Comparison of Manual Evaluation Methods for Assessing the Accessibility of Websites based on EN 301 549

Gisela Kollotzek[1], Gottfried Zimmermann[1], Tobias Ableitner[1] and Anne-Marie Nebe[2]
[1]*Competence Center for Digital Accessibility, Stuttgart Media University, Nobelstr. 10, 70569 Stuttgart, Germany*
[2]*T-Systems Multimedia Solutions GmbH, Competence Center for Digital Accessibility & Software Ergonomics, Riesaer Str. 5, 01129 Dresden, Germany*

Keywords: Comparative Study, Conformance, Disability, Empirical Evaluation, Expert Evaluation, Guidelines, Web Accessibility, Accessibility Evaluation Method, WCAG.

Abstract: The relevance of information technology has increased significantly over the last couple of years and therefore it is important to provide for universal access to it. Accessibility of public sector websites has become legally binding through Directive (EU) 2016/2102 for EU member states. Automatic accessibility evaluation methods can only provide a superficial impression of the accessibility status. Only manual evaluation methods can facilitate a comprehensive accessibility check. So far, there is no systematic comparison of existing manual evaluation methods available that is based on real data. In this paper, we define a generic catalog of 22 criteria for assessing the quality of accessibility evaluation methods and specify individual weights for the criteria. We then compare two representatives of manual evaluation methods: BIK BITV-Test, as a representative of conformance-based methods; and BITV-Audit, as a representative of empiric-based methods. We analyze similarities and differences between these two methods, and identify weaknesses and strengths. Our results show an advantage of BITV-Audit over BIK BITV-Test, but other weightings could yield different results.

1 INTRODUCTION

In recent years, digitization has increased in all areas of life. At the same time, the number of people with disabilities in our society has risen continuously. At the end of 2019, around 7.9 persons with a disability degree of 50% or higher lived in Germany. This was around 136,000 more than at the end of 2017. The share of persons with severe disabilities in the total population in Germany was thus 9.5% (Statistisches Bundesamt (Destatis), 2020). Therefore, accessibility in information technology has become more important over the last few years. It has become legally binding in the public sector through Directive (EU) 2016/2102. Automated evaluation methods can be used to assess accessibility but they cannot adequately evaluate all relevant requirements (Vigo et al., 2013). A study by Burkard et al. (2021) showed that with various tested automatic evaluation tools only 15 to 40% of all existing accessibility problems can be identified in the evaluated objects. The automatic evaluation methods can only provide a superficial impression of the accessibility status. However, they cannot determine whether the website is actually accessible.

Currently, only manual evaluation methods can ensure a comprehensive accessibility check. These evaluation methods involve a great deal of time and the qualification of the evaluators. Therefore, the creation of accessible content as well as the certification of accessibility involve a great deal of effort and associated high costs. Currently, there is still little research in the area of manual evaluation methods compared to automatic evaluation methods. However, the further development of both types of evaluation methods is a necessary step for the provision of more accessible websites.

The underlying research question for our study looks at the differences and similarities between conformance-based and empiric-based manual accessibility evaluation methods for websites. A conformance-based evaluation method assesses the degree of fulfillment of a specified set of requirements. Therefore, it can tell whether the evaluation object as a whole can be classified as conforming or not conforming to a specified standard. In the empiric-based evaluation methods, trained evaluators or people with disabilities perform the accessibility evaluation from the perspectives of affected

Kollotzek, G., Zimmermann, G., Ableitner, T. and Nebe, A.
Comparison of Manual Evaluation Methods for Assessing the Accessibility of Websites based on EN 301 549.
DOI: 10.5220/0010647000003060
In *Proceedings of the 5th International Conference on Computer-Human Interaction Research and Applications (CHIRA 2021)*, pages 24-35
ISBN: 978-989-758-538-8; ISSN: 2184-3244

user groups (e.g., visually impaired users). A benefit of empiric-based methods is, that it is also possible to assess how serious the problems found are for the respective user groups.

This paper provides the following contributions on accessibility evaluation methods:

- We specify a catalog of criteria for the comparison of evaluation methods. These criteria were determined and weighted in a process involving accessibility experts.
- We define up to six metrics for each criterion to quantify them. The catalog can be used to analyze manual evaluation methods.
- Based on this catalog of criteria, we compare two German evaluation methods: BIK BITV-Test and BITV-Audit. We identify their similarities and differences as well as their weaknesses and strengths.

The remainder of this paper is structured as follows: In Section 2, we introduce the necessary background. It considers the legal framework conditions and presents the current state of research in the field of manual evaluation methods. In Section 3, we take a more in-depth look at the two manual evaluation methods selected for comparison. We present the criteria and weighting established by the experts in Section 4. In Section 5, we will elaborate on the comparison of the evaluation methods based on the defined criteria. We discuss the results in Section 6 and finally provide a conclusion of this paper in Section 7.

2 BACKGROUND

2.1 Laws and Guidelines

2.1.1 Directive (EU) 2016/2102

The European Parliament and the Council of the European Union adopted Directive (EU) 2016/2102 (European Parliament and Council of the European Union, 2016) on the accessibility of websites and mobile applications of public sector bodies, also known as the Web Accessibility Directive, on October 26, 2016. Directive 2016/2102 requires public sector bodies to provide accessible web services and mobile applications. Aside from technical requirements, (EU) 2016/2102 also specifies that public sector bodies must provide an accessibility statement for their websites. The EU Directive 2016/2102 becomes effective upon transposition into national law by the European Union member states.

2.1.2 Web Content Accessibility Guidelines 2.1

The Web Content Accessibility Guidelines (WCAG) represent globally recognized recommendations for making website content accessible. The World Wide Web Consortium (W3C) published the most current version 2.1 on June 5, 2018 (W3C, 2018). The WCAG have a hierarchical structure, meaning that the principles are subordinate to the guidelines. These, in turn, are assigned to the success criteria. Each success criterion belongs to one of three conformance levels. 30 success criteria belong to the lowest level (A) and 20 to the middle level (AA). The highest conformance level (AAA) includes the remaining 28 success criteria. These conformance levels correspond to different levels of accessibility implementation. As an example, for a website to comply with conformance level AA, all success criteria of level A and level AA must be met (W3C, 2018).

It is important to note that following WCAG – being a standard – is voluntary, unless regulation prescribes conformance.

2.1.3 EN 301 549

EN 301 549 V2.1.2 (European Telecommunications Standards Institute, 2018) specifies requirements for accessibility of information and communication technologies (ICT). It was "harmonized" with (EU) 2016/2102, i.e., conforming to EN 301 549 v2.1.2 is considered sufficient for fulfilling the technical requirements of (EU) 2016/2102 (European Parliament and Council of the European Union, 2018). EN 301 549 covers all areas of ICT: ICT with two-way voice communication, ICT with video capabilities, hardware, web, non-web documents, software, documentation and support services, and ICT providing relay or emergency service access. Chapter 9 of EN 301 549 focuses on requirements for websites. In this context, EN 301 549 is strongly oriented towards WCAG 2.1 AA concerning web content. Annex A of EN 301 549 defines in detail the relationship between the standard and the requirements of Directive 2016/2102. In particular, table A.1 provides a list of all requirements for web content from all chapters of EN 301 549, including the requirements of WCAG 2.1 AA from chapter 9. Version V2.1.2 was published in August 2018. At the time of writing this paper, a new version V3.2.1 (European Telecommunications Standards Institute, 2020) is available, but not yet harmonized with (EU) 2016/2102. Nevertheless, in this paper, EN 301 549 refers to the new version V3.2.1, as it is expected to become the "harmonized" version eventually.

2.1.4 Barrierefreie-Informationstechnik-Verordnung 2.0[1] (BITV 2.0)

BITV 2.0 (German Federal Ministry of Labour and Social Affairs, 2019) is the regulation for the creation of barrier-free information technology in accordance with the Behindertengleichstellungsgesetz[2] (BGG) in Germany. Version 2.0 replaced the BITV in 2011 and the last amendments were made in May 2019. Through these latest amendments, it implements the requirements of Directive 2016/2102, which are not already included in the BGG.

2.1.5 Relation between Laws and Guidelines

Through harmonization, Directive (EU) 2016/2102 points to EN 301 549 for the technical requirements, and includes additional requirements for non-technical purposes. EN 301 549 adopts the requirements of WCAG 2.1 AA for websites, but adds more requirements of more generic nature in its Annex A. A website is considered accessible if all the requirements specified in Directive 2016/2102 and in EN 301 549 Annex A are met.

Since Directive 2016/2102 is a European Directive, the member states had to implement it in their respective national law. In Germany, this was achieved by an amendment of the BGG and by updating BITV 2.0. In Germany, a public sector website is considered accessible if it meets the requirements as specified in BITV 2.0.

2.2 Related Work

Lang (2004) looks at three different types of evaluation methods for assessing the accessibility of websites in terms of their effectiveness through a literature review: automatic evaluation methods, manual evaluation methods, and user studies involving people with disabilities. The results show that none of the evaluation methods examined can identify all accessibility problems. Lang concludes that a combination of the three types of methods is the most effective way to find the largest possible number of accessibility problems and thus be able to improve accessibility in the best possible way.

The study by Mankoff et al. (2005) compares the evaluation methods already mentioned by Lang (2004). The comparison considers the following criteria: effectiveness, validity, and the number of problems found. It turns out that the conformity-based evaluation method with screen readers performed by multiple evaluators is the most successful in finding different classes of problems. However, Mankoff et al. indicate that each type of evaluation method has its respective weaknesses and strengths.

Manual evaluation methods play an important role in the evaluation of website accessibility. Therefore, Yesilada et al. (2009) aim to investigate the influence of expertise on accessibility evaluation. It turns out that the degree of expertise is an important factor for the quality of the evaluation. Another important influencing factor is the selection of the web pages of a website. A web page sample should be able to represent the degree of accessibility of the whole website. Various studies investigate the influence of the sample selection method and sample size (Velleman and van der Geest, 2013; Brajnik et al., 2007).

A commonly used type of manual evaluation method is the conformance-based evaluation method. Standards and guidelines are the basis of this method type. These are largely responsible for the quality of the evaluation. An internationally recognized and frequently used standard is WCAG. Various studies look at the validity and reliability of WCAG 1.0 as well as 2.0 (Rømen and Svanæs, 2008; Brajnik, 2009; Calvo et al., 2016). Freire (2012), however, observes that the guidelines (WCAG 1.0 and WCAG 2.0) do not cover all issues found by impaired users.

Brajnik (2005) presents a new concept for an evaluation method based on a heuristic walkthrough, which is a method for evaluating the usability of a website. In a later work, Brajnik (2006) investigates this heuristic walkthrough method, then called the barrier walkthrough method. For the assessment of the evaluation methods, Brajnik identifies the following criteria: validity, reliability, usefulness, and efficiency. Brajnik (2008) further defines the following criteria for comparing evaluation methods: effectiveness, efficiency, and usefulness. Effectiveness is broken down more precisely into reliability and validity. Validity in turn is broken down into the two criteria correctness and sensitivity. He uses these criteria (excluding usefulness) to compare two evaluation methods: the conformance-based evaluation (with an Italian standard) and the barrier walkthrough method. The comparison aims to determine the merits of the barrier walkthrough method, whereas the conformance-based evaluation method serves as a control condition. The results show differences in both evaluation methods for validity and reliability. However, the evaluations were performed by less experienced evaluators and might give different results for experienced evaluators.

To summarize, the existing work on evaluation

[1]Translated into English: Barrier-free Information Technology Regulation 2.0

[2]Translated into English: Disability Equality Act

methods mostly looks at the field on a conceptual level and presents new evaluation methods. A comprehensive systematic comparison using real data produced by accessibility experts is currently missing. Therefore, in the following sections, we present our comparison of two types of manual evaluation methods: conformance-based and empiric-based methods. We present our generic catalog of comparison criteria that we developed with experts' input. It can be adapted or extended to meet possible future requirements. Based on this catalog, we compare two concrete evaluation methods: BIK BITV-Test (DIAS GmbH, 2019) and BITV-Audit (Nebe, 2021).

3 METHODS FOR ASSESSING THE ACCESSIBILITY OF WEBSITES

To identify the differences, similarities, advantages, and disadvantages of the different types (conformance-based vs. empiric-based) of manual evaluation methods, we select one representative of each type. There is a variety of evaluation methods. However, to be able to carry out an expert-based comparison, we consider only evaluation methods involving experts as testers. Apart from that, to provide comparable assessment results, both representatives must provide a certificate for complying websites.

We selected the following evaluation methods: the BIK BITV-Test (conformity-based) and the BITV-Audit of T-Systems Multimedia Solutions GmbH (empiric-based). The BIK BITV-Test is the established evaluation method in Germany for the conformity assessment of websites. In the context of this work, we only consider the version as of 2020. Note that the BIK BITV-Test has been significantly revised in its current version, March 2021. We choose the BITV-Audit as a representative of the empiric-based evaluation methods. The evaluation procedure consists of an accessibility evaluation and a conformance-based evaluation. In the empirical expert evaluation, accessibility is considered from the perspective of people with disabilities. Subsequently, the evaluator assigns the problems to the conformity criteria and checks whether any criteria not yet considered have been violated.

3.1 BIK BITV-Test

The BIK BITV-Test is an evaluation method for assessing the accessibility of websites and web applications. The BIK project developed this evaluation method together with accessibility experts, disability associations, and service providers for websites. The procedure is fully disclosed. It was first published in 2004 and has been continuously developed since then (DIAS GmbH, 2020b).

The following standards are the basis of this evaluation method: BITV 2.0, EN 301 549 (Section 9 Web), WCAG 2.1 AA. The evaluation method does not cover the additional requirements of BITV 2.0 which are not included in EN 301 549, such as the need for an accessibility statement, and the need for a description in sign language and easy language. Furthermore, the additional requirements of Table A.1 of EN 301 549 are not included in the 2020 version. There are also no evaluation steps that implement the WCAG 2.1 success criteria of conformance level AAA. This method is not suited to evaluate non-web documents e.g. PDF documents (DIAS GmbH, 2019).

There are three variants of the BIK BITV-Test: *Development-accompanying BITV-Test*, *final BITV-Test*, and *BITV self-assessment*. All three variants have the same scope of evaluation steps but differ in the evaluation procedure. While a certification of conformity can only be achieved by a *final BITV-Test*, the results of the *development-accompanying BITV-Test* and the *BITV self-assessment* may only be used for internal use (DIAS GmbH, 2020c). In the following chapters, reference is made exclusively to the *final BITV-Test*.

In the beginning, the expert and the client select the appropriate evaluation variant depending on the goal to be pursued with the evaluation. They clarify which areas and web pages belong to the evaluated website. Subsequently, the required scope can be estimated. The initial evaluator creates the evaluation data in the BITV-Test tool, assigns a second evaluator, and selects the web pages to be evaluated. A manual evaluation of the whole website is not feasible, because such an evaluation would be too costly. Therefore, a representative web page sample is made. This selection step is done without the client. The size of the sample depends on the complexity of the evaluation object, i.e., the website under investigation. A qualified evaluator of the BIK evaluation association, hereafter referred to as the QA (for "quality assurance"), checks the web page sample and releases it for evaluation. The initial and second evaluators perform the evaluation steps independently of each other. After each evaluator has completed the evaluation steps, they discuss their results in a joint meeting. Subsequently, the QA carries out the quality assurance of the found results. If the QA identifies necessary corrections, the first evaluator revises them. Optionally,

several correction loops can be made here. As soon as this correction work is completed, the QA can generate the evaluation report from the BITV-Test tool and the initial evaluator sends it to the client. The client can expose a link on their website to the published report as a certification of accessibility. A successful BITV-Test entitles to an official seal of approval which can be obtained for compliant websites (DIAS GmbH, 2019).

3.2 BITV-Audit

The BITV-Audit is an evaluation method for assessing the accessibility of websites, mobile applications, documents, and desktop software (T-Systems Multimedia Solutions GmbH, 2020a; Nebe, 2021). T-Systems Multimedia Solutions GmbH (T-Systems MMS) used the evaluation method for the first time in 2009. In 2010, the BITV-Audit was accredited for the first time and has been continuously developed since then (A. Nebe [T-Systems MMS], personal communication, April 12, 2021). A detailed description of the evaluation procedure including the evaluation criteria is currently not publicly available.

The following laws and standards are the basis of the BITV-Audit: BITV 2.0, EN 301 549, WCAG 2.1, EN ISO 9241-171, and ISO 14289-1 (T-Systems Multimedia Solutions GmbH, 2020b).

There are two variants of the evaluation: *simplified audit* and *in-depth audit*. The *simplified audit* records vulnerabilities and their severity in a short protocol with problem descriptions and heuristic solution recommendations based on standard specifications. The *in-depth audit* additionally provides a root cause analysis for each problem, an impact description for each user group, and a detailed recommended action to solve the problem (T-Systems Multimedia Solutions GmbH, 2020a). In the following chapters, reference is made exclusively to the *in-depth audit*.

In the beginning, the expert and the client define the framework conditions. Based on this, they select the appropriate variant of the evaluation method which can be optionally supplemented by a usability test. Finally, they determine which web pages and areas belong to the evaluated website. The evaluator receives a program introduction from the client. For simple websites, this step can be omitted. Based on the website and the agreements made with the client, the relevant evaluation criteria are automatically selected. Furthermore, the expert makes a representative web page selection on which the evaluation is carried out.

The evaluation then consists of two parts: the accessibility evaluation and the conformance-based evaluation. The former involves an empirical evaluation by experts, in which the accessibility of the evaluation object is considered from the perspective of users with disabilities. Five user groups are distinguished for this purpose: visually impaired, blind, motor impaired, hearing impaired, and cognitively impaired users. The expert assigns the problems found to the conformance criteria of the standards and laws involved (see above). Finally, the evaluator examines all criteria that have not yet been considered. The severity of the impact on the affected user groups is the basis for the weighting of the problems. There is a team review with at least one other evaluator. They discuss all problems found with regard to the problem cause, weighting, and recommended action. The evaluator then prepares the evaluation report. Finally, an experienced evaluator formally approves it and sends it to the client. An official seal of approval which can be obtained for compliant websites (T-Systems Multimedia Solutions GmbH, 2020b).

4 CATALOG OF CRITERIA

4.1 Procedure

For the comparison, we derived a pre-selection of 14 criteria through a literature review. Following this, experts discussed and revised this pre-selection in a focus group and supplemented it with further criteria. The nine-member expert panel consisting of members of the BIK evaluation association, T-Systems MMS staff, and other accessibility experts met in a virtual session. In the end, the outcome was a total of 22 criteria, listed in Table 2.

Following the virtual expert panel, each participant was sent a spreadsheet containing the 22 criteria. The experts were allowed to assign as many points as they liked to each criterion. However, the understanding for this was that a criterion received points in relation to the subjective importance it bore for the individual experts. If a criterion was considered completely irrelevant, it would receive 0 points. In the end, eight of the nine experts assigned a weighting. The weightings were normalized and then averaged.

4.2 Criteria

The catalog of comparison criteria consists of 22 items. Each criterion has a weight, as determined by averaging the experts' individual weights. To be able to quantify the results, for every criterion, we define up to six metrics which are equally weighted

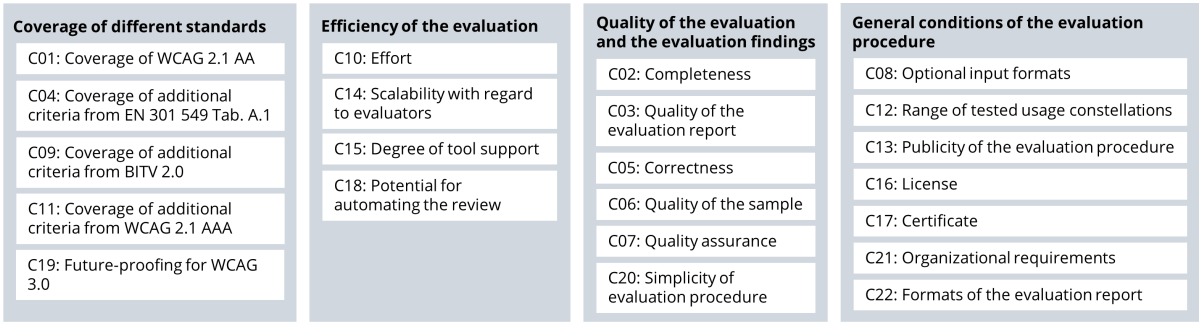

Figure 1: Overview of the 22 criteria divided into 4 groups.

among each other. The total score (TS) of an evaluation method consists of the sum of the products of the individual criterion scores (CS) and their weights (W), as seen in equation (1).

$$TS = \sum_{criterion=1}^{22} \left(CS_{criterion} \cdot W_{criterion} \right) \quad (1)$$
$$with\, 0 \leq TS, CS, W \leq 1$$

Based on the experts' ratings, the following five criteria have the highest weights: "C01: Coverage of WCAG 2.1 AA" ($W = 7.72$), "C02: Completeness" ($W = 7.39$), "C03: Quality of the evaluation report" ($W = 6.67$), "C04: Coverage of additional criteria from EN 301 549" ($W = 6.64$), and "C05: Correctness" ($W = 6.32$). An overview of the 22 criteria and their weights is shown in Table 2. The criteria can be divided into the following four groups: Coverage of different standards, efficiency of the evaluation, quality of the evaluation and the evaluation findings, and general conditions of the evaluation procedure. The mapping is shown in Figure 1.

5 STUDY RESULTS

The procedure descriptions and created evaluation reports are the basis of the comparison of the evaluation methods. Furthermore, accessibility experts processed the created evaluation reports. In addition, the subjective assessments of accessibility experts on specific aspects, as obtained in interviews, are taken into account.

5.1 Procedure

In a first step, we selected the websites to be used as case studies for both evaluation methods: the website of the Hochschule der Medien[3](HdM) and the website of the Bundesfachstelle für Barrierefreiheit[4](BfB). These two websites were selected under the assumption that they differ in the degree of their accessibility. This allows us to compare the results for both a less accessible website and a more accessible website. We decided that only the two start pages of the websites should be used as samples to reduce the scope of the evaluations. The start pages contained a selection of layouts, navigation menus as well as other elements. Thus, many evaluation steps were applicable. An employee of T-Systems MMS carried out the evaluations since he is officially qualified to perform both evaluation methods. Each website was evaluated with both the BIK BITV-Test and the BITV-Audit. The order of evaluation was alternated to avoid bias effects.

The three criteria "C02: Completeness", "C05: Correctness", and "C20: Simplicity of evaluation procedure" are based on a classification of the problems found by experts into the following four categories:

- False positives: The evaluation method identifies a problem as such, but it is not a real problem.
- False negatives: The evaluation method does not identify a problem as such, but it is a real problem.
- True positives: The evaluation method identifies a problem correctly as a problem.
- Mistakes: The evaluation method would identify a problem as such, but the evaluator does not identify it when applying the evaluation procedure.

We defined WCAG 2.1 AA as the reference point for true positives because it is the legal requirement in the European legislation (European Parliament and Council of the European Union, 2018). It has also been shown that the classification of the problems must additionally be carried out from the user's point

[3]translated into English: Stuttgart Media University; accessible at https://www.hdm-stuttgart.de/

[4]translated into English: Federal Agency for Accessibility; accessible at https://www.bundesfachstelle-barrierefreiheit.de/DE/Home/home_node.html

Table 1: Overview of prepared results of the evaluation reports.

	BIK BITV-Test		**BITV-Audit**	
	HdM	BfB	HdM	BfB
False positives (WCAG 2.1 AA)	3	2	7	11
False positives (user perspective)	1	3	0	1
False negatives (WCAG 2.1 AA)	12	5	6	3
False negatives (user perspective)	20	17	9	4
True positives (WCAG 2.1 AA)	62	24	68	26
True positives (user perspective)	64	23	75	36
Mistakes	7	7	12	3

of view since these two perspectives can lead to numerous divergent results. The results are shown in Table 1

A total of 74 problems (the sum of false negatives and true positives) according to WCAG 2.1 AA and 84 problems from the user's perspective were identified on the HdM website. On the BfB website, 29 problems were found according to WCAG 2.1 AA and 40 problems from the user's point of view. The results confirm our assumption that the two websites differ in their accessibility.

The results of the expert classification show that ten problems on the HdM website would not be assessed as a problem by WCAG 2.1, but still represent a usability problem for users. For the BfB website, there are 13 problems which are not identified as a problem by WCAG 2.1 but still represent a usability problem for users. However, there are also two problems on the BfB website that would be assessed as a problem by WCAG 2.1 but do not represent usability or accessibility problems for the users.

For the following four of the total 22 criteria, the subjective assessment of accessibility experts are taken into account in the evaluation: "C03: Quality of the evaluation report", "C14: Scalability with regard to evaluators", "C15: Degree of tool support" and "C20: Simplicity of evaluation procedure". For this purpose, we conducted virtual semi-structured expert interviews in the period from 03/08/2021 to 03/22/2021. The interviews were recorded with the consent of the experts. All data were stored anonymously. A total of six experts participated. Three members of the BIK evaluation association were interviewed about the BIK BITV-Test and three employees of T-Systems MMS about the BITV-Audit. One employee of T-Systems MMS could also participate in the interview for the BIK BITV-Test, as he is also a member of the BIK evaluation association. The interview partners assigned a value to each item on a 7-point Likert scale and justified their decision in a qualitative manner. To avoid bias effects, the items were presented to the interviewees in differing order.

5.2 Results

The BITV-Audit from T-Systems MMS achieves a total score of 0.69, the BIK BITV-Test only 0.50. An overview of the results is shown in Table 2. All results are rounded to two digits after the period. For both evaluation methods, the criteria "C18: Potential for automating the review" and "C19: Future-proofing for WCAG 3.0" were not evaluated, as no valid quantification could be defined here.

5.2.1 Similarities and Differences of the Evaluation Methods

In the following, we provide an overview on our findings regarding similarities and differences between the two methods under investigation.

> With a difference ($Diff$) of less than 0.10 between the criterion scores of the two evaluation methods ($Diff < 0.10$), only minor deviations are observed. The evaluation methods show great similarities for this criterion.
>
> At a difference ($Diff$) of at least 0.50 between the criterion scores of the two evaluation methods ($Diff \geq 0.50$), strong deviations are observed. The evaluation methods show large differences for this criterion.

For the following four criteria, the evaluation methods have a lot in common: "C01: Coverage of WCAG 2.1 AA" ($Diff = 0.00$), "C05: Correctness"($Diff = 0.04$), "C15: Degree of tool support" ($Diff = 0.08$), and "C20: Simplicity of evaluation procedure" ($Diff = 0.05$).

For the following seven criteria, the evaluation methods differ greatly: "C04: Coverage of additional criteria from EN 301 549 Tab. A.1" ($Diff = 0.71$; in favor of BITV-Audit), "C09: Coverage of additional criteria from BITV 2.0" ($Diff = 1.00$; in fa-

Table 2: Results of all criteria for the BIK BITV-Test and the BITV-Audit.

No.	Criterion	Weight (in %)	BIK BITV-Test	BITV-Audit
C01	Coverage of WCAG 2.1 AA	7.72	1.00	1.00
C02	Completeness	7.39	0.75	0.90
C03	Quality of the evaluation report	6.67	0.30	0.79
C04	Coverage of additional criteria from EN 301 549 Tab. A.1	6.64	0.02	0.74
C05	Correctness	6.32	0.94	0.90
C06	Quality of the sample	5.96	0.73	0.93
C07	Quality assurance	5.58	0.73	0.60
C08	Optional input formats	5.57	0.00	1.00
C09	Coverage of additional criteria from BITV 2.0	5.53	0.00	1.00
C10	Effort	4.77	0.17	0.29
C11	Coverage of additional criteria from WCAG 2.1 AAA	4.70	0.00	1.00
C12	Range of tested usage constellations	4.19	0.63	0.74
C13	Publicity of the evaluation procedure	4.07	1.00	0.00
C14	Scalability with regard to evaluators	4.04	0.63	0.80
C15	Degree of tool support	3.87	0.52	0.60
C16	License	2.96	0.50	0.00
C17	Certificate	2.90	1.00	0.75
C18	Potential for automating the review	2.90	0.00	0.00
C19	Future-proofing for WCAG 3.0	2.80	0.00	0.00
C20	Simplicity of evaluation procedure	2.44	0.54	0.59
C21	Organizational requirements	1.51	1.00	0.00
C22	Formats of the evaluation report	1.48	0.50	0.25
	Total score	**100**	**0.50**	**0.69**

vor of BITV-Audit), “C11: Coverage of additional criteria from WCAG 2.1 AAA” ($Diff = 1.00$; in favor of BITV-Audit), “C08: Optional input formats” ($Diff = 1.00$; in favor of BITV-Audit), “C16: License” ($Diff = 0.50$; in favor of BIK BITV-Test), “C13: Publicity of the evaluation procedure” ($Diff = 1.00$; in favor of BIK BITV-Test) and “C21: Organizational requirements” ($Diff = 1.00$; in favor of BIK BITV-Test).

5.2.2 Strengths and Weaknesses of the Evaluation Methods

In the following, we provide an overview on our findings regarding strengths and weaknesses of the two methods under investigation.

> With a criterion score (CS) of less than 0.25 ($CS < 0.25$), less than a quarter of the possible score is achieved. This criterion is a weakness of the evaluation method.

> With a criterion score (CS) of at least 0.75 ($CS \geq 0.75$), at least three-quarters of the possible score are achieved. This criterion is a strength of the evaluation method.

The following five criteria form the weaknesses of the BIK BITV-Test: “C04: Coverage of additional criteria from EN 301 549 Tab. A.1” ($CS = 0.02$), “C08: Optional input formats” ($CS = 0.00$), “C09: Coverage of additional criteria from BITV 2.0” ($CS = 0.00$), “C10: Effort” ($CS = 0.17$), and “C11: Coverage of additional criteria from WCAG 2.1 AAA” ($CS = 0.00$). The following six criteria form the strengths of the evaluation method: “C01: Coverage of WCAG 2.1 AA” ($CS = 1.00$), “C02: Completeness” ($CS = 0.75$), “C05: Correctness” ($CS = 0.94$), “C13: Publicity of the evaluation procedure” ($CS = 1.00$), “C17: Certificate” ($CS = 1.00$), and “C21: Organizational requirements” ($CS = 1.00$). The weaknesses and strengths of the BIK BITV-Test are shown in Table 3.

The weaknesses of the BITV-Audit lie in the following three criteria: “C13: Publicity of the evaluation procedure” ($CS = 0.00$), “C16: License” ($CS =$

0.00), and "C21: Organizational requirements" ($CS =$ 0.00). The following ten criteria are the strengths: "C01: Coverage of WCAG 2.1 AA" ($CS = 1.00$), "C02: Completeness" ($CS = 0.90$), "C03: Quality of the evaluation report" ($CS = 0.79$), "C05: Correctness" ($CS = 0.90$), "C06: Quality of the sample" ($CS = 0.93$), "C08: Optional input formats" ($CS =$ 1.00), "C09: Coverage of additional criteria from BITV 2.0" ($CS = 1.00$), "C11: Coverage of additional criteria from WCAG 2.1 AAA" ($CS = 1.00$), "C14: Scalability with regard to evaluators" ($CS = 0.80$), and "C17: Certificate" ($CS = 0.75$). The weaknesses and strengths of the BITV-Audit are shown in Table 3.

5.2.3 Observed Anomalies on the Websites

In the following, we provide an overview on our findings regarding anomalies that occur depending on the accessibility level of the website.

> If there is a difference ($Diff_A$) of at least 0.10 between the normalized values of a metric regarding the two websites HdM and BfB ($Diff_A \geq 0.10$), there are anomalies in the evaluation method. These anomalies occur depending on the accessibility of the evaluation object.

In the BIK BITV-Test such anomalies could be observed in at least one metric of the following criteria: "C02: Completeness" ($Diff_A = 0.18$; in favor of HdM), "C05: Correctness" ($Diff_A = 0.10$; in favor of HdM), "C03: Quality of the evaluation report" ($Diff_A = 0.13$; in favor of HdM), and "C20: Simplicity of evaluation procedure" ($Diff_A = 0.18$; in favor of HdM). In the BITV-Audit, anomalies were observed in at least one metric of the following criteria: "C05: Correctness" ($Diff_A = 0.21$; in favor of HdM) and "C03: Quality of the evaluation report" ($Diff_A = 0.23$; in favor of BfB).

6 DISCUSSION

6.1 Catalog of Criteria

In the context of this work, a comprehensive, generic criteria catalog was created based on the expertise of diverse accessibility experts. The criteria catalog consists of 22 criteria with a total of 41 metrics and is thus more comprehensive than existing criteria catalogs in related work.

Table 2 lists the criteria and their weights. "C01: Coverage of WCAG 2.1 AA", "C04: Coverage of additional criteria from EN 301 549 Tab A.1", and "C09: Coverage of additional criteria from BITV 2.0" have received high weights in the averaged expert weighting. The requirements of these standards are mandatory for the public sector in Germany. Because the underlying standard is decisive for the results of a conformity check, these criteria could be considered particularly important. Since checking non-web documents is included in the standards, this could explain the high weighting of "C08: Optional input formats". The high values for "C02: Completeness", "C05: Correctness" and "C07: Quality assurance" are not surprising, as this is about the validity of the results. The valid statement about the accessibility of the checked web pages is the goal of a manual evaluation method. The high weight of the "C06: Quality of the sample" can be explained by the fact that the conformity statement of the checked web pages should be representative of the entire website. It is not feasible to evaluate the whole website, so the selection of web pages is very important. The "C10: Effort" is also perceived as an important criterion by the experts. An evaluation that provides ideal results but is not feasible in reality is not a good approach. Furthermore, the aim of a manual evaluation is not only to determine the accessibility of a website but also to contribute to improving accessibility. Therefore, the experts may have found the criterion "C03: Quality of the evaluation report" important.

Finally, the comparison of two testing methods was performed on real data using this set of criteria. The results of the comparison show that various dimensions of the evaluation methods can be captured and quantified using the catalog of criteria.

6.2 Implications of the Study Results

Interesting insights can be gained from the results. One of the most promising is the combination of empirical evaluation and subsequent conformity evaluation. It was shown that conformance-based evaluation ensures that a diverse base of issues can be fully and correctly identified. It is important that the evaluation method covers as many critical standards as possible. However, the results also show that an empirical evaluation approach can further increase completeness. This can identify problems that are not covered by the guidelines but are nevertheless problems that inhibit access for users with disabilities. The combination of empirical evaluation with subsequent conformity evaluation, as used by the BITV-Audit, promises valid results for assessing the accessibility of a website. This approach promises a comprehensive accessibility evaluation, beyond the limits of the standards and puts the human in the center. The combination

Table 3: Weaknesses and strengths of the evaluation methods.

	Weaknesses	Strengths
BIK BITV-Test	- Coverage of additional criteria from EN 301 549 Tab. A.1 - Coverage of additional criteria from BITV 2.0 - Coverage of additional criteria from WCAG 2.1 AAA - Optional input formats - Effort	- Coverage of WCAG 2.1 AA - Completeness - Correctness - Publicity of the evaluation procedure - Certificate - Organizational requirements
BITV-Audit	- Publicity of the evaluation procedure - License - Organizational requirements	- Coverage of WCAG 2.1 AA - Coverage of additional criteria from BITV 2.0 - Coverage of additional criteria from WCAG 2.1 AAA - Completeness - Correctness - Quality of the evaluation report - Quality of the sample - Optional input formats - Scalability with regard to evaluators - Certificate

of conformance testing and empirical evaluation is also to be pursued in the design of WCAG 3.0 (W3C, 2021).

In addition, detailed evaluation reports with screenshots and recommended solutions are very important to create a comprehensive understanding on the part of the reader and thus facilitate the removal of the barriers. It is essential to develop user-friendly tools that efficiently support evaluators in creating evaluation reports.

Furthermore, the interviews unveiled that the amount of training required to become an experienced evaluator should not be underestimated, and conducting evaluations is not trivial. Good training programs and a common exchange between evaluators can have a supporting effect.

The BIK BITV-Test with its tandem procedure has performed better in quality assurance than the BITV-Audit. However, this procedure does not seem to lead to significantly better results in terms of correctness and completeness. Therefore, the tandem procedure in the evaluation could be avoided and a comprehensive quality assurance of the results could compensate for this lack. This would reduce the effort of the evaluation procedure. This has already been implemented in the BIK BITV-Test with the new version (DIAS GmbH, 2020a).

A central and important aspect in the area of manual evaluation methods is the scalability of the evaluation procedures. Websites are becoming more and more complex and the selection of web pages is crucial for the representativity of accessibility for the whole website. However, scalability is also critical to the feasibility of such manual evaluations. The scope has to be reduced to a manageable size and still provide a valid statement about the accessibility of the entire website. The WCAG 3.0 working group is also concerned with this problem. As shown in the expert interviews, collaboration in teams can be helpful to efficiently conduct evaluations of complex websites. For example, experts for different user groups could work together on an evaluation. Above all, the level of tool support must be increased where possible, as well as the automation of processes and evaluation steps. Evaluators require tools which will efficiently support them during performing and documenting evaluations. This should be taken into consideration when designing a new evaluation method.

6.3 Limitations

In looking at results of our study, the following limitations should be considered. The results of this study would probably have been more reliable and generic if these limitations had not been in place.

The identification and weighting of the criteria were done by accessibility experts, but the metrics contributing to the criteria do not have an expert-based weighting. Such weighting could lead to a different result.

In order to keep the effort for the study in a feasible frame, we had to cut down on some evaluation parameters which would have otherwise been performed in a more comprehensive fashion: (1) The study used only two websites as evaluation samples. (2) Only one web page was taken as a sample for each website. For an evaluation that leads to official certification, the evaluator has to select a representative sample of the website. (3) Both websites were evaluated by only one evaluator. (4) Furthermore, both websites were information-oriented. The results could differ for evaluations of complex web applications.

It should also be noted that the evaluator was an employee of T-Systems, the company that owns the BITV-Audit. This could have led to a bias towards this method. On the other hand, the person acting as QA for the BIK BITV-Test, was an employee of HdM.

In the context of this study, primarily the perspective of the evaluators and organizations conducting the evaluations were taken. It could be argued that by incorporating the client's perspective to a greater extend, the results would be more comprehensive.

7 CONCLUSION

In this paper, a systematic comparison of two existing manual evaluation methods was conducted using real-world data involving accessibility experts and two exemplary websites. For this purpose, we created a generic catalog of comparison criteria based on the expertise of various accessibility experts. On the basis of this catalog, we compared two evaluation methods in terms of their suitability and effectiveness: The BIK BITV-Test as one of the best-known conformance-based evaluation methods in Germany and the BITV-Audit of T-Systems MMS as an example of the empiric-based evaluation methods.

In comparison, the BITV-Audit performs better than the BIK BITV-Test based on the defined catalog and the specific weightings determined by the accessibility experts involved in the study. However, it should be noted that no universally valid weighting can be defined for all possible situations. Therefore, if necessary, our weights may be replaced by individual weights and so used to recalculate the comparison for both methods, on a case-by-case basis. Thus, this paper can assist in deciding which evaluation method is more appropriate in a particular situation. Additionally, the discussed weaknesses and strengths of each method can assist in making a decision.

Also, the results show the following major similarities between the two evaluation methods: Both fully cover the WCAG 2.1 success criteria of conformance level A and AA. Furthermore, they have similar values in the criteria of tool support and simplicity of the evaluation procedure. Strong differences are found in the following areas: Coverage of additional criteria from various standards (EN 301 549, WCAG 2.1 AAA, BITV 2.0; in favor of BITV-Audit), the scope of optional input formats (in favor of BITV-Audit), publicity of the evaluation procedure (in favor of BIK BITV-Test), licensing conditions (in favor of BIK BITV-Test), and organizational requirements in order to gain permission to use the evaluation method (in favor of BIK BITV-Test).

Moreover, we have observed that WCAG 2.1 does not consider all usability problems which were identified through expert assessments.

In this work, we carefully defined a catalog of criteria based on 22 criteria and common standards together with experts. Nevertheless, it is conceivable that further criteria and metrics may be of high relevance in the future. The results of this work can always serve as a basis for possible future extensions in this area.

REFERENCES

Brajnik, G. (2005). Accessibility assessments through heuristic walkthroughs. *SIGCHI-Italy*, page 77.

Brajnik, G. (2006). Web accessibility testing: When the method is the culprit. In *International Conference on Computers for Handicapped Persons*, pages 156–163. Springer.

Brajnik, G. (2008). A comparative test of web accessibility evaluation methods. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '08, page 113–120, New York, NY, USA. Association for Computing Machinery.

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, page 131–138, New York, NY, USA. Association for Computing Machinery.

Brajnik, G., Mulas, A., and Pitton, C. (2007). Effects of sampling methods on web accessibility evaluations. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '07, page 59–66, New York, NY, USA. Association for Computing Machinery.

Burkard, A., Zimmermann, G., and Schwarzer, B. (2021). Monitoring systems for checking websites on accessibility. *Frontiers in Computer Science*, 3:2.

Calvo, R., Seyedarabi, F., and Savva, A. (2016). Beyond web content accessibility guidelines: Expert accessibility reviews. In *Proceedings of the 7th International Conference on Software Development and Technolo-*

gies for Enhancing Accessibility and Fighting Info-Exclusion, DSAI 2016, page 77–84, New York, NY, USA. Association for Computing Machinery.

DIAS GmbH (2019). Beschreibung des prüfverfahrens. https://www.bitvtest.de/bitv_test/das_testverfahren_im_detail/verfahren.html. Accessed October 14, 2020.

DIAS GmbH (2020a). Anpassung des bitv-/wcag-test prüfverfahrens 2021. https://www.bitvtest.de/bitv_test/bitv_test_beauftragen/infos_preise/preise_abschliessend.html. Accessed March 04, 2021.

DIAS GmbH (2020b). Bitv-test: Kurzvorstellung. https://www.bitvtest.de/bitv_Test/einfuehrung/kurzvorstellung. Accessed October 14, 2020.

DIAS GmbH (2020c). Informationen und preise. https://www.bitvtest.de/bitv_test/bitv_test_beauftragen/infos_preise.html. Accessed October 14, 2020.

European Parliament and Council of the European Union (2016). Directive (eu) 2016/2102 of the european parliament and of the council of 26 october 2016 on the accessibility of the websites and mobile applications of public sector bodies.

European Parliament and Council of the European Union (2018). Commission implementing decision (eu) 2018/2048 of 20 december 2018 on the harmonised standard for websites and mobile applications drafted in support of directive (eu) 2016/2102 of the european parliament and of the council.

European Telecommunications Standards Institute (2018). En 301 549 v2.1.2 - accessibility requirements for ict products and services.

European Telecommunications Standards Institute (2020). Draft - en 301 549 v3.2.1 - accessibility requirements for ict products and services.

Freire, A. P. (2012). *Disabled people and the Web: User-based measurement of accessibility*. PhD thesis, University of York.

German Federal Ministry of Labour and Social Affairs (2019). Verordnung zur schaffung barrierefreier informationstechnik nach dem behindertengleichstellungsgesetz (barrierefreie-informationstechnik-verordnung - bitv 2.0).

Lang, T. (2004). Comparing website accessibility evaluation methods and learnings from usability evaluation methods.

Mankoff, J., Fait, H., and Tran, T. (2005). Is your web page accessible? a comparative study of methods for assessing web page accessibility for the blind. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, page 41–50, New York, NY, USA. Association for Computing Machinery.

Nebe, A.-M. (2021). Bitv audit. https://gpii.eu/bfit/ag03/handreichung/pruefverfahren/bitv-audit.html. Accessed May 16, 2021.

Rømen, D. and Svanæs, D. (2008). Evaluating web site accessibility: Validating the wai guidelines through usability testing with disabled users. In *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, NordiCHI '08, page 535–538, New York, NY, USA. Association for Computing Machinery.

Statistisches Bundesamt (Destatis) (2020). 7,9 millionen schwerbehinderte menschen leben in deutschland.

T-Systems Multimedia Solutions GmbH (2020a). Bitv-prüfmethodik - dakks akkreditierter testprozess expertenevaluation. Internal document viewed on October 14, 2020.

T-Systems Multimedia Solutions GmbH (2020b). Verfahrensbeschreibung. Internal document viewed on October 14, 2020.

Velleman, E. and van der Geest, T. (2013). Page sample size in web accessibility testing: How many pages is enough? In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, New York, NY, USA. Association for Computing Machinery.

Vigo, M., Brown, J., and 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*, pages 1–10.

W3C (2018). Web content accessibility guidelines (wcag) 2.1. https://www.w3.org/TR/WCAG21/. Accessed October 27, 2020.

W3C (2021). W3c editor's draft 22 march 2021 - w3c accessibility guidelines (wcag) 3.0. https://w3c.github.io/silver/guidelines. Accessed April 08, 2021.

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, page 203–210, New York, NY, USA. Association for Computing Machinery.