# Measuring and Comparing the Reliability of the Structured Walkthrough Evaluation Method with Novices and Experts


Christopher Bailey
AbilityNet
Unit 25 Angel Gate Office Village
City Road
London, EC1V 2PT, UK.
+44 (0)7971 902 049
chris.bailey@abilitynet.org.uk

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

Voula Gkatzidou
Brunel University
St. Johns 201
Information Systems and Computing
Uxbridge, UB8 3PH, UK.
+44 (0)1895 267237
voula.gkatzidou@brunel.ac.uk



## ABSTRACT

Effective evaluation of websites for accessibility remains problematic. Automated evaluation tools still require a significant manual element. There is also a significant expertise and evaluator effect. The Structured Walkthrough method is the translation of a manual, expert accessibility evaluation process adapted for use by novices. The method is embedded in the Accessibility Evaluation Assistant (AEA), a web accessibility knowledge management tool. Previous trials examined the pedagogical potential of the tool when incorporated into an undergraduate computing curriculum. The results of the evaluations carried out by novices yielded promising, consistent levels of validity and reliability. This paper presents the results of an empirical study that compares the reliability of accessibility evaluations produced by two groups (novices and experts). The main results of this study indicate that overall reliability of expert evaluations was 76% compared to 65% for evaluations produced by novices. The potential of the Structured Walkthrough method as a useful and viable tool for expert evaluators is also examined.



## Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology*. 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.


## 1. INTRODUCTION

The Accessibility Evaluation Assistant (AEA) [1] is a web accessibility knowledge management tool that has been designed specifically to guide the novice auditor through the process of conducting an accessibility evaluation. Evaluation reports have a positive educational and motivational aspect on those who do not have expertise in web accessibility, [15] so a tool developed to assist a novice auditor could have 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) and has been used by over 100 students over a period of three years as part of their in course assessment.

The Structured Walkthrough evaluation method is embedded into the AEA. It represents a translation of an expert evaluation process adapted for use by novices. Based on the Barrier Walkthrough method [8], the Structured Walkthrough method breaks down the process of evaluating accessibility heuristics into a number of separate components that guide the novice auditor through the process of completing a manual accessibility evaluation. The rationale for each check is defined by a description of the barriers that individuals with disabilities may encounter. This helps to encapsulate the expertise and introduce the novice to the fundamental principles of accessible design. The AEA has previously been trialled with computing students and those experiments found that evaluations produced by novices using the Structured Walkthrough method have consistent levels of reliability and promising levels of validity when compared to an evaluation produced by an expert [4]. The evaluations produced by novices using the Structured Walkthrough method when compared to a WCAG 2.0 Conformance Review was found to be more effective, with higher levels of reliability and validity. Analysis of the qualitative data indicated that the students perceived the Structured Walkthrough method to be more useful and viable [5].

This paper presents the results of a third experiment designed to measure and compare the effectiveness of evaluations produced using the AEA by a number of experts as well as novices. The purpose of the study is to further establish the effectiveness of the Structured Walkthrough method for novices and measure the reliability of evaluations produced by a number of experts using the tool with the goal of demonstrating the potential of the Structured Walkthrough Method for use by experts.

The rest of the paper presents the current landscape of accessibility evaluation tools and methods (Section 2), and discusses the indicative problems emerging from the relevant research agenda. The overarching principles of the AEA tool and



*W4A '14,* April 07 - 09 2014, Seoul, Republic of Korea



http://dx.doi.org/10.1145/2596695.2596696

its embedded Structured Walkthrough method are presented in Section 2.1. This is followed by a description of the experiment (Section 3) with a discussion of the preliminary analysis (Section 4). The paper concludes with a presentation of the study results and a discussion of the findings and how these influence future research directions of accessibility evaluation (Section 6).

# 2. ACCESSIBILITY EVALUATION TOOLS AND METHODS

A range of evaluation support tools and methodologies have been developed to address the complexity of conducting accessibility evaluations. These have primarily been developed to support WCAG conformance review.

Tools to support evaluators include the MAGENTA tool [13] which was developed as a semi-automatic evaluation tool which checks a website against a specified set of guidelines and allows the user to conduct an accessibility evaluation from a range of guideline sets. OceanAcc integrates an automated evaluation tool with accessibility metrics [14] to provide an application with a semi-automatic evaluation process which is claimed to simplify and speed up the evaluation process. The Web Accessibility Assessment Tool (Waat) [19] 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. HERA was developed to assist in a manual WCAG 1.0 evaluation by including some explanatory text for each checkpoint, the ability to view an annotated version of the page being evaluated and the ability to view the page source code and functionality to generate a report. HERA 2.0 included an automatic preliminary analysis which assigned a value of pass, fail, not applicable or needs checking (indicating a manual examination is required) to each WCAG 1.0 Checkpoint [6]. Despite their apparent usefulness, a study measuring the effectiveness of 6 commercially available automated tools found that manual evaluation and human judgement is still a significant requirement in ensuring accessibility [18].

The Unified Web Evaluation Methodology is specifically aimed at expert evaluators (UWEM) [16] 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. While the UWEM methodology has been designed to support reliability, this has not been verified. Although the UWEM defines a structured test procedure, it is limited in that it while it describes what to test and provides criteria for pass/fail, it does not describe how to perform the test. The WAB Cluster developed a migration plan to incorporate WCAG 2.0 into the methodology [17]. WAB acknowledges the difficulties in standardising an evaluation methodology for UWEM 2.0 due to unstable W3C documentation that could be frequently amended. At the time of writing, development of UWEM 2.0 remains on-going and no formal documentation has been published. Barrier Walkthrough is itself an adaptation of heuristic walkthrough. The benefits of Heuristic walkthrough are that it can be adopted as a more reliable alternative to guideline conformance review and heuristic evaluation as it better constrains and guides the evaluator [7].

One of the benefits of adapting an expert evaluation into one that could be used effectively by novices is the potential to reduce the evaluator effect. The evaluator effect 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 can never be 100% [12]. The fundamental cause of the evaluator effort is that the process of an evaluation is a complex cognitive activity that requires evaluators to exercise difficult judgements and therefore even with guidance, the interpretation of any subjective element is heavily dependent on individual experience and background.

Most automatic evaluation support tools have been designed to support a conformance review evaluation; none have been developed purely to support manual, heuristic evaluation. One possible limitation of both Barrier Walkthrough and UWEM is that they do not support the auditor in performing a check or test to verify the presence of a given barrier. While originally conceived as a method to be used by novices, it is the integration of a defined checking technique into the Structured Walkthrough Method, as well as steps to assist verification in the presence of a barrier which makes it distinct from both Barrier Walkthrough and UWEM. These elements remove some of the requirements for individual judgement that we see has potential to reduce the evaluator effect in both novice and experienced evaluators

The primary rationale for development of the Structured Walkthrough method was to enable novices to produce an effective accessibility evaluation and, in the process, to increase their knowledge and awareness of web accessibility.

Recent work has focussed on the impact the evaluator's expertise has on the results of on accessibility evaluation. The W3C consider checkpoints to be reliably testable if 80% of knowledgeable evaluators would agree on the conclusion. Brajnik 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 [9]. 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 [2]. 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 [11]. An evaluation of the Barrier Walkthrough method with experts and non-experts concluded that the auditors' level of expertise is an important factor in the quality of an accessibility evaluation [21]. Expert judges were more effective at finding true accessibility barriers and spent significantly less time conducting their evaluation. These finding are supported by a study that evaluated the testability and validity of WCAG 2.0 with both experts and non-experts [10]. 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. The overall findings demonstrated that 50% of WCAG 2.0 Success Criteria failed to meet the W3C agreement threshold

of 80%. The study concluded that the reasons for this and possible solutions for the problem were subjects for further investigation. The results of these studies suggest an expertise gap when comparing the results of novice and expert evaluations regardless of the evaluation method adopted. There is a specific requirement for a higher level of understanding of accessibility when using more advanced tools and methods. This indicates a need for a tool or method which can support novices in developing 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).

The Structured Walkthrough method has the potential to reduce the evaluator effect by encapsulating expertise into the method itself and further guiding the evaluator by defining the specific checking technique.

## 2.1 THE ACCESSIBILITY EVALUATION ASSISTANT (AEA) TOOL

The AEA tool 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. The web-based tool allows the auditor to conduct an evaluation in a variety of contexts, such as by user group, by site features and by check categories. As a result, the auditor can carry out an effective audit without the need to go through a full set of checks - which streamlines the evaluation process and eliminates redundancy.

## 2.2 Evaluation by Check Categories

The Check Categories function supports a comprehensive accessibility evaluation using all 48 checks and is the primary function of the tool. The checks are broken down into five categories to make the evaluation process more manageable by grouping related checks in a meaningful way for novices. For instance, accessibility requirements addressed by Global Checks could be considered at the very start of a web development project whereas UX practitioners could consider accessibility requirements addressed by Design Checks when concept designs or wireframes are produced. The five categories are:

- **Design Checks:** Concerned with aspects of general presentation, the use of text and colour and the layout and positioning of items.
- **User Checks:** Practical checks with some degree of subjectivity checks that 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).
- **Structural Checks**: Concerned with the way content is structured (e.g. ensuring HTML Heading elements are used and implemented correctly to structure the content of a page).
- **Technical Checks:** Concerned with coding elements such as validating the HTML and CSS mark-up used to produce a webpage.
- **Global Checks:** Referring to issues that apply to the entire website (e.g. providing a Site Map) or refer to specific functionality (e.g. providing options for user customisation).

The checks are presented to the auditor in a list, along with a brief text summary. Many checks require the auditor to manually examine the website or webpage being checked and as such are not suitable for a solely automated process but the novice is supported with advice on the checking procedure. The AEA is not an automated evaluation tool, but does utilise existing resources - primarily the Web Accessibility Toolbar [20] - to simplify and support the process of testing and verification.

## 2.3 Evaluation by User Group

The Check by User Group function currently allows the auditor to prioritise checks based on the needs of 10 different user groups (e.g. Screen Reader User, Older Web User). 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.

## 2.4 Evaluation by Other Functions

The evaluation 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 is addressed using the AEA 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. For a fuller description of these components of the AEA see [3].

## 2.5 The Structured Walkthrough Evaluation Method

Based on the Barrier Walkthrough method [8], the Structured Walkthrough method breaks the process of evaluating accessibility heuristics down into a number of separate components (e.g. principle, short summary), guiding the novice auditor through the process of completing a manual accessibility evaluation. It provides the rationale for each check by defining and describing the barriers that individuals with disabilities may encounter, thus introducing the novice to the fundamental principles of accessible design and thus encapsulating expertise. The Structured Walkthrough method defines checks based on the specific heuristics and supports the novice evaluator with guidance and tutorials. The checks for each heuristic is broken into a number of components: the title of the accessibility principle (heuristic), a short summary, a general description of the check's importance in terms of the user group(s) affected and the nature of the barrier or problem caused, a description of the method and step-by-step instructions to perform the check, with the steps to verify and record the result, and a video demonstration of the check being performed in context. Figure 1 illustrates an example of the typical instructions provided for the auditor – in this case for checking image text alternatives. Integrating the rationale for each check into the sequence aims to improve the educational aspect of the evaluation method.

**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.**

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 one of the key elements of the AEA as an expert system. A short video demonstration of an expert evaluator performing the check including 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 is also included.

# 3. EXPERIMENT METHODOLOGY

The aim of this experiment was to measure and compare the quality of evaluations produced by both novices and experts using the Structured Walkthrough Method. The experiment had two elements; the aim of the first element was to measure the reliability of evaluations produced by novices. The aim of the second element was to measure the reliability of evaluations produced by experts in order to compare the two. The decision which the majority of experts agreed on was used to measure validity of the novice evaluations.

## 3.1 Participants

The first element of this experiment for novices was conducted with 28 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. The second part of the experiment involved six experienced accessibility practitioners. Participants were recruited through AbilityNet, an organisation dedicated to supporting disabled people using IT. All participants had over 10 years' experience working in an accessibility related role, primarily working as accessibility and usability consultants. They conduct WCAG audits on a regular basis as part of their normal working duties.

## 3.2 Materials and methods

Two websites were used for this study:

- Fitness First: http://www.fitnessfirst.co.uk
- Pure Gym: http://www.puregym.co.uk

The websites were chosen as they had both previously been identified as containing a number of potential accessibility barriers. An accessibility expert (the author) had previously conducted an accessibility review of each homepage in order to ensure a range of accessibility problems were present and could be detected using the information in the Structured Walkthrough Method. Some accessibility barriers were present on both websites, while others were unique to one. The participants would evaluate only a limited sub-set of checks from the total available. This was to make the experiment more efficient and eliminate redundancy. All of the selected checks were relevant to both websites to ensure uniformity in the checking procedure and to allow direct comparisons to be drawn between evaluations. 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; these are presented in Table 1.

**Table 1:** Checks Used During Evaluation.

| Check No. | AEA Heuristic |
|---|---|
| 1 | Images of Text |
| 2 | Colour Contrast |
| 3 | Moving Elements |
| 4 | Text Size |
| 5 | Keyboard Navigation |
| 6 | Link Names |
| 7 | Skip Navigation Link |
| 8 | Text Alternatives |
| 9 | Link Titles |
| 10 | Headings and Sub-Headings |
| 11 | Form Labels |
| 12 | Identify Language of Text |
| 13 | Validate (X)HTML Code |
| 14 | Site Map |
| 15 | Accessibility Information |

The novice users were divided into two groups and the experiment broken down into three separate activities conducted over three

weeks. In the first week one half of the cohort would evaluate the Home Page of one website and complete an evaluation template, in the second week they would evaluate the Home Page of the second website. Novices had previously been exposed to WCAG 2.0 documentation and had conducted WCAG Conformance checks in tutorial sessions but in this study we did not focus on WCAG 2.0 evaluations as part of the experiment. Novice users 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.

The experts were given the same evaluation report template as the novices and were required to perform an evaluation using the same 15 checks. They were only required to perform an evaluation on one of the homepages – Fitness First – in order to encourage participation thus maximising the response rate. They were given a briefing by email and instructed that for each check, they should judge if the requirements of the check were met, part met or not met and provide a short justification. Their judgement should be based on the instructions provided by the Structured Walkthrough method and their interpretation of the information contained in the AEA. Experts were requested to return their evaluations electronically within one week. The experts were asked to provide comments in free text responses describing their experience of the Structured Walkthrough Method compared to their normal process for conducting a WCAG 2.0 Conformance Review. They were also invited to comment on their impression of the AEA

## 3.3 Evaluation Method Comparison Metrics

Evaluation methods can be compared on a number of quality attributes such as effectiveness, efficiency and usefulness. To be accurately measured, these quality attributes may be customised to the individual contextual circumstances of an experiment [8]. Effectiveness is defined as 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.
3. Efficiency or viability refers to the amount of resources (e.g. time, skills, money, 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.
4. Usefulness is the effectiveness and usability of the results produced (with respect to those who assess, fix, or manage the accessibility of a website).

This would be the first instance that the AEA and the Structured Walkthrough Method had been used by more than one experienced accessibility practitioner to conduct a live evaluation. As such, we would also gain an insight into both the potential usefulness and viability of the AEA as an instructional tool and the Structured Walkthrough Method to be used more widely by experts.

For the purposes of this study quantitative data was used to measure validity and reliability. Reliability is defined as the extent to which evaluators reach the same decision and is applicable both to evaluations produced by the novices and by the experts. Validity is applied only to evaluations produced by novices and is defined as the extent to which novices made a decision that matched the majority of the experts. In effect, it refers to how 'correct' they were in their judgements. This is appropriate as the novices were mimicking the evaluation process of the expert and making the same subjective decisions as to whether the criteria for the AEA heuristic Criteria was Met, Not Met, or Partly Met. Qualitative data in the form of collated free text participant responses was used to gauge evaluator opinions of the AEA and Structured Walkthrough method and its perceived usefulness and viability.

# 4. RESULTS

The preliminary analysis of the results of the novice evaluations focuses on measuring their reliability and validity. In the case of the expert evaluations, the focus is solely on reliability.

## 4.1 Calculating Reliability and Validity

Reliability refers to the extent to which different evaluations of the same page lead to the same results. In the context of this study it is defined as the extent to which the participants agree on the result of a check. Validity refers to how 'correct' the novice evaluators were in their judgements, measured as the extent to which they reached the same judgment decisions as the majority of experienced evaluators. Using the data collated from the results of the evaluations the reliability and validity is calculated in the manner described in the following example.

Table 2 shows the results of five checks taken from the collated results of evaluations of the Fitness First Homepage. The numbers correspond to the number of novice evaluators who made each decision. The decision which the majority of expert evaluators made is marked in bold and with an asterisk.

Reliability (R) is calculated as the decision that the highest number of evaluators agreed upon. It is then expressed as a proportion of the maximum possible value. Using the data in Table 2, for the 'Colour Contrast' check, 15 out of 28 novice participants recorded a decision of 'met', while 13 out of 28 recorded a result of 'part met'. Therefore, the reliability of this one particular check is 15/28 or 54%. To calculate the overall figure, for each check the value of the decision that the highest number of evaluators agreed upon was used. In Table 2 the value taken from the 'Colour Contrast' check is 15; the 'Text Size' check is 23, etc.

**Table 2:** Example check results (Fitness First).

| Check | Decision | | | Reliability (R) | Validity (V) |
|---|---|---|---|---|---|
| | Met | Part Met | Not Met | | |
| Colour Contrast | 15 | **13*** | 0 | 15/28 | 13/28 |
| Text Size | **23*** | 5 | 0 | 23/28 | 23/28 |
| Text Alternatives | 2 | 16 | **10*** | 16/28 | 10/28 |
| Link Titles | 2 | 4 | **21*** | 21/28 | 21/28 |
| Language of Text | **23*** | 3 | 2 | 23/28 | 23/28 |

We express the overall figure of reliability as a proportion of the maximum possible value; in this case the total number of decisions made. In the example presented in Table 2, we have 28 evaluators performing 5 checks; this gives a total of 140 decisions. We calculate overall reliability (R) as follows:

$$R= (15 + 23 + 16 + 21 + 23) /140$$

The overall reliability of the 5 checks is 98/140 or 70%.

To calculate validity (V) the number of novice decisions that matched that of the majority of experienced evaluators was used. Using the data in Table 2, for the 'Language of Text' check, 23 out of 28 novice evaluators recorded a decision of 'met'; this matched the judgement of the majority of experienced evaluators. As 23 of the novice evaluators correctly matched the majority decision of experts, the validity of this particular check is 82%. The contribution of the result of this check towards the overall figure for validity is given the value of 23. Again, the overall figure is expressed as a proportion of the maximum possible value (the total number of decisions made). The overall validity (V) is calculated as follows:

$$V= (13 + 23 + 10 + 21 + 23) 90/140$$

The overall validity of the 5 checks is 90/140 or 64%.

Using the results of all 15 checks conducted by each evaluator, the overall reliability and validity figures for each page were calculated in this manner.

## 4.2 Reliability

In a study conducted by Yesilada et al. (2010) a heuristic was considered reliably evaluated if a majority of evaluators recorded the same result. A figure of 50% agreement is used as a consensus and the baseline for a minimum acceptable level of reliability. We examine the extent to which this occurs for each of the checks the novices conducted. Table 3 shows the figures of reliability for each check from the collated novice evaluations of the Fitness First homepage. In all but one of the checks (Headings and Sub-Headings) the level of reliability was over 50%. Similar to the study conducted by Alonso et al. (2010), we also set two other thresholds of 60% and 70% to measure levels higher than this minimum. Seven of the checks (47%) had a reliability of 60% or greater, while four of the checks (27%) had a reliability level of 70% or greater. Only two checks (13%) meet or exceed the W3C agreement threshold of 80%. Table 4 shows the figures of reliability for each check from the collated novice evaluations of the Pure Gym homepage.

**Table 3:** Reliability of checks of Fitness First Homepage (Novices).

| Check | Reliability (R) |
|---|---|
| Images of Text | 60% |
| Colour Contrast | 54% |
| Moving Elements | 57% |
| Text Size | 82% |
| Keyboard Navigation | 75% |
| Link Names | 57% |
| Skip Navigation Link | 68% |
| Text Alternatives | 57% |
| Link Titles | 75% |
| Headings and Sub-Headings | 39% |
| Form Labels | 50% |
| Identify Language of Text | 82% |
| Validate (X)HTML Code | 68% |
| Site Map | 57% |
| Accessibility Information | 50% |

**Table 4:** Reliability of checks of Pure Gym Homepage (Novices).

| Check | Reliability (R) |
|---|---|
| Images of Text | 43% |
| Colour Contrast | 61% |
| Moving Elements | 43% |
| Text Size | 86% |
| Keyboard Navigation | 64% |
| Link Names | 68% |
| Skip Navigation Link | 100% |
| Text Alternatives | 50% |
| Link Titles | 50% |
| Headings and Sub-Headings | 54% |
| Form Labels | 61% |
| Identify Language of Text | 79% |
| Validate (X)HTML Code | 100% |
| Site Map | 54% |
| Accessibility Information | 100% |

In all but two of the checks (Images of Text and Moving Elements) the level of reliability was over 50%. Nine of the checks (60%) had a level of reliability of 60% or greater, while 5

of the checks (33%) had a level of reliability of 70% or greater. Four of the checks (27%) meet or exceed the W3C agreement threshold of 80%. Table 5 shows the overall figures for reliability of the individual checks for novices averaged across both websites.

**Table 5:** Overall reliability of individual checks performed by novices.

| Check | Reliability (R) |
|---|---|
| Images of Text | 52% |
| Colour Contrast | 57% |
| Moving Elements | 50% |
| Text Size | 84% |
| Keyboard Navigation | 70% |
| Link Names | 63% |
| Skip Navigation Link | 84% |
| Text Alternatives | 54% |
| Link Titles | 63% |
| Headings and Sub-Headings | 46% |
| Form Labels | 55% |
| Identify Language of Text | 80% |
| Validate (X)HTML Code | 84% |
| Site Map | 55% |
| Accessibility Information | 75% |

**Table 6:** Reliability of checks of Fitness First Homepage (Experts).

| Check | Reliability (R) |
|---|---|
| Images of Text | 66% |
| Colour Contrast | 83% |
| Moving Elements | 83% |
| Text Size | 100% |
| Keyboard Navigation | 50% |
| Link Names | 66% |
| Skip Navigation Link | 66% |
| Text Alternatives | 83% |
| Link Titles | 66% |
| Headings and Sub-Headings | 66% |
| Form Labels | 50% |
| Identify Language of Text | 100% |
| Validate (X)HTML Code | 100% |
| Site Map | 66% |
| Accessibility Information | 83% |

Looking at the overall figures in Table 5, all but one check (Headings and Sub-Headings) has a reliability level of 50% or more. Eight (53%) have a reliability level of 60% or above, while six (40%) have a reliability level of 70% or greater. Four checks (27%) meet or exceed the W3C agreement threshold of 80%.

Table 6 shows the figures of reliability of each check from the collated expert evaluations of the Fitness First homepage. All 15 checks performed by experts have a figure of reliability of 50% or above meaning if the same criteria as Yesilada et al. (2010) is used the checks can be considered to have an acceptable level of reliability. 13 (87%) of the checks have a level of reliability of 60% or above, while seven (47%) have reliability of 80% or above.

## 4.3 Summary of Validity

For this experiment validity is defined as the extent to which the novices' decisions matched that of the majority of the expert evaluators. The data in Table 7 shows the extent to which this occurred, along with the overall figure. As the experts evaluated only one of the homepages, validity refers to the novice evaluations of the Fitness First page only.

**Table 7:** Validity of novice checks of Fitness First homepage.

| Check | Validity (V) |
|---|---|
| Images of Text | 14% |
| Colour Contrast | 46% |
| Moving Elements | 57% |
| Text Size | 82% |
| Keyboard Navigation | 75% |
| Link Names | 0% |
| Skip Navigation Link | 1% |
| Text Alternatives | 36% |
| Link Titles | 75% |
| Headings and Sub-Headings | 39% |
| Form Labels | 50% |
| Identify Language of Text | 92% |
| Validate (X)HTML Code | 68% |
| Site Map | 39% |
| Accessibility Information | 43% |
| **Overall** | **48%** |

When looking at the levels of validity in the results of the novice evaluations the results could be considered disappointing. Seven of the 15 checks (47%) had a level of validity of 50% or more. The levels of validity for three of the checks (Images of Text, Link Names and Skip Navigation Link) were particularly low. An analysis of the justifications provided can explain this.

In the case of Images of Text, the majority of experts recorded a result of 'not met'. This is because there were a number of instances of embedded text in images. The majority of novices recorded a decision 'part met'; they did identity that a barrier was

present, but recorded a different decision. In the case of Link Names, the experts correctly identified the presence of a small number of generic link names and the majority recorded a decision of 'not met'. The majority of novices recorded a decision of 'met' with many stating that generic link names were avoided; the assumption is that they were not thorough enough when performing their evaluation. This is similar to the findings of other studies [2]. In the case of the Skip Navigation Link check, the majority of experts correctly determined that although a Skip Navigation link was present, it was not fully functional and recorded a decision of 'not met'. The majority of novices recorded a decision of 'met' stating that the link was provided. Again, it appears that the novices were not thorough enough during the evaluation but we must also consider the possibility that the AEA did not provide them with enough information, for example, not fully describing how a skip navigation link should function properly.

## 4.4 Discussion

Examination of the figures for reliability on checks conducted by novices, are encouraging in that there was only one check that did not reach the 50% agreement threshold that has been used in previous studies as the benchmark for which a check can be considered reliably testable. Table 8 shows the overall figures for reliability presented by webpage and level of evaluator expertise.

**Table 8: Overall Reliability**

| Website | Reliability (R) | |
|---|---|---|
| | Novice Evaluations | Expert Evaluations |
| **Fitness First** | 62% | 76% |
| **Pure Gym** | 67% | - |
| **Overall** | 65% | 76% |

The reliability level of novice evaluations of the Fitness First homepage was 62% while the reliability of the evaluations of the Pure Gym homepage was 67%; this gives an overall figure of 65%. This is consistent with the overall level of reliability of evaluations produced by novices using the Structured Walkthrough Method in previous studies which ranged from 63% - 78% [4, 5]. When looking at the results of the individual checks, generally, reliability was higher for experts, but in the results of three checks (Keyboard Navigation, Skip Navigation Link and Form Labels) reliability was higher in the novice evaluations.

The overall figure of validity of 48% of novice evaluations in this study is, however, lower than that found in those previous studies. Previously, figures of validity ranged from 56% - 73%. The figure in this study has been reduced by the results of three checks that produced extremely low levels of validity and the possible reasons for this have been presented in Section 4.3. The levels of reliability shown in the results of the evaluations produced by experts are particularly encouraging. All 15 checks had a reliability level of 50% or more, with seven having a figure of over 80%. The overall figure was 76%. When considering the reliability of novice evaluations, whilst this 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, these figures are promising. With further refinement and testing, we are confident that overall figures for evaluations produced by experts could reach the 80% threshold. An analysis of feedback given by the expert practitioners provides some insight into how this might be achieved.

# 5. QUALITATIVE ANALYSIS

Qualitative data was gathered with the aim of gaining feedback from the experts on their experience of using the Structured Walkthrough Method with a view to examining its appropriateness, usefulness and viability as an expert evaluation method that would reduce the evaluator effect. The feedback from the experts was examined and a primary thematic analysis was conducted and the data coded in order to identify key trends. The expert evaluators commented favourably on the method itself. The experts found it easy to understand and quick to use; this suggests the method has potential for viability. This is reflected in the following comments:

*"Simple to understand and well structured. Could easily follow the steps based on the instructions provided."*

*"It was easy and succinct. Found it pretty useful."*

*"Much simpler (than WCAG 2.0) and more directed."*

The experts commented favourably on the actual presentation, content and structure of the checks. Elements that drew specific comment were:

*"The way the checks are divided or grouped along with instructions including videos."*

*"The information about why it (a check) is important and how to check it."*

*"Although not used during this test, I liked the ability to be able to prioritise the user groups so that the more relevant checks were completed."*

*"It was also good to see that within the checks they were separated out into critical, important and minor."*

The experts commented on the potential usefulness of the AEA and the Structured Walkthrough Method for use by less experienced evaluators or for conducting a preliminary inspection.

*"The evaluation tool would be very useful for someone with little accessibility experience. They would be able to evaluate a web page using the instructions and video provided."*

*"It was different in that there was a lot less hands on with the code. I felt it to be a more top level assessment using mainly the WAT toolbar rather than going into the code to find out what the actual problem was."*

*"I don't think it could replace a WCAG 2.0 audit but it does have the benefit of being a quick way to evaluate a number of pages to provide indicators as to where problem areas are before conducting a more in-depth WCAG 2.0 audit once the top level issues have been fixed."*

*"The videos provided a good background too and demonstrated the instructions well. This would be particularly helpful for testers who have little accessibility experience."*

One expert commented that due to their specific evaluation

technique, the AEA was not useful to them in its current form as it was not comprehensive enough:

*"Works well if only Internet Explorer is used however in my testing I will use Firefox inspectors and assistive technology to verify issues."*

This issue will be resolved with the provision of testing instructions for other tools (e.g.: WAVE Firefox extension), this would represent significant expansion of the AEA as a knowledge management tool. One expert commented that the lack of documentation or information on resolving the detected barriers could be detrimental to its usefulness:

*"Whilst there was an emphasis on detailing where there were errors, there was no emphasis on detailing how the issue could be resolved – i.e. with the addition of code snippets."*

One expert commented that they felt the instructions in the Structured Walkthrough Method were not complete enough or were not appropriate:

*"Some aspects not covered (colour\sensory reliance), heading interpretation is too strict. Not sure on coverage of link title attributes and how much of an impact adhering to this checkpoint would have in practical terms."*

Conversely, the requirements for some checks drew praise due to their specificity:

*"Requirement for a sitemap explicitly was good rather than the vaguer, WCAG 2.0 equivalent."*

One expert perceived the inclusion of a 'part met' judgement to be useful:

*"Ability to grade issues as partially met\not met felt useful. Checkpoints seemed quite broad, allowing for some degree of flexibility when interpreting."*

The concern experts had regarding this option was the subjectivity of the decision making process and the lack of clear decision support. This is reflected in the following comments:

*"It felt more subjective given the met\partially met\not met scores for each issue than a strict WCAG audit."*

*"….judgement was still required as to how to classify a check. If one of the points was not met does that mean 'part met' or 'not met'? How much common sense and judgement should be applied? However this is still much better than WCAG 2.0 where guidance at this level is a really big issue."*

*"Sometimes determining if an issue is not as cut and dried as yes or no although there is a 'Part Met' option."*

These comments suggest that an area of enhancement would be to provide clearer rules or guidance on the decision making process. This could increase the effectiveness of the Structured Walkthrough Method in reducing the evaluator effect.

The main criticisms of the tool were related to the visual design and poor usability of the AEA interface.

*"I felt the information was a little too close together and would benefit from more spacing between the heuristics. On a number of occasions I actually selected the wrong 'more' link as it was in closer proximity to the wrong paragraph."*

*"I didn't particularly like the double scrolling – one for the main page and the other for the main information."*

These issues are consistent with issues reported by novices in previous studies [4, 5]. While this is clearly an issue that need resolving, the AEA itself is a working prototype and many issues will be resolved with a redevelopment of the user interface.

# 6. CONCLUSION

Overall, the study adds further evidence of the usefulness of the Structured Walkthrough as an evaluation support tool for novices. The overall level of reliability using the AEA in evaluations produced by novices was 65% - which is consistent with previous studies. One limitation of the study is the comparatively small number of expert evaluations used. While a larger number of experts were contacted, gaining data from professional practitioners proved extremely problematic. Despite this, the level of expertise was high as all had over 10 years' experience as accessibility practitioners. For this reason, it is appropriate to draw some conclusions on the potential of the Structured Walkthrough Method to be used by experts.

The most important finding from the quantitative data from the element of the study conducted with experts demonstrates the potential for the Structured Walkthrough Method to be further developed as an expert evaluation method. We are encouraged by the overall level of reliability in the expert evaluations which was 76%. Qualitative data gained from the experts highlights elements of the method which need to be enhanced such as increasing the range of checks and proving more comprehensive guidance on the recording of a decision. Generally, the experts felt that the AEA tool would not be useful for them in conducting WCAG 2.0 evaluations in its current form as it was not complete or comprehensive enough. They did recognise its usefulness as a tool for those who did not have as a high a level of expertise or if conducting a preliminary review.

The quantitative data from this experiment and previous work demonstrate that novice evaluations using the Structured Walkthrough Method have consistent levels of reliability and support its continued use in this context. The reliability of evaluations produced by experts was high and the feedback from the experts indicates that this measure could be enhanced with guidance on how to record a decision for each check and that this would, in turn, reduce the evaluator effect. The qualitative data suggests the AEA tool itself is not a useful or appropriate means to deliver the method to experts.

The next stage of this work is to examine whether it would be possible to implement a Structured Walkthrough approach to WCAG 2.0 Success Criteria. Further development of the AEA would see each heuristic explicitly related to a relevant WCAG 2.0 Success Criteria. This could be integrated into the existing text for each heuristic, or more practically, added as an additional evaluation support function. By defining a testing and verification procedure and including information for how to record the results of a check the reliability of expert WCAG 2.0 evaluations could be improved and the evaluator effect reduced. Additionally, the provision of completed, expert evaluations would enhance the functionality of the AEA as a knowledge management tool for

training novices. This will be considered as a feature when the tool undergoes redevelopment.

## 7. REFERENCES


[1] Accessibility Evaluation Assistant. *Accessibility Research Centre.* http://arc.tees.ac.uk/aea/. Accessed: 5/1/14.

[2] 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. http://doi.acm.org/10.1145/1805986.1806000

[3] 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

[4] 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. http://doi.acm.org/10.1145/1969289.1969293

[5] Bailey, C. & Pearson, E. 2012. Evaluation of the effectiveness of a tool to support novice auditors. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility* (W4A '12). ACM, New York, NY, USA, Article 33, 10 pages. http://doi.acm.org/10.1145/2207016.2207044

[6] Benavidez, C., Fuertes, J., Gutierrez, E., & Martinez, L. 2006. Teaching web accessibility with Contramano and HERA. *Computers Helping People with Special Needs, 4061*, 341-348.

[7] Brajnik, G. 2005. Accessibility assessments through heuristic walkthroughs. *Proceedings of HCItaly 2005.* Rome, Italy.

[8] 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

[9] 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. http://doi.acm.org/10.1145/1639642.1639666

[10] 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. http://doi.acm.org/10.1145/1878803.1878813

[11] 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. http://doi.acm.org/10.1145/1969289.1969294

[12] 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.

[13] 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

[14] 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. http://doi.acm.org/10.1145/1805986.1805996

[15] 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.

[16] 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. http://www.wabcluster.org/uwem1_2/UWEM_1_2_CORE.pdf. Accessed: 5/1/14.

[17] Velleman, E., Meerveld, C., Strobbe, C., Koch, J., Verlasco, C. A., Snaprud, M. & Nietzio, A. 2007. Migration Plan from UWEM 1.2 to UWEM 2.0. http://www.wabcluster.org/uwem1_2/MigrationPlanFinal.pdf Accessed: 5/1/14.

[18] Vigo, M., Brown, J., & Conway, V. 2013. Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (W4A '13). ACM, New York, NY, USA, Article 1, 10 pages.

[19] Web Accessibility Assessment Tool (WaaT). http://www.accessible-eu.org/index.php/waat.html. Accessed: 5/1/14.

[20] Web Accessibility Toolbar 2.0. *The Paciello Group.* http://www.paciellogroup.com/resources/wat. Accessed: 5/1/14.

[21] 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