of different formats provided.

Regarding the RQs explored in this study, results from quantitative data show low levels of agreement with 42 fair agreements and 23 moderate agreements for K1, and six fair agreements and 41 moderate agreements for K2, indicating 90% and 65% of criteria for each Kappa show a significant level of disagreement (i.e., below 80%). Previous research included expert and nonexpert judges indicating that expertise matters [Brajnik et al., 2011]. Even though the effect of expertise varies depending on the metric used to measure quality, the level of expertise is an important factor in the quality of the evaluation. When pages are evaluated with nonexperts, there exists a drop in validity and reliability. The same authors in a different study indicated that among 22 experts and 27 novice evaluators, perfect agreement was rare in both groups [Brajnik et al., 2018], confirming WCAG conformance cannot be tested only by human inspection to a level where it is believed that perfect agreement can be obtained.

Results from the study reported in this paper show different interpretations of the evaluation criteria. One solution on how to address these differences is to encourage team-based evaluation which facilitates dialogue and discussion for a better understanding of the criteria and a more accurate result, also reported in the literature [Brajnik et al., 2016]. All the areas where raters disagree are potential opportunities for WCAG improvement, mostly within the perceivable and operable principles, but one key aspect is that several of those criteria with a fair agreement including 1.2.9 and 2.2.2 and 2.2.3 do not apply to the context of the evaluation. Therefore, the first reasonable step is to provide comprehensive and clear instructions about the applicability (or not) of some criteria (*"What to test for"* and *"Testing method"* information).

It is not surprising that five out of six criteria with fair agreement at K2 are flagged as not evaluated correctly by automatic tools (i.e., 1.2.7, 1.2.9, 2.2.1, 2.2.2 and 2.2.3). In total 62 (86%) criteria are marked as not being correctly addressed with an overlap of those showing false positives, while 25 criteria (34%) are indicated as difficult to

evaluate manually. Previous research has indicated that automatic tools only analyse half of the success criteria and only four out of 10 are caught at the further risk of generating false positives [Vigo et al., 2013]. It is not only about improving the instructions but the use of a combination of tools that are clearly stated for their use for different criteria. For example, there are tools specifically designed for checking the colour contrast or screen readers. In general, for *"all in one"* tools, it is preferred those which support their use through the browser session (i.e., via plugins) such as WAVE offer a better choice than installed tools like TAW. Identifying the limitations and strengths of automatic tools allows a better selection to be included in evaluation protocols [Treviranus et al., 2019].

A limitation of this research is that students do not have enough experience in the assessment of accessibility and, therefore, cannot be considered experts, even though they are trained for one month in the use of WCAG heuristic evaluation and in the use of automatic tools. Finally, we believe that future research for the evaluation of educational resources cannot be limited to Web accessibility [Iniesto, 2020] alone: if we are considering a learning environment, we need to include user experience, quality, and learning design criteria alongside accessibility. Our research shows there is a need for improving the evaluation protocols and changes should be considered in the publication of WCAG 3.0. This future version should incorporate content from the user agent accessibility guidelines (UAAG) and authoring tool accessibility guidelines (ATAG) to develop a holistic understanding of the accessibility requirements, leading to the adoption of a broader approach incorporating *"Guidelines, Outcomes, Methods and Tests"*.

Acknowledgements

We are grateful for the time and dedication of the participants in this study. We want to thank Alejandro Rodríguez-Ascaso and Emilio Letón-Molina for the access to the MOOC "Accessible digital materials" and Shailey Minocha for proofreading the article.

References

[Alajarmeh, 2022] Alajarmeh N. The extent of mobile accessibility coverage in WCAG 2.1: Sufficiency of success criteria and appropriateness of relevant conformance levels pertaining to accessibility problems encountered by users who are visually impaired. Universal Access in the Information Society. 2022 Jun;21(2):507-32.

[Alonso et al., 2010] Alonso F, Fuertes JL, González ÁL, Martínez L. Using collaborative learning to teach WCAG 2.0. International Conference on Computers for Handicapped Persons 2010 Jul 14 (pp. 400-403). Springer, Berlin, Heidelberg.

[Baker et al., 2020] Baker CM, El-Glaly YN, Shinohara K. A systematic analysis of accessibility in computing education research. InProceedings of the 51st ACM Technical Symposium on Computer Science Education 2020 Feb 26 (pp. 107-113).

[Bohman, 2012] Bohman PR. Teaching accessibility and design-for-all in the information and communication technology curriculum: Three case studies of universities in the United States, England, and Austria. Utah State University; 2012.

[Brajnik et al., 2016] Brajnik G, Vigo M, Yesilada Y, Harper S. Group vs individual web accessibility evaluations: effects with novice evaluators. Interacting with Computers. 2016 Nov 19;28(6):843-61.

[Brajnik et al., 2010] Brajnik G, Yesilada Y, Harper S. Testability and validity of WCAG 2.0: the expertise effect. Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility 2010 Oct 25 (pp. 43-50).

[Brajnik et al., 2011] Brajnik G, Yesilada Y, Harper S. The expertise effect on web accessibility evaluation methods. Human–Computer Interaction. 2011 Aug 30;26(3):246-83.

[Brajnik et al., 2012] Brajnik G, Yesilada Y, Harper S. Is accessibility conformance an elusive property? A study of validity and reliability of WCAG 2.0. ACM Transactions on Accessible Computing (TACCESS). 2012 Mar 30;4(2):1-28.

[Duran, 2017] Duran M. What we found when we tested tools on the world's least-accessible webpage. UK Government. 2017. https://accessibility.blog.gov.uk/2017/02/24/what-we-found-when-we-tested-tools-on-the-worlds-least-accessible-webpage/

[Gavin, 2008] Gavin H. Understanding research methods and statistics in psychology. Sage; 2008 Mar 6.

[Gay et al., 2017] Gay G, Djafarova N, Zefi L. Teaching accessibility to the masses. Proceedings of the 14th International Web for All Conference 2017 Apr 2 (pp. 1-8).

[Iniesto, 2020] Iniesto F. An investigation into the accessibility of massive open online courses (MOOCs). Doctoral dissertation. Open University (United Kingdom); 2020.

[Iniesto and Rodrigo, 2016] Iniesto F, Rodrigo C. A preliminary study for developing accessible MOOC services. Journal of accessibility and design for all. 2016 Nov 30;6(2):126-50.

[Ingavélez-Guerra et al., 2020] Ingavélez-Guerra P, Robles-Bykbaev V, Teixeira A, Otón-Tortosa S, Hilera JR. Accessibility Challenges in OER and MOOC: MLR Analysis Considering the Pandemic Years. Sustainability. 2022 Mar 12;14(6):3340.

[Kawas et al., 2019] Kawas S, Vonessen L, Ko AJ. Teaching accessibility: A design exploration of faculty professional development at scale. InProceedings of the 50th ACM Technical Symposium on Computer Science Education 2019 Feb 22 (pp. 983-989).

[Kumar et al., 2021] Kumar S, Shree DV J, Biswas P. Comparing ten WCAG tools for accessibility evaluation of websites. Technology and Disability. 2021 Jan 1;33(3):163-85.

[Landis and Koch, 1977] Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977 Jun 1:363-74.

[Lewthwaite et al., 2020] Lewthwaite S, Coverdale A, Butler-Rees A. Teaching accessibility in computer science and related disciplines: a systematic literature review and narrative synthesis protocol. Social Science Protocols. 2020 May 16;3:1-1.

[Lewthwaite and Sloan, 2016] Lewthwaite S, Sloan D. Exploring pedagogical culture for accessibility in Computing Science. Retrieved from. 2016;10(2899475.2899490).

[Myers and Powers, 2017] Myers KK, Powers SR. Mixed methods. The International Encyclopedia of Organizational Communication. 2017 Feb 21:1-1.

[Petrie and Bevan, 2009] Petrie H, Bevan N. The evaluation of accessibility, usability, and user experience. The universal access handbook. 2009;1:1-6.

[Petrie et al., 2015] Petrie H, Savva A, Power C. Towards a unified definition of web accessibility. Proceedings of the 12th International Web for All Conference 2015 May 18 (pp. 1-13).

[Putnam et al., 2016] Putnam C, Dahman M, Rose E, Cheng J, Bradford G. Best practices for teaching accessibility in university classrooms: cultivating awareness, understanding, and appreciation for diverse users. ACM Transactions on Accessible Computing (TACCESS). 2016 Mar 18;8(4):1-26.

[Restrepo et al., 2012] Restrepo EG, Benavidez C, Gutiérrez H. The challenge of teaching to create accessible learning objects to higher education lecturers. Procedia Computer Science. 2012 Jan 1;14:371-81.

[Rodrigo et al., 2020] Rodrigo C, Iniesto F, García-Serrano A. Reflections on Instructional Design Guidelines From the MOOCification of Distance Education: A Case Study of a Course on Design for All. UXD and UCD Approaches for Accessible Education 2020 (pp. 21-37). IGI Global.

[Rodríguez-Ascaso and León-Molina, 2018] Rodríguez-Ascaso A, León-Molina E. Materiales digitales accesibles. 2018. http://e-spacio.uned.es/fez/view/bibliuned:EditorialUNED-aa-EDU-Arodriguez-0003

[Rosewell and Jansen, 2014] Rosewell J, Jansen D. The OpenupEd quality label: benchmarks for MOOCs. INNOQUAL: The International Journal for Innovation and Quality in Learning. 2014;2(3):88-100.

[Sauer et al., 2020] Sauer J, Sonderegger A, Schmutz S. Usability, user experience and accessibility: towards an integrative model. Ergonomics. 2020 Oct 2;63(10):1207-20.

[Seale et al., 2019] Seale, J., Burgstahler, S. and Fisseler, B., 2019. Tackling the Inaccessibility of Websites in Postsecondary Education. Web accessibility: A foundation for research, pp.263-279.

[Shinohara et al., 2018] Shinohara K, Kawas S, Ko AJ, Ladner RE. Who teaches accessibility? A survey of US computing faculty. Proceedings of the 49th ACM Technical Symposium on Computer Science Education 2018 Feb 21 (pp. 197-202).

[Treviranus et al., 2019] Treviranus J, Richards J, Clark C. Inclusively Designed Authoring Tools. InWeb Accessibility 2019 (pp. 357-372). Springer, London.

[UNESCO , 2020] United Nations Educational, Scientific and Cultural Organization (UNESCO). Global education monitoring report 2020: Inclusion and education: All means all. 92310038. 2020 Feb 4.

[Vigo et al., 2013] Vigo M, Brown J, Conway V. Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests. Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility 2013 May 13 (pp. 1-10).

[Waller et al., 2009] Waller A, Hanson VL, Sloan D. Including accessibility within and beyond undergraduate computing courses. InProceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility 2009 Oct 25 (pp. 155-162).

[Weichbroth, 2020] Weichbroth P. Usability of mobile applications: a systematic literature study. IEEE Access. 2020 Mar 19;8:55563-77.

Appendix

Principle	Guideline	Criteria	Total
Perceivable: The site must provide text alternatives for non-text content, alternatives for time-based media, layout alternatives for related or sequential content, and generally make sure all content is easy to see and hear	1.1 Text Alternatives	1	29
	1.2 Time-based Media	9	
	1.3 Adaptable	6	
	1.4 Distinguishable	13	
Operable: The site must provide keyboard access, enough time to read and use content, orientation, clear navigation, and organised content. A site must also operate safely without flashing	2.1 Keyboard Accessible	4	29
	2.2 Enough Time	6	
	2.3 Seizures	3	
	2.4 Navigable	10	
	2.5 Input modalities	6	
Understandable: Content must be readable, consistent, and predictable. Instructions must be clear and helpful	3.1 Readable	6	17
	3.2 Predictable	5	
	3.3 Input Assistance	6	
Robust: Content must be compatible with a variety of user agents and assistive technologies.	4.1 Compatible	3	3
Total			78

Accessibility principles and guidelines of WCAG 2.1