Evaluation of the Effectiveness of a Tool to Support Novice Auditors

Christopher Bailey
Teesside University
Accessibility Research Centre
School of Computing
Middlesbrough, TS1 3BA, UK.
+44 (0)1642 384648
c.p.bailey@tees.ac.uk

Dr. Elaine Pearson
Teesside University
Accessibility Research Centre
School of Computing
Middlesbrough, TS1 3BA, UK.
+44 (0)1642 342656
e.pearson@tees.ac.uk

ABSTRACT
The Accessibility Evaluation Assistant (AEA) is a web accessibility knowledge management tool designed specifically to assist novice auditors in conducting an accessibility evaluation. The software incorporates a bespoke structured walkthrough approach designed to guide the auditor through a series of checks based on established accessibility principles with the goal of identifying accessibility barriers. A previous trial examined the effectiveness of the AEA and explored the pedagogical potential of the tool when incorporated into the undergraduate computing curriculum. The results of the evaluations carried out by the novices yielded promising levels of validity and reliability. This paper presents the results of a second experiment designed to test the overall efficacy of the AEA when compared to a WCAG 2.0 conformance review. The results of evaluations produced using both AEA and Conformance Review methods were examined and comparisons made of quality factors such as effectiveness, reliability, efficiency and usefulness. Quantitative and qualitative data from the experiment support continued use of the AEA in an educational context, highlighting the benefits compared to WCAG 2.0 and gives further insight into the complex nature of developing accessibility evaluation skills in novices.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology*. K.3.2 **[Computers and Education]**: Computer and Information Science Education – *Curriculum*. K.4.2 [**Computers and Society**]: Social Issues – *Assistive Technologies for persons with disabilities*.

General Terms
Measurement, Human Factors, Verification.

Keywords
Web Accessibility Evaluation, Web Accessibility Guidelines, Accessibility Education.

1. INTRODUCTION
Digital products and services, including websites, should be as inclusive and accessible to the widest range of users as possible. The accessibility of a website is influenced by its context of use and has been discussed elsewhere [15]. Expertise in accessibility is required to fully appreciate contextual factors. An explanation of how this work supports novices in appreciating contextual factors is provided in [7], [8].

The study of accessibility is part of the required body of knowledge for the ACM curricula of Computer Science (CS2008), Information Systems (IS2010) and Information Technology (IT2008) [2]. Lack of knowledge and understanding among designers and developers and insufficient implementation of techniques to support accessibility contributes to the continuing presence of accessibility barriers on websites [21]. Limited exposure to accessibility during training of I.T. professionals and computing graduates has also been identified as an issue [17]. In many institutions there is little space in the computing curricula for a standalone course in accessibility. If included as an additional syllabus topic it is often optional. Unless a set of accessibility topics is clearly defined there is the risk of repetition or exclusion of some components [14]. Work remains on-going in individual institutions to integrate accessibility topics throughout the undergraduate curriculum [23], [20].

When accessibility is not specifically included in the curriculum, computing students may have limited knowledge and understanding of disability and accessibility issues. An institutional review of practical student projects in the area concluded that students' conducting website accessibility evaluations was an effective means to expose students to the issues and techniques required to create accessible websites [14]. One of the problems encountered when teaching web accessibility evaluation specifically to undergraduate students is the lack of proper educational tools that support students in learning about accessibility barriers [3].

Undergraduate computing students within our Institution design and develop rich internet applications and websites and carry out projects for 'live' clients. Testing and evaluating their work for accessibility can be a significant issue. Students lack knowledge and understanding of accessibility guidelines, have problems interpreting the results of automated evaluation tools, have limited access to disabled end users, may not have access to expert reviewers and have little time to dedicate to accessibility in the wider context of their assignments. Undergraduate and postgraduate students increasingly need skills in accessible design to prepare them for work placement projects and employment. They need to understand accessibility beyond the basics, and to apply it in real-world situations, such as developing specific content and solutions for different audiences.

W4A2012 - Technical, April 16-17, 2012, Lyon, France. Co-Located with the 21st International World Wide Web Conference.

The Accessibility Evaluation Assistant (AEA) [1] is a web accessibility knowledge management tool designed specifically to guide the novice auditor through the process of conducting an accessibility evaluation. It utilises a structured walkthrough method to identify barriers on a website and assists the auditor in the production of an evaluation report. Evaluation reports have a positive educational and motivational aspect on those who do not have expertise in web accessibility [21] so a tool developed to assist a novice auditor is considered to have the potential for strong pedagogical value. The tool has been incorporated into a final year elective module, Accessibility and Adaptive Technologies, for students studying a range of computing degrees (e.g. Computing, Web Design and Development, Creative Digital Media).

A previous trial examined the pedagogical potential of the AEA when incorporated into the undergraduate computing curriculum [20] and a comparison of the results of novice and expert evaluations yielded promising levels of validity and reliability [8]. In order to support continued use of the tool and to establish the benefits of using the AEA with novices when compared to other evaluation methods (e.g. WCAG Conformance Review) further investigation was required. This paper presents the results of a second experiment designed to test the overall efficacy of the AEA when compared to a WCAG 2.0 conformance review. The results of evaluations produced using both the AEA and WCAG Conformance Review methods were examined and comparisons made of quality factors (e.g. effectiveness, reliability, efficiency and usefulness) between the two methods. Analysis of the quantitative and qualitative data suggests the AEA provides a positive educational experience for the student and offers further insight into the complex nature of developing accessibility evaluation skills in novices.

The rest of the paper discusses the rationale behind the continued development and use of the tool, related work, a summary of the AEA functionality and structured walkthrough approach, the evaluation methodology and finally, a discussion of the results.

2. RATIONALE

Recent work has focussed on the impact the level of expertise of the evaluator has on the results of on accessibility evaluation. Brajnik [11] evaluated the validity and reliability of 21 checkpoints taken from WCAG 1.0 and 2.0 with 35 inexperienced evaluators and found that neither of the guideline sets have checkpoints whose reliability is definitely higher than the W3C recommended threshold. The W3C consider checkpoints to be reliably testable if 80% of knowledgeable evaluators would agree on the conclusion. A study which examined the testability of the 25 highest priority level 'A' success criteria using manual evaluation techniques found that only 8 could be considered reliably human testable when the auditors were novices [4]. These finding are supported by Brajnik [12] who evaluated the testability and validity of WCAG 2.0 with both experts and non-experts. The results for non-experts showed that the agreement level was 6% below that of the experts, they produced 42% false positives and missed 49% of the true accessibility problems.

A study with a small sample of novices conducting a WCAG 2.0 conformance review accessibility evaluation of a single Home page with and without the assistance of the Hera-FFX 2 tool found that the accuracy of the results of the checks for approximately 50% of WCAG 2.0 success criteria improved with use of the tool. The authors noted the use of HERA-FFX improved the novices' skills in some aspects of evaluation more than others, but even with the use of the tool the novices mistakes were due to a knowledge gap caused by limited prior exposure to WCAG 2.0 [13]. An evaluation of the Barrier Walkthrough method with experts and non-experts [26] concluded that the auditors' level of expertise is an important factor in the quality of an accessibility evaluation. Expert judges were more effective at finding true accessibility barriers and spent significantly less time conducting their evaluation.

The results of these studies suggest an expertise gap when comparing the results of novice and expert evaluations regardless of the evaluation method adopted and there is a specific requirement for a higher level of understanding when using more advanced tools and methods. There is a need for a tool or method which can support novices in gaining an understanding of accessibility evaluation, assist them in identifying barriers and prepare them for use of established evaluation methods (e.g. WCAG 2.0 Conformance Review and Barrier Walkthrough.

3. RELATED WORK AND TOOLS

A range of methods and support tools have been developed to address the complexity of conducting accessibility evaluations and WCAG Conformance Review in particular. Baguma et al. propose a framework for filtering and presenting web accessibility guidelines according to different contexts of use [6]. The MAGENTA tool [18] was developed as a semi-automatic evaluation tool which checks a website against a specified set of guidelines. The user can carry out an accessibility evaluation from a range of pre-defined guideline sets, and can select which individual guidelines within a set are used.

The Accessibility Guidelines Management Framework [5] is a repository of different sets of guidelines including general web accessibility, as well as those for different application types, end-users and specific user and application type. The framework consists of a web application which allows the user to search for lists of guidelines, create new lists, or edit and update existing lists. OceanAcc integrates an automated evaluation tool with accessibility metrics [19] to provide an application with a semi-automatic evaluation process which simplifies and quickens the evaluation process. User intervention is required to filter false positives and results which are not applicable to the context of the website. The Web Accessibility Evaluation Tool (Waat) [24] allows the auditor to conduct a comprehensive evaluation against WCAG 2.0 checkpoints and tailor the evaluation by impairment or disability type using the ACCESSIBLE harmonised methodology.

The Unified Web Evaluation Methodology is specifically aimed at expert evaluators (UWEM) [22] and can be adopted by organisations to assist interpretation of WCAG 1.0 and 2.0. The documentation contains a range of procedures to validate WCAG checkpoints with applicability criteria, expected results for pass or fail and information if the check is fully automatable.

The Accessibility Example Generator (AEG) tool [3] contains a repository of example accessibility barriers coded in HTML, CSS and JavaScript, based on a list of common failures of WCAG 2.0 Checkpoints in W3C documentation. The failures can be evaluated by students and used by instructors to demonstrate example accessibility barriers. The tool is designed to be used as one of the first steps in teaching accessibility. The authors hypothesise that by presenting examples of bad practice will help students to develop practical skills in accessibility evaluation. One

institutional approach [9] to educate students studying web accessibility courses uses a fictional Contramano website and HERA evaluation tool. An 'incorrect' version of the Contramano website features content designed to fail checkpoints requirements of some WCAG 1.0 guidelines and a 'correct' version is presented with all accessibility problems solved. The 'incorrect' version allows the tutor to demonstrate barriers and allows the students to evaluate the website using a conformance review method. The 'correct' version demonstrates how the barriers can be removed by implementing different design and development techniques. While these approaches clearly have value, one potential limitation is that when conducting evaluations in a 'live' context, a checkpoint failure or accessibility barrier may not be so obvious. In many cases it is not a case of a simple 'pass' or 'fail' - many barriers are subjective, or require interpretation and an appreciation of related contextual factors.

The AEA shares some of the same concepts of existing tools such as filtering recommendations by type of disability or content features. However, much of the existing work caters for the needs of web developers, IT professionals and other accessibility stakeholders, such as project managers – in short, those who could be expected to possess at least a basic level of knowledge about accessibility and related issues. Conformance review against WCAG 2.0 can be too complex for a novice auditor as they may not have a full understanding of all the issues needed to interpret them, the guidelines need to be applied with informed judgement and the results of any tool used to assist the evaluation need to be interpreted correctly. The AEA is designed specifically as tool to support manual, heuristic evaluation. By guiding the novice auditor through the process of an evaluation, it supports the student in developing skills in accessible web design and development before they use more advanced evaluation tools. The process of conducting an accessibility evaluation can be complex; the AEA aims to simplify this process so it can be followed and understood by those without in-depth knowledge of accessibility guidelines. The AEA is complementary to evaluation methods such as Barrier Walkthrough and UWEM by supporting novices in gaining a fundamental knowledge of the principles of accessibility before using more advanced evaluation methodologies and tools.

4. TOOL FUNCTIONALITY AND DESIGN

The AEA [1] contains a database of 48 separate accessibility checks for heuristics based on established accessibility principles taken from a range of guidelines, established evaluation methodologies proposed by accessibility practitioners and the authors' personal experience of identifying barriers when conducting evaluations on a range of website in the private, public and higher education sectors. Upon launching the tool from their browser the auditor is presented with the option of conducting an evaluation in one of three different contexts:

- Check Categories
- User Group
- Site Features

By presenting the relevant guidelines according to a specific context the auditor can carry out an effective audit without the need to go through a full set of checks. This streamlines the evaluation process and eliminates redundancy.

4.1 Check Categories

The Check Categories function supports a comprehensive accessibility evaluation using all 48 checks and is the primary function of the tool. Introducing the auditor to a greater range of checks for potential accessibility barriers supports a more holistic approach to accessibility. The checks are broken down into five categories to make the evaluation process more manageable and groups related checks in a meaningful way for novices:

- **Design Checks:** 11 checks examine the visual appearance of a website. They are concerned with aspects of general presentation, the use of text and colour and the layout and positioning of items. They are generally conducted by a visual examination of the website, e.g. testing the colour contrast of foreground and background colour combinations used on a page.
- **User Checks:** 15 practical checks which require manual human testing and interaction with the website in order to conduct these checks, (e.g. ensuring that navigation elements on a page are accessible using only the keyboard). Although automated tools can help, human intervention is required as some checks are subjective, and require judgement (e.g. ensuring the copy text has an acceptable level of readability and is appropriate for the site's target audience).
- **Structural Checks**: 11 checks concerned with the way content is structured on a webpage and ensuring semantic information about content is delivered to the user (e.g. ensuring HTML Heading elements are used and implemented correctly to structure the content of a page).
- **Technical Checks:** Five checks concerned with coding elements such as validating the HTML and CSS mark-up used to produce a webpage. Technical checks also deal with the metadata elements of a webpage (e.g. specifying a DOCTYPE and HTML language attribute).
- **Global Checks:** Six checks which refer to issues which apply to the entire website (e.g. providing a Site Map) or refer to specific functionality (e.g. providing options for user customisation). Those that refer to specific content and functionality can sometimes be verified by examining a single page (usually the home page) and need only be conducted once.

Grouped by category, the checks are presented to the auditor in a list, along with a brief text summary. Many checks recommended by the tool require the auditor to manually examine the website or webpage being checked and as such are not suitable for a solely automated process. The AEA is not an automated evaluation tool, but does utilise existing resources - primarily the Web Accessibility Toolbar 2.0 [25] - to simplify the process of testing and verification.

4.2 User Group

The Check by User Group function currently allows the auditor to prioritise checks based on the needs of 10 different disability groups (e.g. Screen Reader User, Older Web User). Individual accessibility issues may have a greater impact on one user group when compared to another. By enabling the novice auditor to filter and prioritise the accessibility checks according to the needs of different user groups they become familiar with the common principles of accessible design, identify specific exceptions and learn about the needs of diverse user groups. The AEA provides a function to directly compare prioritised checks for two user

groups. The AEA defines three priority levels for the checks for each User Group; Critical Checks, Important Checks and Minor Checks. Unlike WCAG 1.0 and 2.0 this priority level is not fixed but changes depending on the relative potential impact it could have on that user group. For a full list of User Groups in the AEA and a description of the priority levels see [8].

4.3 Site Features

The Site Features function allows the auditor to filter the checks based on specific elements of a website (Forms, Images, Cascading Style Sheets, Links, Multimedia, Semantic HTML, Tables). This feature was not examined in this experiment. When using a conformance review evaluation, accessibility guidelines for a single element or site feature (e.g. forms) can be spread across two or three different priority levels. This could be confusing for a novice evaluator and makes the evaluation process overly complex. This complexity can be addressed by grouping checks together based on the element or site feature they refer to, presenting the relevant checks to the auditor, and increasing the usability of the checking process.

4.4 The Structured Walkthrough Approach

Having selected checks based on Categories, User Group or Site Feature the second element of the AEA provides a step-by-step walkthrough for each check. The structured walkthrough approach to evaluation is based on the Barrier Walkthrough method [10] but adapts the method with the aim of making it more appropriate for use by novices. The structured walkthrough approach defines checks based on specific heuristics the auditor is evaluating and supports them with guidance and tutorials. Each heuristic is broken into a number of components:

1. The title of the accessibility principle (heuristic).
2. A short summary, in the form of lead text.
3. General description of the checks' importance in terms of the user group(s) affected and the nature of the barrier or problem caused.
4. Description of the method to perform the check, with step-by-step instructions (utilising functions of the Web Accessibility Toolbar (WAT) or other checking tools).
5. Steps to verify and record a result for the check.

Integrating the rationale for each check into the sequence aims to improve the educational aspect of the evaluation method and the definition of an exact procedure for checking and verifying the issue is considered a key feature of the AEA as a tool to support novices. The procedure for checking and verifying may be manual, automatic or a combination of both. Where the check directs the user to an automated check or functionality provided by the WAT, instructions are given for which element or function to use and advice is given on interpreting the results; this is considered to be one of the key elements of the AEA as an expert system. A short video demonstration of an expert evaluator performing the check is also provided. This includes a commentary describing the check procedure, highlighting the accessibility barriers found, and gives advice on interpreting the results of the automated elements of the WAT. Figure 1 illustrates an example of the typical instructions provided for the auditor – in this case for checking image text alternatives.

Image Text Alternatives
Check that all images, and similar elements, have an appropriate text alternative that accurately and concisely describes its content and/or function.

Why this is important
Text alternatives are important for screen reader users as the text is read aloud by the software. If written properly they describe the content or function of an image. They also act as a tooltip as some browsers display the text alternative when the user hovers over the image. A null text alternative of empty quotation marks can be used if the image is purely decorative as this will instruct the screen reader to ignore the image.

How to check this
The Web Accessibility Toolbar can assist with this check but it must be manually verified:
1. Select "Images" > "Remove Images"
2. Images will be removed from the page, and the text alternative will be displayed
3. Where there is no text alternative a warning of "No Alt!" will be displayed

To verify the check:
1. Check and record if all images have a text alternative
2. Check that the text alternative is concise, accurately describes the content of the image and is related to the content of the page
3. If the image acts as a link (or has a function) the text alternative should state the function or page it links to
4. If the image is purely decorative, is used to add visual appeal to the page or is a spacer image, check that it has a null text alternative.

Figure 1: Instructional Information

5. EXPERIMENT AIMS

The aim of this experiment was to compare the usefulness of the AEA for novices against another established evaluation method, specifically WCAG 2.0 conformance review. We also wanted to determine the students' perception of the tool as a scaffold for learning about accessibility evaluations. The establishment and definition of quality attributes means it is possible to compare the effectiveness and viability of each evaluation method depending on the context in which is it to be used. The aim of the experiment was to compare the effectiveness of evaluations produced by novices when using both the AEA and WCAG and to examine the students' experience of both tools from both a pedagogical and usability aspect. Evaluation methods can be compared on a number of attributes such as effectiveness, efficiency and usefulness. To be accurately measured, these quality attributes may be customised to the individual circumstances of an experiment and some of the criteria can be further sub-divided to be made more specific [10]. Effectiveness is the extent to which the method can be used to deliver results with appropriate levels of accuracy and completeness. This can be further divided into validity and reliability.

1. Validity is defined as the extent to which the problems detected during an evaluation are also those that show up during real-world use of the system.

2. Reliability is the extent to which evaluations conducted independently will produce the same result.

Depending on the context of the individual study, validity can be separated into two different measures:

1. Correctness: the percentage of reported problems that are true problems, also referred to as precision.
2. Sensitivity or Thoroughness: the percentage of the true problems that are reported.

For the purposes of this study we have used quantitative data to measure validity and reliability. While we can use the above definition of reliability for our study, we must customise the definition of validity. Validity is measured as the extent to which novices made a decision which matched that of the expert. The novices were mimicking the evaluation process of an expert and making the same subjective decisions as to whether the criteria for the AEA heuristic or WCAG Success Criteria was Met, Not Met, or Partly Met. Further analysis would measure how well novices identified accessibility barriers present on a page.

In this study we must consider the evaluator effect; this occurs during any review process where multiple evaluators may detect different sets of problems when examining the same interface. It affects both novice and experienced evaluators – in short, reliability is highly unlikely to reach 100% [15]. The fundamental cause of the evaluator effort is that the process of an evaluation is a complex cognitive activity that requires evaluators to excise difficult judgements and therefore even with guidance, the interpretation of any subjective element is heavily dependent on individual experience and background.

Other qualities which the experiment would examine are efficiency or viability, which refers the amount of resources (e.g. time, skills, money, and facilities), required to conduct an evaluation. This is related to the level of effectiveness and usefulness required by the evaluation. For example, an evaluation may be considered efficient if it can be conducted in a very short period of time; however the result of this may mean that it will be relatively ineffective in that it may only detect a small number of barriers. Usefulness is the effectiveness and usability of the results produced (with respect to those who assess, fix, or manage the accessibility of a web site). These qualities are measured qualitatively, and rather relate more to the novices' experience of using the methods, rather than effectiveness.

6. EXPERIMENT METHODOLOGY

The evaluation was conducted with 37 final year undergraduate students from a range of computing degrees enrolled on the Accessibility and Adaptive Technologies module. The students were all new to accessibility evaluation. Two websites were used for this study:

- Harley Davidson UK: http://www.harley-davidson.com/en_GB/Content/Pages/home.html
- Sunsail UK: http://www.sunsail.co.uk/

The websites were chosen as they had both previously been identified as containing a number of potential accessibility barriers. An accessibility expert conducted independent evaluations of each Home Page using both the AEA and WCAG evaluation methods to ensure accurate comparisons could be made. Some of the barriers the students were required to check were present on both websites, while others were unique to one or the other. The participants were divided into four groups (Table 1) and the experiment broken down into three separate activities conducted over three weeks. In the first week they would evaluate the Home Page of one website using one evaluation method (e.g. SunSail using the AEA) then in the second week they would evaluate the Home Page of the second site using the alternative method (e.g. Harley Davidson using WCAG 2.0). In the third week they were required to write a short reflective piece on their experience of using the two methods

Table 1: Group Order of Evaluation Method/Website.

Group	Week 1		Week 2	
	Evaluation Method	**Website**	**Evaluation Method**	**Website**
1	AEA	Harley Davidson	WCAG 2.0	Sunsail
2	WCAG 2.0	Harley Davidson	AEA	Sunsail
3	AEA	Sunsail	WCAG 2.0	Harley Davidson
4	WCAG 2.0	Sunsail	AEA	Harley Davidson

In this experiment the participants would evaluate only a limited sub-set of checks from each method. This was to make the experiment more efficient, eliminate redundancy and to reduce potential for confusion. All of the selected checks were relevant to both websites to ensure uniformity in the checking procedure and to allow direct comparisons between methods to be drawn. 15 of the accessibility heuristics from all of the five categories of AEA checks were proportionally represented to ensure a range of checks were covered. The nearest equivalent WCAG 2.0 Success Criteria were investigated and selected for evaluation. This was to ensure the novices were checking for the same issue using each evaluation method which would allow for a more direct comparison to be made. In most cases it was possible to relate each AEA heuristic directly to the requirements of a single WCAG 2.0 Success Criteria. In other cases it was necessary to include multiple WCAG 2.0 Success Criteria to fulfil the evaluation requirements of an AEA heuristic (Table 2).

Table 2: Checks Used During Website Evaluation.

Check No.	AEA Heuristic	Check No.	WCAG 2.0 Success Criteria
1	Images of Text	1	1.4.5 Images of Text
2	Colour Contrast	2	1.4.3 Contrast
3	Use of Colour	3	1.4.1 Use of Colour
4	Text Size	4	1.4.4 Resize text
5	Keyboard Navigation	5a	2.1.1 Keyboard
		5b	2.4.3 Focus Order
		5c	2.4.7 Focus Visible

6	Link Names	6	2.4.4 Link Purpose (In Context)
7	Skip Navigation Link	7	2.4.1 Bypass Blocks
8	Text Alternatives	8	1.1.1 Non-text Content
9	Order of Content	9	1.3.2 Meaningful Sequence
10	Headings and Sub-Headings	10a	2.4.10 Section Headings
		10b	1.3.1 Info and Relationships
11	Form Labels	11	3.3.2 Labels or Instructions
12	Identify Language of Text	12	3.1.1 Language of Page
13	Validate (X)HTML Code	13	4.1.1 Parsing
14	Site Map	14	2.4.5 Multiple Ways
15	Search Function		

The students were provided with a blank evaluation report template for each part of the experiment. They were instructed to carry out an accessibility evaluation of the Home Page using their designated method and given 24 hours to submit their evaluation electronically. For each check the students were required to decide whether the requirements were Met, Not Met or Partly Met. Students were also required to provide some comments or justification to support their decision. This would assist in the analysis of the result by helping to identify false positives, erroneous decisions or cases where the student had misunderstood the requirements for the check.

For the final part of the experiment the students were asked to provide a short piece of text. They were asked to reflect on their experience of using both evaluation methods, consider the advantages and disadvantages, describe any issue or problems they encountered and suggest how each process might be improved.

7. QUANTATIVE RESULTS

The preliminary analysis of the results of the students' evaluations focuses on comparing two quality attributes of methods, reliability, or the extent to which different evaluators achieved the same result, and validity. In this study we focus on correctness as a sub-measure of validity. As both the AEA and WCAG evaluations method were utilised, the results would allow us to compare the effectiveness of the methods when used by novices.

7.1 Reliability

For this experiment we define reliability as the extent to which the novices' decisions matched each other. Table 3 shows the figures broken down by each group so we can examine if results are influenced by one of the independent variables: website or method used.

Table 3: Measure of Reliability by Group

Group	Week 1		Week 2	
	Evaluation Method/ Website	**Reliability**	**Evaluation Method**	**Reliability**
1	AEA - Harley Davidson	73%	WCAG 2.0 - Sunsail	63%
2	WCAG 2.0 - Harley Davidson	67%	AEA - Sunsail	78%
3	AEA - Sunsail	71%	WCAG 2.0 - Harley Davidson	59%
4	WCAG 2.0 - Sunsail	70%	AEA - Harley Davidson	63%

Table 4 shows the overall figure for reliability averaged across each tool and site.

Table 4: Overall Reliability

Method	Week 1		Week 2		Overall	
	AEA	WCAG 2.0	AEA	WCAG 2.0	AEA	WCAG 2.0
Reliability	72%	68.5%	70.5%	61%	71.25%	64.75%

In three of the four conditions in which the AEA was used the reliability was over 70% (in once case 78%), while the overall figure was 71.25% - 6.5% higher than WCAG. Looking at the individual figures for each evaluation scenario, there was only one evaluation condition where the use of the AEA did not lead to higher reliability figure than WCAG. Whilst figures for reliability of the AEA do not meet the level of 80% required by the W3C for knowledgeable evaluators, given that the auditors were accessibility novices and completing their first evaluation, we consider these figures promising. In fact, in all cases, the performance of the novices could be considered acceptable.

7.2 Summary of Validity

For this experiment we define validly as the extent to which the novices' decisions matched that of the expert evaluator. The data in the table shows the extent to which this occurred. Table 5 shows the figures broken down by each group so we can examine if results are influenced by one the independent variables, website or method used.

Table 5: Expert Comparison by Group

Group	Week 1		Week 2	
	Evaluation Method/ Website	**Validity**	**Evaluation Method**	**Validity**
1	AEA - Harley Davidson	73%	WCAG 2.0 - Sunsail	49%
2	WCAG 2.0 - Harley Davidson	59%	AEA - Sunsail	73%
3	AEA - Sunsail	66%	WCAG 2.0 - Harley Davidson	49%
4	WCAG 2.0 - Sunsail	50%	AEA - Harley Davidson	62%

Table 6 shows the overall figure for validity.

Table 6: Overall Expert Comparison

Method	Week 1		Week 2		Overall	
	AEA	WCAG 2.0	AEA	WCAG 2.0	AEA	WCAG 2.0
Validity	69.5%	54.5%	67.5%	49%	68.5%	51.75%

If we examine the four conditions in which the AEA was used we can see that in all cases the overall validity was higher than any condition where WCAG was used. The overall figure for validity of the AEA was 68.5% - 16.75% higher than WCAG.

8. QUALITATIVE ANALYSIS

Qualitative data was gathered with the aim of gaining feedback from the novices about their comparative experiences of using the both AEA and WCAG Conformance review evaluation methods. As this was provided in a piece of reflective text and submitted electronically it was possible to use text analysis software to identify trends in the vocabulary used by the novices in articulating their experience of conducting an evaluation. This was then supplemented by a manual analysis. The data was divided into positive and negative comments on each evaluation method. Table 7 summarises the general categories of positive comments of the AEA.

Table 7: Positive Aspects of AEA

Comment	Frequency
Ease of Use/Simplicity	27
Easy to Understand/Clear Terminology	22
Explanation Guides User	15
Categorisation/Grouping of Checks	14
User Group Checks	11
Helpful Videos	9
Speed of Check Process	6
Instructions Concise	5
Easy to Compare Websites	4
Supports Direct Analysis	3

The novices found the AEA simple and easy use highlighted by their comments regarding the instructions being concise and easy to understand, and the provision of clear instructions for performing checks. The novices also found the grouping of checks in the AEA intuitive and the videos helpful.

Table 8: Common Criticisms of AEA.

Comment	Frequency
Check Explanation Too Brief	6
Videos Too Small	6
Checks Required Individual Judgement	6
Poor Usability of AEA	4
Hard to Judge Met/Not Met	4
Bugs and Errors	4
Problems With WAT	3
Confusing Wording	3
No Advice on Solving Problems	3
Hard to Use	3

Common criticisms of the AEA were related to the design, functionality and usability of the AEA interface; many of which can be resolved in further development of the tool. The students commented that they felt some checks still required a high degree of judgement – however, this reflects the subjective nature of conducting accessibility evaluations in a live context. Table 8 summarises common criticisms students had of WCAG.

Table 8: Common Criticisms of WCAG.

Comment	Frequency
Confusing/Difficult to Understand	38
Complex/Hard to Use	16
Too Long/Detailed	12
More Knowledge/Experience Required	7
Hard to Judge Relevance of Check	6
No Explanation for Performing Check	5
Too Technical	3

The students overwhelmingly found the language used in WCAG confusing and difficult to understand. They found the documentation too complex, overly detailed and difficult to use. The novices appreciated that a higher level of knowledge and experience is required to use WCAG correctly and indicated that explanations on how to perform checks for conformance may improve this. The students did show a balanced viewpoint and recognised the positive aspects of the WCAG, summarised in Table 9.

Table 9: Positive Aspects of WCAG

Comment	Frequency
Detailed Explanations	14
Linked to Regulation/Industry	11
Thorough	9
Real Examples in Documentation	5
Fewer Individual Judgements Required	4
Easy to Navigate Documentation	3

The novices recognised that use of WCAG is linked to use for regulatory purposes and in industry. They found the success criteria explanations detailed and through, if somewhat difficult to understand, and appreciated the inclusion of real world examples in the Techniques for WCAG 2.0 documentation.

The overall consensus was that the students preferred the experience of using the AEA when compared to WCAG. When relating these to the more subjective qualities of evaluation methods we can suggest that the fact the novices explicitly referred to the AEA as being easy and simple to use, provided clear explanation and guidance and the intuitive categorisation of checks supports this. The students' perception appears to be that they found the AEA more efficient and useful than WCAG, with the proviso that this was not measured quantitatively. Many of the criticisms of WCAG such as complexity, difficulty in understanding language and the perception for the requirement for a higher degree of knowledge and understanding to use WCAG correctly are related to the findings of similar studies [4].

Taking these factors into consideration, coupled with the overall trend that both reliability and validity was be higher when using the AEA we can propose that in this study, the AEA was the more effective method for novices to use. We are encouraged, however, that the novices did recognise the authoritative and comprehensive nature of WCAG.

8.1 Student Feedback

Generally, students' comments illustrate their preference for the use of the AEA as an evaluation method and provide some qualitative evidence for the AEA having a higher level of efficiency and usefulness when compared to WCAG 2.0. The primary reason for this, as articulated by the students, is due to issues of comprehension and ease-of-use. Excerpts taken directly from the students' reflective text provide further insight into their experience of both methods.

Comments regarding the AEA:

"I think that the AEA tool is very east to understand, follow and implement the checks. The step-by-step instructions are not only helpful, but informative and made my understand why I should be performing the check".

"Almost straight away it became apparent that the AEA tool was tailored for the novice user by detailing step-by-step instructions with videos. WCAG contained far more detail."

"The advantages of the AEA tool were that the terminology used was simple, the explanations concise and the instructions are easy to understand. The simple steps set it out like a guide and meant that users are not bombarded with lengthy explanations and jargon."

"I found the AEA approach was easier to use as the language used throughout was relatively simple. The AEA informed me of the of the exact way to perform each check and helped me develop my own thoughts on each checkpoint and analyse features I may not otherwise have looked for".

"I found the AEA uncomplicated to use, the videos provided helpful supplement to the written instructions and simplified complicated and complex routines. In my opinion the AEA was more effective since it was less complex and time consuming improving my understanding of accessibility checks more than WCAG 2.0."

"I found the AEA tool significantly easier and more intuitive than the WCAG 2.0 approach. There is a lot less repetition in the AEA, it guides the tester to a relevant section, explains what to look for in an easy to understand manner as well as explaining the purpose of the test."

"The advantages of the AEA tool are speed, simplicity and ease-of-use. Using the AEA definitely felt more effective as it follows the WCAG 2.0 guidelines but streamlines the majority of checks allowing you bypass the technical jargon and bureaucracy in WCAG."

"The AEA tool was more effective because it was easier to and less complicated to use due to how it was set out (each check under headings) and there was also support when performing the checks (step-by-step instructions and small video)."

"It was much easier to understand and perform the check with the AEA tool than with WCAG 2.0 guidance. The main advantage of the AEA is that is provides a very straightforward way to understand the importance of the check and how to meet it using existing tools."

"The advantages using the AEA method are that you are told precisely what you are checking clearly and concisely, making the process faster and simpler. With WCAG 2.0 the sentences are extremely long and I found whilst performing the check I kept forgetting what I was checking for and had to repeatedly read the check."

Comments regarding WCAG:

"It is apparent that WCAG 2.0 involves a more in-depth approach, whereas the AEA tool provides a concise yet compact package that appears to reference WCAG 2.0 strategies but focus on the novice user."

"The WCAG 2.0 Checkpoints were significantly harder to understand, it was text heavy, which, when combined with the use

of language makes it overwhelming to read and difficulty to understand."

"WCAG 2.0 not explain the steps taken to actually perform the checks and because it is so detailed and technical it can be time consuming to use and confusing."

"The WCAG approach provides lots of in-depth information on each checkpoint. This is well organised, which makes it easier to navigate. It can grant you a fuller understanding of accessibility issues and their solutions; however, the content is overwhelming. The language often left me confused and I spent a lot of time re-reading text and searching for a simpler explanation. I spent more time on the WCAG site than I did on the one I was testing."

9. CONCLUSION

The data gained from this experiment supports the observations made in previous work about acceptable levels of validity and reliability produced in evaluations conducted by novices using the structured walkthrough method in the AEA. While firm conclusions cannot be drawn from a single study, using the custom definitions of validity and reliability with the context of this experiment, both figures for validity and reliability were higher with evaluations produced using the AEA. Overall validity of the AEA was 68.5%, compared to 51.75% for WCAG 2.0, while overall reliability was 71.25% for the AEA compared to 64.75% for WCAG 2.0. In terms of validity, further analysis of the quantitative data will verify the overall conclusions drawn on both evaluation methods in this paper and allow accurate figures to be calculated for both precision (percentage of reported problems that are true problems) and sensitivity (percentage of the true problems that are reported).

The results of the experiment do not suggest that using the AEA for a single evaluation will lead to an improved performance when performing a subsequent WCAG 2.0 conformance evaluation. As in this study the AEA heuristics and WCAG Checkpoint Criteria were related, it was possible that this trend would be noted. Qualitative feedback from some students did acknowledge they were able to determine the relationship between them, but this requires further investigation. One possible explanation was that it was due to the novices confusing evaluation methods (e.g. attempting to use the AEA to check and verify WCAG Success Criteria without appropriate guidance. Further development of the AEA would see each heuristic explicitly related to a relevant WCAG 2.0 Success Criteria.

Stronger conclusions could have been drawn from if it would have been possible use control groups, or establish a baseline to measure the students' existing knowledge of accessibility before conducting their first evaluation. Gaining such data is problematic when conducting experiments in a live situation and it would be unethical to provide students with a different learning experience.

The value of conducting further studies to examine the impact that providing tool support specifically for the WCAG 2.0 conformance evaluation would have on the results is acknowledged. However, this would be dependent on finding a suitable tool for use by novices, or providing them with appropriate guidance during the experiments.

It would be useful to measure how knowledge of accessibility improves with multiple uses of the each evaluation method. A more longitudinal study would examine the effectiveness or quality of the students' evaluations with repeated use of the same evaluation method, but again, limited access to participants makes this difficult.

Overall, the study adds further evidence that accessibility evaluation requires expert judgement, although the combined results of this first trial and this experiment indicate that some of this expertise can be incorporated into the AEA. We are encouraged that with continued maintenance and development we can continue to incorporate the AEA into the teaching of accessibility alongside the use of other tools and methods to support students in developing skills in accessibility evaluation.

10. REFERENCES

[1] Accessibility Evaluation Assistant. *Accessibility Research Centre.* http://arc.tees.ac.uk/aea/. Accessed: 27/11/10.

[2] ACM. Current Curricula. Available Online: http://www.acm.org/education/curricula/. Accessed: 14/11/11.

[3] Al-Khalifa, A.S. & Al-Khalifa, H.S. 2011. An educational tool for generating inaccessible page examples based on WCAG 2.0 failures. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (W4A '11). ACM, New York, NY, USA. Article 30, 4 pages. DOI=10.1145/1969289.1969328. http://doi.acm.org/10.1145/1969289.1969328

[4] Alonso, F., Fuertes, J.L., Gonzalez, L.A. and Martinez, L. 2010. On the testability of WCAG 2.0 for beginners. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (W4A '10). ACM, New York, NY, USA. DOI=10.1145/1805986.1806000 http://doi.acm.org/10.1145/1805986.1806000

[5] Arrue, M., Vigo, M., Aizpurua, A. and Abascal, J. 2007. Accessibility Guidelines Management Framework. *In C. Stephanidis (Ed.). Universal Access in HCI, Part III, HCI International 2007* (Beijing, China, July 22-27, 2007). LNCS 4556, 3--10, Springer, 2007.

[6] Baguma, R., Stone, R. G., Lugega, J. T., and van der Weide, T. P. 2009. A framework for filtering web accessibility guidelines. In *Proceedings of the 2009 international Cross-Disciplinary Conference on Web Accessibililty (W4a)* (Madrid, Spain, April 20 - 21, 2009). W4A '09. ACM, New York, NY, 46-49. DOI= http://doi.acm.org/10.1145/1535654.1535663

[7] Bailey, C., and Pearson, E. 2010. An educational tool to support the accessibility evaluation process. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (W4A '10). ACM, New York, NY, USA. DOI=10.1145/1805986.1806003 http://doi.acm.org/10.1145/1805986.1806003

[8] Bailey, C. & Pearson, E. 2011. Development and trial of an educational tool to support the accessibility evaluation process. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (W4A '11). ACM, New York, NY, USA, , Article 2 , 10 pages. DOI=10.1145/1969289.1969293. http://doi.acm.org/10.1145/1969289.1969293

[9] Benavídez, C., Fuertes, J.L., Gutiérrez, E., & Martínez, L. 2006. Teaching Web Accessibility with "Contramano" and Hera. In: Miesenberger, K., Klaus, J., Zagler, W.L.,

Karshmer, A.I. (eds.) ICCHP 2006. LNCS, Vol. 4061, pp. 341–348. Springer, Heidelberg.

[10] Brajnik, G. 2008. A comparative test of web accessibility evaluation methods. In *Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada, October 13 - 15, 2008). Assets '08. ACM, New York, NY, 113-120. DOI= http://doi.acm.org/10.1145/1414471.1414494

[11] Brajnik, G. 2009. Validity and reliability of web accessibility guidelines. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility* (Assets '09). ACM, New York, NY, USA, 131-138. DOI=10.1145/1639642.1639666 http://doi.acm.org/10.1145/1639642.1639666

[12] Brajnik, G., Yesilada, Y. and Harper, S. 2010. Testability and validity of WCAG 2.0: the expertise effect. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '10). ACM, New York, NY, USA, 43-50. DOI=10.1145/1878803.1878813 http://doi.acm.org/10.1145/1878803.1878813

[13] Fuertes, J.L., Gutiérrez, E., & Martínez, L. 2011. Developing Hera-FFX for WCAG 2.0. In Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (W4A '11). ACM, New York, NY, USA, Article 3, 9 pages. DOI=10.1145/1969289.1969294. http://doi.acm.org/10.1145/1969289.1969294

[14] Gellenbeck, E. 2005. Integrating accessibility into the computer science curriculum. *J. Comput. Small Coll.* 21, 1 (October 2005), 267-273.

[15] Kelly, B., Sloan, D., Brown, S., Seale, J., Petrie, H., Lauke, P and Ball, S. 2007. Accessibility 2.0: people, policies and processes. In *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)* (W4A '07). ACM, New York, NY, USA, 138-147. DOI=10.1145/1243441.1243471 http://doi.acm.org/10.1145/1243441.1243471

[16] Hertzum, M., & Jacobsen, N.E. 2001. The evaluator effect: A chilling fact about usability evaluation methods. International Journal of Human-Computer Interaction, 13, 421-443.

[17] Law, C., Jacko, J., and Edwards, P. 2005. Programmer-focused website accessibility evaluations. In *Proceedings of the 7th international ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, MD, USA, October 09 - 12, 2005). Assets '05. ACM, New York, NY, 20-27. DOI= http://doi.acm.org/10.1145/1090785.1090792

[18] Leporini, B., Paternò, F., Scorcia, A. 2006. Flexible tool support for accessibility evaluation, *Interacting with Computers*, v.18 n.5, p.869-890, September, 2006 DOI=10.1016/j.intcom.2006.03.001

[19] Naftali, M. 2010. Analysis and integration of web accessibility metrics. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* (W4A '10). ACM, New York, NY, USA. DOI=10.1145/1805986.1805996 http://doi.acm.org/10.1145/1805986.1805996

[20] Pearson, E., Bailey, C., & Green, S. 2011. A tool to support the web accessibility evaluation process for novices. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education* (ITiCSE '11). ACM, New York, NY, USA, 28-32. DOI=10.1145/1999747.1999758. http://doi.acm.org/10.1145/1999747.1999758

[21] Sloan, D. 2006. The Effectiveness of the Web Accessibility Audit as a Motivational and Educational Tool in Inclusive Web Design. Ph.D. Thesis, University of Dundee, Scotland. June, 2006.

[22] Velleman, E., Meerveld, C., Strobbe, C., Koch, J., Velasco, C. A., Snaprud, M. and Nietzio, A. 2007. *D-WAB4 Unified Web Evaluation Methodology*. Web Accessibility Benchmarking Cluster. Retrieved 14th February 2010: http://www.wabcluster.org/uwem1_2/UWEM_1_2_CORE.pdf

[23] Waller, A., Hanson, V., and Sloan, D. 2009. Including accessibility within and beyond undergraduate computing courses. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility* (ASSETS '09). ACM, New York, NY, USA, 155-162. DOI=10.1145/1639642.1639670. http://doi.acm.org/10.1145/1639642.1639670

[24] WCAG 2.0 Assessment Tool. ACCESSIBLE Applications Design and Development Project. Accessed 14th February 2011: http://www.iti.gr/accessible/WCAG2.0_WebAssessmentTool_v2.0.zip

[25] Web Accessibility Toolbar 2.0. *The Paciello Group.* http://www.paciellogroup.com/resources/wat-ie-about.html. Accessed: 27/11/10.

[26] Yesilada, Y. Brajnik, G. and Harper, S. 2009. How much does expertise matter?: a barrier walkthrough study with experts and non-experts. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility* (Assets '09). ACM, New York, NY, USA, 203-210. DOI=10.1145/1639642.1639678 http://doi.acm.org/10.1145/1639642.1639678