Exploring Mobile Device Accessibility: Challenges, Insights, and Recommendations for Evaluation Methodologies

Letícia Seixas Pereira
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
lspereira@ciencias.ulisboa.pt

Maria Matos
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
mgmatos@lasige.di.fc.ul.pt

Carlos Duarte
LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal
caduarte@edu.ulisboa.pt

ABSTRACT

With the ubiquitous use of mobile applications, it is paramount that they are accessible, so they can empower all users, including those with different needs. Determining if an app is accessible implies conducting an accessibility evaluation. While accessibility evaluations have been thoroughly studied in the web domain, there are still many open questions when evaluating mobile applications. This paper investigates mobile accessibility evaluation methodologies. We conducted four studies, including an examination of accessibility reports from European Member-states, interviews with accessibility experts, manual evaluations, and usability tests involving users. Our investigations have uncovered significant limitations in current evaluation methods, suggesting that the absence of authoritative guidelines and standards, similar to what exists for the web, but tailored specifically to mobile devices, hampers the effectiveness of accessibility evaluation and monitoring activities. Based on our findings, we present a set of recommendations aimed at improving the evaluation methodologies for assessing mobile applications' accessibility.

CCS CONCEPTS

• **Human-centered computing** → Accessibility; Accessibility design and evaluation methods.

KEYWORDS

Accessibility, Mobile accessibility evaluation, Accessibility evaluation methodologies, mobile applications

ACM Reference Format:
Letícia Seixas Pereira, Maria Matos, and Carlos Duarte. 2024. Exploring Mobile Device Accessibility: Challenges, Insights, and Recommendations for Evaluation Methodologies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642526

1 INTRODUCTION

Mobile devices have become an indispensable part of our daily lives while introducing specific challenges such as their limited display size, unique interaction methods, diverse usage contexts, and lack of platform consistency [19]. These challenges can also impact the usage of people with disabilities. Failing to meet accessibility requirements in the content delivered on mobile devices not only hinders equal opportunities but also restricts the full participation of individuals with disabilities in society, thereby limiting their independence and overall well-being. Assessing accessibility, particularly in the context of mobile devices, poses distinct challenges requiring specialized methodologies and approaches [19].

Significant progress has been made globally in promoting the rights of individuals with disabilities in digital environments. The widespread adoption of the United Nations Convention on the Rights of Persons with Disabilities (CRPD) [52] amplified awareness and commitment to digital accessibility. Web Content Accessibility Guidelines (WCAG) [22] played a crucial role in fostering harmonization of accessibility standards. These guidelines have been embraced as standards in various regions, including Europe and the United States. Within the European Union (EU), the implementation of the Web Accessibility Directive (WAD) [53] includes a monitoring process that conducts comprehensive assessments and evaluations to examine the accessibility status of websites, digital content, and ICT (Information and Communication Technology) products and services, including mobile applications.

This research aligns with recent initiatives aimed at enhancing accessibility, specifically within the domain of mobile applications (apps). Our aim is to explore this scenario, with a focus on identifying the main challenges and opportunities within current methodologies and practices. Accordingly, we have formulated the following research questions:

- What are the current practices used by evaluators in mobile accessibility, and how do they impact the outcomes of evaluations?
- How can current methodologies for evaluating mobile accessibility be enhanced?

To accomplish this, four distinct studies were conducted. We began by analyzing 26 reports on the outcomes of monitoring activities, published during the initial monitoring period, as mandated by the WAD. These reports provide valuable insights into the current state of accessibility in the European public sector and include data from various accessibility evaluations conducted in diverse contexts. Our analysis revealed a lack of consistency in how countries evaluate mobile applications, particularly in the (limited) inclusion of user testing. Since a methodology combines various techniques, this absence of a standardized approach poses a significant challenge. Varying methodologies can result in different outcomes, hindering the assurance of quality and comparability in the monitoring process.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642526

The second study consisted of interviewing five accessibility evaluators to understand their tasks when assessing mobile applications. Through these interviews, two significant challenges emerged: the absence of authoritative evaluation standards and guidance tailored for mobile applications, and the lack of automated tools to assist their work. These challenges result in a labor-intensive process susceptible to errors. The absence of clear guidance requires additional time to interpret and apply the requirements current in place. Additionally, the limited availability of automated tools demands the manual execution of various evaluation steps.

Subsequently, we conducted two additional studies to gather different perspectives through practical application. First, we conducted a manual evaluation of four mobile applications. This assessment integrated insights from the prior studies, and information obtained from the support documentation provided by standardization bodies. We applied a methodology informed by both the WCAG-EM and the European Standard (EN) Methodology, as also employed by the interviewed accessibility experts. This approach enabled a more thorough exploration of the issues highlighted by evaluators, providing a comprehensive grasp of current methodologies and their limitations. Throughout this process, we corroborated the findings derived from the conducted interviews and identified further challenges on interpreting and applying current standards. Subsequently, we discuss three crucial areas requiring further refinement and guidance: sample selection for evaluation, assessment of this sample, and results reporting.

The final study involved conducting user tests with people with disabilities. Its main goals were to gather participant feedback on main accessibility issues in mobile applications, evaluate coverage of current guidelines and standards, and identify existing methodological challenges.

Our results suggest that finding a harmonized evaluation methodology specifically tailored for mobile devices and the absence of clear and authoritative guidelines for evaluating mobile accessibility poses obstacles in this domain. This is particularly challenging as evaluating accessibility is essential not only to identify barriers and improve accessibility of a product or service, but also to monitor progress. Without harmonized approaches, the lack of comparability in results hinders effective monitoring efforts, making it difficult to assess overall accessibility trends and make informed decisions for further enhancements. User testing also represents a gap in this process, as many accessibility issues only surface during user interactions. Moreover, the absence of clear guidance on how to conduct this process may contribute to the low adherence in such assessments.

Based on the insights derived from this research, recommendations are proposed to enhance current methodologies for conducting accessibility evaluations for mobile applications. While previous work has explored the accessibility of mobile applications from different perspectives, this paper aims to provide a broader perspective, both in terms of accessibility criteria evaluated and in terms of methodologies. Our primary **contributions** include the following:

- A report on the existing challenges within methodologies for conducting accessibility evaluations of mobile applications.
- A discussion on the extent to which main accessibility standards and guidelines address accessibility in mobile applications.
- Recommendations for refining methodologies for assessing accessibility of mobile applications.

2 RELATED WORK

Ensuring accessibility involves several critical components and processes, including compatibility with assistive technologies, integrating accessibility considerations throughout the design and development cycles, compliance with accessibility requirements, and more [1]. These measures are essential for enhancing the product's usability across diverse user profiles. This research, however, specifically emphasizes the evaluation phase, entailing an assessment of the product's effectiveness for a broad user base, whether in development or already completed. In this section, we address three key topics pertinent to this context: 1) techniques and tools for evaluation support or execution; 2) existing standards, guidelines, and recommendations; and 3) current methodologies for assessing accessibility.

2.1 Accessibility testing techniques and tools

Testing techniques are essential for assessing accessibility, with three primary types commonly used: automated, manual, and user testing [1]. Automated testing employs specialized tools to analyze code, identifying potential accessibility issues based on specific criteria. While efficient, it may not catch all problems [27, 28]. Hence, additional methods are vital for ensuring accessibility. Manual testing involves experts thoroughly examining websites or applications to uncover issues, which automated tools might miss [27]. User testing provides valuable insights into user experiences, revealing accessibility barriers and deepening our understanding of user challenges [27, 28], leading to more effective improvements. By combining these techniques, comprehensive evaluation methodologies can address a wide range of barriers.

Ensuring mobile accessibility is challenging also due to diverse devices impacting content display. Varying sizes and interaction methods add complexity. The absence of a constant keyboard limits input commands, while gestures add complexity. Assessing mobile accessibility requires then a comprehensive understanding of factors such as user capabilities, device characteristics, interface design, app functionality, and the context of use [19]. Silva et al. [16] assessed tool support for mobile accessibility and categorized automated tools into static and dynamic analysis. Static analysis like Android Lint [54] focuses on source code but may miss runtime flaws. Dynamic evaluation, involving manual exploration, is adopted by tools like IBM Equal Access Accessibility Checker [21], Accessibility Scanner [55], and Accessibility Inspector [56]. However, manually exploring complex apps can be laborious and error-prone, especially for frequently updated apps.

Efforts in the literature aim to develop more efficient tools for automated accessibility analysis in mobile apps. For instance, Salehnamadi et al. [34] introduced A11yPuppetry, a semi-automated record-and-replay technique for Android apps using TalkBack. They also proposed Latte [33], employing Use-Case Specifications and Android's Switch Access. Groundhog [35], a tool replicating interaction patterns of users with disabilities, operates independently from testing frameworks. Kashif et al. [23] focused on automated tools for mobile accessibility evaluation during design and prototyping. Park et al. [30] proposed a tool for evaluating alternative texts for images in Android apps. Additionally, Alotaibi et al. [5] developed

an automated technique for detecting barriers in interactions mediated by TalkBack. Eler et al. [16] created MATE, exploring apps while checking for accessibility issues related to visual impairments.

Automated techniques are crucial for identifying accessibility issues. However, no single approach offers a complete assessment. Our research aims to explore various techniques, discussing their strengths and limitations. We emphasize the complementary nature of these methods, aiming to contribute to the ongoing discussion on improving current methodologies for a more comprehensive assessment of accessibility.

2.2 Standards, Guidelines and Recommendations

Accessibility standards are integral in guiding the creation of content that is usable for individuals with disabilities. They help determine the necessary accessibility provisions and evaluate the level of compliance with these standards [1].

Worldwide, WCAG forms the bedrock of legal standards for web accessibility. In Europe, it is reflected in EN 301 549 [17], while in the United States, it is embedded within Section 508 [57]. Many regions have adopted WCAG as the cornerstone of their web accessibility standards [40]. While still focused on web content, WCAG has expanded its criteria to address the challenges posed by the increasing use of smartphones.

To better address the mobile context, the Web Accessibility Initiative (WAI) introduced initiatives like the Mobile Web Best Practices (MWBP) [58] and the Guidance on Applying WCAG 2.0 to Non-Web Information and Communications Technologies (WCAG2ICT) [41]. Notably, WCAG 2.1 introduced 17 new success criteria, specifically targeting accessibility for users with low vision, cognitive and learning disabilities, as well as mobile accessibility [42]. The recent WCAG 2.2 draft proposes 9 additional success criteria, aimed at enhancing user experiences by ensuring easier access to support resources and compatibility with a wider range of assistive technologies [43]. Of significance, two of the new criteria pertain to mobile devices, addressing movements for drag and drop actions, and emphasizing target size to prevent inadvertent clicks on adjacent buttons due to limited spacing.

It is noteworthy that other initiatives have emerged to expedite the development of guidelines tailored to the mobile context. For instance, in 2012, Funka published a document based on WCAG 2.0 containing guidelines for the development of accessible mobile interfaces [59]. Additionally, The BBC Mobile Accessibility Guidelines [60] present a set of technology-agnostic best practices applicable to mobile web content, hybrid apps, and native apps. The Mobile Accessibility Checklist [50], developed by the Mozilla MDN Web Docs project, also serves as a practical resource for ensuring mobile accessibility and is intended to be continuously updated to incorporate emerging patterns.

Given the unique opportunities and limitations presented by mobile applications, it is imperative to adopt new approaches for evaluating their accessibility and usability [7]. Despite the significant improvements in mobile accessibility coverage in WCAG 2.1 compared to its predecessor, WCAG 2.0, adherence to these accessibility requirements by developers, content authors, and service providers remains inconsistent [4]. Furthermore, there are several issues that WCAG 2.1 does not yet address. Alajarmeh et al. [4] highlighted the insufficient conformance levels for many success criteria related to moderate and severe accessibility issues faced by users who are blind or visually impaired. To rectify this, one crucial step is to enhance the adoption and effectiveness of automated evaluation. This can be achieved by broadening the range of accessibility guidelines that can be supported by automated methods [37].

For example, Silva et al. [37] noted that BBC's guidelines are less subjective and more amenable to testing, but they do not provide explicit details on how automated evaluation can be conducted. Siebra et al. [36] find that current popular tools do not sufficiently support the evaluation of the implementation of various accessibility requirements in mobile applications. In their analysis of mainstream automated tools, Silva et al. [37] reveal that only approximately 13% of accessibility guidelines, based on BBC's and W3C's standards, are covered by all the tools combined.

The automated evaluation of mobile accessibility remains highly limited due to the small subset of accessibility guidelines that current tools can address [37]. The scarcity of automated tools for mobile accessibility checking can be attributed to the relative newness of the field and the lack of awareness among developers regarding the specific accessibility requirements their apps should meet, as well as the limited understanding of how existing automated tools can assist developers in satisfying these requirements [37].

Conducting conformance testing can pose challenges for both evaluators and developers. It often becomes time-consuming and intricate when issues fall outside established guidelines, necessitating thorough analysis and interpretation by an experienced consultant to establish connections [15]. Nevertheless, there are tangible advantages to reporting issues within the framework of these standards. Doing so assists technical teams, including developers and managers, in gaining a deeper comprehension of why a specific issue hinders accessibility [15]. For instance, adherence to the harmonized technical standard EN 301 549 supports a common understanding of the term 'accessible' in this context [53]. This study will focus on assessing accessibility within the context of WCAG and EN 301 549. We will explore limitations and offer insights for further enhancements.

2.3 Methodologies used for accessibility evaluation

Evaluation methodologies are crucial for assessing and monitoring the accessibility of websites and mobile applications. They provide structured, standardized approaches, ensuring consistent and comparable results. In this section, we explore some widely used methodologies for evaluating accessibility.

The Web Accessibility Directive [17], initiated by the European Commission, is an effort aimed at creating an inclusive Europe accessible to all. Its objective is to enhance the usability of websites and mobile applications of public services for people with disabilities by ensuring compliance with the EN 301 549 [13]. Member States are required to monitor compliance using the methodology specified by the Commission. The European Standard (EN) Methodology outlines the frequency of monitoring, sampling procedures for web pages and mobile applications, provisions for automated, manual, and usability testing, guidelines for determining compliance, and a mechanism to support public sector bodies in addressing identified deficiencies. The directive introduces two evaluation

processes: simplified testing and detailed testing. Simplified testing involves automated tools and manual checks to assess a small section of website or app. Detailed testing provides a more comprehensive examination, testing against WCAG success criteria using assistive technology, and combining manual and automated methods.

WCAG-EM (Website Accessibility Conformance Evaluation Methodology) [23] is widely used for supporting the evaluation of websites. It provides a structured approach for determining compliance with WCAG standards. Trusted Tester [61] is another established methodology for evaluating web accessibility, ensuring compliance with Section 508 requirements. The IBM Equal Access Toolkit [62] offers a comprehensive methodology for assessing web content, providing clear and concise guidance throughout the development process.

In recent years, there has been a growing focus on methodologies for evaluating mobile applications. Appt-EM [51], derived from WCAG-EM, is designed for assessing mobile applications. Appt-EM applies 31 out of 50 WCAG 2.1 success criteria, excluding six and adjusting definitions for 13. It also provides valuable recommendations, such as contextual information during evaluation, capturing screenshots of screens with errors, and structuring reports based on success criteria and screens.

Furthermore, Google and Apple have developed their own methodologies for testing accessibility in their respective mobile systems [63, 64]. Android offers a methodology for developers to test an application's accessibility by experiencing it from a user's perspective. Google recommends a combination of manual testing, analytic tools, automated testing, and user testing for comprehensive results. Apple focuses on ensuring users can complete critical tasks in the app, regardless of how they interact with their devices. Testing critical user processes with accessibility features turned on provides insights into potential difficulties and areas for improvement. Apple suggests enabling features like VoiceOver, Reduce Motion, or Large Text Size during testing.

Various methodologies have emerged in literature. Billi et al. [7] introduced a two-step approach: first, evaluating accessibility early on to guide developers, and then, assessing usability after addressing accessibility issues. They recommended at least three evaluators for diverse perspectives, incorporating WCAG 1.0 [14], WCAG 2.0 (working draft at the time), and User Agents Accessibility Guidelines (UAAG) [39]. Mobile-specific usability heuristics were also integrated with accessibility guidelines. Acosta-Vargas et al. [3] proposed a manual testing method combining WCAG 2.1 with the Accessibility Scanner tool, emphasizing the need for multiple evaluation methods and considering diverse user abilities and scenarios. In a subsequent work [2], they combined automatic and manual reviews based on WCAG 2.1 guidelines. Silva et al. [38] utilized an observation method for assessing app accessibility for visually impaired users, involving participant selection, task definition, case study implementation, and results analysis. Joshi et al. [25] suggested a straightforward methodology encompassing quick automated testing, screen reader testing, magnification/zooming testing, and switch access and keyboard testing, accessible to any member of the development team. Finally, Mateus et al. [28] conducted a study comparing evaluation methodologies, underscoring the importance of user testing alongside automated techniques.

While research on mobile accessibility has grown significantly, many studies maintain narrow foci, often limited to specific user groups or operating systems. Additionally, some methodologies discussed here serve as interim steps, involving tool evaluations or gathering user feedback on accessibility barriers. This paper aims to provide a perspective on conducting comprehensive accessibility evaluations.

3 STUDY 1: MONITORING REPORTS

In the European Commission's monitoring methodology [17], Member States generate reports to document the results of accessibility evaluations conducted on websites and mobile applications [49]. These reports provide insights into the shortcomings, challenges, and findings encountered during the evaluation process and also serve as a valuable resource to identify the different methods, strategies and tools used by them. By analyzing these reports, we aim to gain a comprehensive understanding of the mobile evaluation landscape, uncover areas for improvement, and learn from the experiences and approaches shared by different Member States.

3.1 Method

We employed content analysis [29] to examine reports that document the outcomes of the Member States monitoring activities. Among the 27 Member States, plus the UK, we examined 26 reports. It is important to note that, at the time of the analysis, two Member States had not submitted their reports, and another had not yet conducted mobile application evaluations. This analysis centered on extracting relevant data from the reports, including the methods employed by each country for evaluating mobile applications, the quantity of apps assessed by each, along with challenges or observations documented in these reports.

3.2 Findings

In the following, we describe the main findings obtained from the analysis of the reports regarding accessibility monitoring in Europe.

3.2.1 Sample size and operating systems. In accordance with the methodology outlined by the European Commission, the sample size of mobile applications included in the evaluation of each country should be proportionate to its population size, encompassing six applications plus one additional application per million inhabitants. Considering the number of mobile applications assessed in countries that have conducted mobile evaluations and submitted reports, the average stands at 13 mobile applications evaluated per country.

Concerning operating systems, the reports highlighted the distinction between Android and iOS. Out of a total of 366 mobile applications assessed, the majority were Android applications (n=213). Some countries expressed their choices in their reports, citing reasons such as not evaluating applications for iOS due to its limited usage among people with disabilities in the country or because their official language is incompatible with the screen reader used for iOS. Two Member States omitted specific counts of operating system, providing only the total number of apps evaluated. One country deviated from the norm by abstaining from the evaluation of any applications for Android, yet they did not provide an explanation for this decision in their report.

3.2.2 Evaluation methods applied to evaluate mobile applications. Concerning the assessment of mobile applications, detailed testing, i.e., manual evaluations, emerged as the predominant method among the countries (n=24). It's important to note that in the context of the WAD, detailed testing is the only method required for monitoring the accessibility of mobile applications. To a lesser extent, some countries also conducted simplified testing (n=9), i.e., integrated automated testing, using tools such as Accessibility Scanner, Evinced tool, and Accessibility Insights, while only a limited number incorporated user testing as part of their monitoring activities (n=3) – both methods defined as optional. Furthermore, most reports do not specify how many participants were involved in these tests; only one country provides a total count, with Portugal including 4 individuals with disabilities.

This research primarily focuses on mobile evaluation; however, these reports also encompass evaluations conducted on websites. To provide a clearer understanding of the scope limitations, in comparison to the numbers derived from website assessments, manual and user testing reveal little disparity. Yet, when automated testing is compared, the figures present a significant contrast. While only 9 countries utilized automated tests for mobile applications, all Member States employed automated testing for websites.

Moreover, there was no discernible pattern in the techniques employed by each country, as evident in prior data. Considering that a methodology is formed through the combination of various techniques, this absence of a standardized methodology poses a notable challenge. Divergent methodologies can yield disparate results, thereby hindering the assurance of quality and comparability in the monitoring process. Finally, it is concerning to observe the limited number of user tests conducted in these evaluations.

3.2.3 Challenges in accessibility compliance and monitoring. The primary challenge identified from this analysis relates to the low level of accessibility found in the mobile applications evaluated and reported by the Member States. It was observed that some key success criteria were frequently missed by these applications, with some of them being highlighted as the criteria that most often failed in the evaluated apps. For example, criteria 1.3.1 and 4.1.2 were noted in 12 reports, while criteria 1.1.1 and 1.4.3 appeared in 11 reports. Additionally, criteria 1.3.4 and 1.4.11 were found in 7 reports.

It's important to note that not all reports provided this level of detailed information. Some only reported the final conformity result without specifying which criteria passed or failed. Further evidencing the current lack of commitment to accessibility, there was a notable scarcity of accessibility statements in these apps – despite their requirement by the WAD. Out of the 366 applications evaluated, only 8 had an accessibility statement.

Taking a broader perspective, based on the analysis of all the reports, several challenges were also identified within the scope of the monitoring process itself. As mentioned, not all reports included detailed information about the evaluations conducted. While the WAD explicitly provides a guideline for the development of these reports, including sections such as executive summary, background about the evaluation, scope of review, review process, results and recommended actions, references, and appendices, these guidelines often leave room for the omission of detailed information. This results in each Member State defining its own criteria for providing such details. These findings highlight significant areas for improvement both in the applications' accessibility features and in the overall approach to monitoring and evaluation.

3.3 Discussion

The first study conducted in the scope of this research involved the analysis of reports created by EU Member States for their first monitoring exercise. This analysis aimed to better understand the mobile evaluation landscape through the examination of 26 reports, identifying areas for improvement and learning from diverse approaches.

Our findings revealed that **the majority of evaluated apps were Android applications**, with **manual evaluations being the most employed** method among the countries, while **user testing was less frequently conducted**. Furthermore, a **lack of accessibility statements** was noted in most of these apps.

It's noteworthy that detailed information was gathered from only a limited number of reports due to **the inconsistent levels of detail provided**. Furthermore, there is a **discrepancy in the level of detail between evaluations conducted in websites and apps**, highlighting differing priorities. Lastly, despite guidance on the methodology to be followed and the development of the final report, it is still possible to observe that while the reports often mention the methods used (such as detailed testing or simplified methods), there is a **lack of comprehensive information on how these evaluations were conducted**. This deficiency not only hinders our analysis but also hampers a deeper understanding of the reliability of the reported results. Moreover, from the perspective of advancing the field, it also becomes challenging to propose improvements due to the difficulty in identifying any challenges that may have arisen during the evaluations.

4 STUDY 2: INTERVIEW WITH EVALUATORS

Building upon the insights gained from the monitoring reports, this phase of the study aimed to further investigate these findings by connecting with accessibility experts. We conducted semi-structured interviews with professionals involved in assessing the accessibility of mobile applications. Our goal was to better understand their perspectives about the primary challenges they encounter. Additionally, we aimed to explore the methods and tools they use in the evaluation process and the reasons behind their choices. This study was conducted with approval from our University's Ethics Committee.

4.1 Method

4.1.1 Participants. Five accessibility evaluators from Europe were recruited through the research team's professional network, prioritizing people in different locations to achieve a more diverse sample, as detailed in Table 1. Participation's single requirement was having previous experience in evaluating mobile applications. Among the participants, including two blind people, all had experience levels ranging from 1 to 9 years, with an average of three years. They also had experience with both iOS and Android, except one of them who had previous experience only evaluating Android mobile applications.

4.1.2 Procedure. The recruitment of participants involved initially reaching out to them through a first email, which provided a brief

Table 1: Interview participants demographic data.

Participant	Country	Type of disability	Years of experience with mobile accessibility evaluation	Operating systems used in mobile accessibility evaluation
PE1	Portugal	Blindness	9	iOS and Android
PE2	Denmark	Blindness	3	iOS and Android
PE3	Norway	-	1	iOS and Android
PE4	Portugal	-	1	Android
PE5	Netherlands	-	2	iOS and Android

introduction to the research. Once their initial interest was confirmed, participants were sent detailed information about the study, along with an Informed Consent Form to review at their own pace. To ensure convenience, suitable interview times were arranged, taking into consideration the availability of both the researcher and the participant. The interviews were conducted via the Zoom platform, with most participants granting permission to record. However, in one case where authorization was not given, the researcher relied on taking notes.

Each interview began with an introduction to the study, followed by addressing any questions or concerns that participants may have had. Participants were then asked questions related to their experience in evaluating mobile applications and encouraged to provide feedback on relevant aspects of the study. These questions included challenges faced when conducting evaluations, methodologies employed, tools used, opinions on the ideal methodology, and suggestions for improving this process. Additionally, participants were invited to share any additional insights they deemed relevant.

4.1.3 Data analysis. In our analysis of the interview data, we followed the guidelines for reflexive thematic analysis provided by Braun and Clarke [10, 11]. We integrated both inductive and deductive coding approaches. Inductive coding was used to generate initial themes. The deductive aspect focused on gaining specific insights, particularly around evaluators' challenges, their methodological perspectives, and the tools used in evaluations. The initial coding was manually conducted by a principal coder and then refined in regular team discussions. In the initial sessions, the discussion centered around the identification of themes and verification of codes. In follow up sessions, we focused on the nuanced interpretation of themes. In between sessions, the principal coder reflected these outcomes in the analysis. The process continued until the team agreed on a common understanding that respected the complexity of our data.

4.2 Findings

From the analysis conducted, three main themes were generated regarding the lack of clear guidance for mobile accessibility assessment, automation of methodology's process, and challenges in conveying errors to developers in an effective way.

4.2.1 Lack of clear guidance for mobile accessibility assessment. A major challenge faced by evaluators is the lack of methodologies tailored for mobile devices (PE3, PE4: "as it is not adapted, as it was not made on purpose for mobile applications, there are some things that do not make so much sense"). They note that existing documents primarily focus on web accessibility, leading to varying interpretations among evaluators (PE3). Evaluators feel there is a dearth of information (PE1: "There is no specific thing for native apps") and examples when it comes to evaluating apps (PE3). This leads to the scenario observed where some evaluators developed their own methodologies based on existing ones (PE5: "so we are testing 44 criteria and besides that we rewrote the WCAG-EM and did some pointers as well"), while others have created their methodologies from scratch (PE2: "we've actually built out our own methodology"). In an EU context, according to them, there should be a common methodology applicable to all countries (PE3). This is especially important considering the implementation of the WAD and its aim to harmonize accessibility compliance.

However, despite using their own methodologies, evaluators are unsure if they are following the correct approach to assess certain requirements on a mobile context (PE1: "I can use my methodology all right, but for what it's worth, there is no standard") due to the absence of concrete documentation (PE1: "what applies to mobile? There is no specific thing for native applications"). They also highlight discrepancies between the European Standard and the WCAG (PE3), emphasizing that user needs differ across the two documents (PE3). Ensuring an accurate interpretation of the standard and the WCAG is deemed crucial (PE5: "I think you should first start with EN standard with a good interpretation of WCAG because otherwise it doesn't make sense"). Some evaluators argue that adopting the user's perspective can provide valuable insights during the evaluation process (PE2: "what we have to do then is to put ourselves in the place of the user, which is in principle really, really good because what we are evaluating is not so much compliance but more the actual user experience").

Finally, some challenges arising from this scenario were also mentioned, such as the differences between operating systems, both in terms of the criteria imposed by each system (PE1: "I think Apple does something better, which is that when they make an application, they have to meet certain criteria") and the availability of automatic tools (PE2: "I mean they are only available for Android primarily"). Participants also mentioned the advantage of having access to the code to identify issues that may go unnoticed when relying solely on the app (PE3).

4.2.2 Automation of methodology's processes. Given the existing issues with current methodologies, there are emerging ideas regarding the automation of the evaluation processes (PE4: "Ideally it would be all automatic, that's impossible, but automating as much as possible"). This automation is seen as a positive step to reduce time and resources (PE3), even if the evaluation process remains semi-automatic and certain criteria need manual assessment (PE3). To automate the methodology processes, various tools can

be utilized, including both automatic and support tools, as they assist in the evaluation process. However, evaluators agree that there is a scarcity of tools available for assessing mobile application accessibility. Automatic tools are either non-existent (PE3) or they only address specific issues, and the results may not always be accurate (PE1: "sometimes you find things that turn out to be right"). Nonetheless, evaluators express their desire for an all-encompassing automatic tool (PE2: "I mean anyone would wish that there was an automatic tool that could do anything"), or at the very least, a tool capable of detecting errors that are not always apparent during manual validation (PE1: "Detect errors that are not so noticeable in manual validation"). Support tools are either underutilized or have limited scope, with the common mention being the color contrast analyzer (PE2: "we use color contrast checkers"). Another idea proposed by evaluators is the potential use of Artificial Intelligence (PE2: "I also believe that we can use the power of AI to some extent, but still, it will require manual assessments").

4.2.3 Challenges in conveying errors to developers in an effective way. From a practical perspective, evaluators also highlighted prevalent accessibility issues persisting in mobile applications. These included the absence of labels on elements (PE1: "Labels, labels of the fields and the buttons"), inadequate contrast between elements (PE5: "about contrast elements, contrast of text"), challenges related to application navigation (PE4: "the navigation, the fact that they are adapted from web pages makes the navigation a lot worse"), and difficulties encountered when using a screen reader (PE5: "The focus order for the screen reader that goes wrong quite often").

Given this challenging scenario, evaluators also raised concerns about effectively communicating these issues to developers. They argue that if developers lack knowledge of accessibility, comprehending the reports and the information contained within them could also be challenging (PE2: "typically the recipient of an evaluation report like this, they don't have knowledge enough about, you know, what the reasons would be for that symptom"). Furthermore, evaluators advocated for a shift in the methodology's focus, emphasizing that it should not solely concentrate on issue identification but should also be oriented toward problem-solving (PE5: "it's all about WCAG and finding the issues. And I think that's wrong because it's about fixing the issues").

4.3 Discussion

In this study, we conducted in-depth interviews with five professionals specialized in the assessment of mobile application accessibility. Our primary aim was to gain nuanced insights into the challenges they face, their methodological approaches, and the tools they rely on. One primary challenge that emerged was the **absence of clear mobile-specific guidelines**, leading to varied interpretations among evaluators, with some of them developing their own assessment methodologies. Additionally, differences among mobile operating systems, versions, and devices add complexity, exacerbated by the limited access to source code of these apps. **Tool scarcity for assessing mobile application accessibility** was also a recurrent issue. Automated options are limited, and support tools have narrow scopes. Nevertheless, there was consensus among professionals on the potential benefits of automation, including the use of AI. Despite ongoing efforts, accessibility issues persist in mobile applications. **Communicating these challenges effectively to developers** remains challenging due to their limited knowledge of accessibility practices. Some evaluators highlight the need for a shift in current approaches to prioritize problem-solving alongside issue identification, aiming for more comprehensive advances in mobile application accessibility assessment.

5 STUDY 3: MANUAL EVALUATION

During interviews with evaluators, we observed the challenges they currently encounter when assessing mobile applications, especially concerning the application and interpretation of existing standards such as WCAG and EN 301 549. To gain a more comprehensive understanding of these challenges, we followed the same process they reported. We conducted manual evaluations of mobile applications, adhering to these standards while adapting them for a mobile context. Our objective was to gain a deeper insight into the challenges of this process, to provide a more thorough assessment of its performance, including both strengths and weaknesses.

5.1 Method

5.1.1 Mobile applications selection. We conducted an initial screening of a diverse sample of mobile applications, focusing on four key domains: 1) public services (e.g. transportation, healthcare), 2) public entities (e.g. government agencies, municipal offices), 3) museums and cultural entities, and 4) social networks. Our aim was to include applications with varied elements and interaction patterns.

To ensure a thorough assessment of all WCAG 2.1 success criteria, we focused on identifying the most common errors and barriers encountered by people with disabilities when using mobile apps, as reported in prior studies [6, 12, 28, 32, 38], monitoring reports, and interviews. Barriers include absence of subtitles in videos, inadequate alternative text for images, a lack of section headings to organize content, small font size or inappropriate font choices, low color contrast, difficulties in accessing specific information or navigating using the keyboard, auto-playing videos, small-sized buttons and click areas, inaccessible Captchas, and limited ability to zoom in on elements.

Taking this list into consideration, selection criteria favored apps that contained elements commonly associated with these accessibility errors. The goal was to include a variety of elements in our sample, ensuring that we covered the full spectrum of potential problems rather than duplicating the same issues. Furthermore, we prioritized apps known for common usage or relevance in the country within these domains. Our research focused on apps developed in the same country as the research, facilitating a deeper understanding of the context in which these apps are used. Additionally, apps were required to be available on both iOS and Android platforms.

Based on that, we initially inspected 12 mobile applications, selecting eight (both Android and iOS) for detailed study. The final selection was completed when we observed repetition of the issues, and no new errors emerged. Detailed app features and accessibility issues are provided in the appendices.

5.1.2 Procedure. To assess the selected mobile applications through manual evaluation, a methodology based on WCAG-EM

and the EN 301 549, in particular the requirements for mobile applications, was defined and employed. While WCAG-EM primarily addresses websites assessments, interviews with evaluators revealed their practice of adapting it for mobile applications, adjusting its structure and specifics. Following this adaptation approach, we employed WCAG-EM while tailoring it for mobile contexts, as delineated in the following steps. It is important to emphasize that due to the challenge in finding explicit directives for certain success criteria within a mobile context, we referred to the recommendations provided by the Appt-EM to effectively adapt WCAG-EM. The following steps describe the methodology followed:

1. **Define the equipment to use:** To cover the two primary mobile operating systems used by users, an iPhone and an Android smartphone were employed for running the applications. In the tests, the respective screen readers for each operating system – VoiceOver for iPhone and TalkBack for Android – were utilized, alongside a Bluetooth keyboard.
2. **Define the screen sample:** Given the WCAG-EM's specific focus on websites, our approach primarily relied upon EN 301 549, which outlines the criteria for selecting screen samples. To effectively evaluate the accessibility of the application's content, considering its core purpose and the typical user interaction pathways, we determined that particular attention should be given to the accessibility of the following screens, where applicable: home screen, login screen, site map, contact, help and terms and conditions screens, at least one screen for each type of service, and other for primary uses, accessibility statement screen, feedback mechanism screen, any distinctive screen or content, one downloadable document for each type of service provided, any other screen considered relevant, and randomly selected screens – corresponding to 10% of the sample already established up to this point. Finally, if any of the screens selected correspond to a stage in a process, all the screens in the process must be included.
3. **Define success criteria to be evaluated:** A total of 43 requirements were assessed across each application. However, certain WCAG criteria were excluded from evaluation: 11.2.4.1 - Bypass Blocks, 11.2.4.2 - Page Titled, 11.2.4.5 - Multiple Ways, 11.3.1.2 - Language of Parts, 11.3.2.3 - Consistent Navigation, and 11.3.2.4 - Consistent Identification. According to the EN 301 549, these criteria are not applicable to mobile applications. The evaluation of success criterion 4.1.1 Parsing was limited to Android platforms due to technical constraints within the mobile context. Mobile applications fall into three main categories [24]: web content, native mobile applications, and hybrid mobile applications, which integrate native and web elements. This success criterion holds relevance for hybrid and web applications. However, pinpointing these cases posed a challenge, as precise identification requires access to their source code. Utilizing Solid Explorer[1] we detected application files with a html extension and assessed their compliance with the criteria through the W3C Markup Validation Service[2]. A comparable application or method was not identified for iOS applications.
4. **Define the tools to be used in the evaluation:** The WCAG-EM Report Tool [44] was employed to gather data regarding accessibility issues within each mobile application and to generate corresponding reports. Additionally, the WebAIM Contrast Checker [46] and a color picker app (Color Picker for Android[3] and Pixel Picker for iOS[4]) were used as supportive tools for evaluating color contrast.
5. **Evaluate the sample:** The chosen mobile applications were used on the selected devices to navigate through designated screens. The applications were then assessed for adherence to success criteria, employing a reporting tool to collect individual outcomes. Appt-EM provided simplified criterion descriptions, outlining the process for evaluating compliance. This information guided the evaluation of the following success criteria: 1.2.1 - Audio-only and Video-only (Prerecorded); 1.2.3 - Audio Description or Media Alternative (Prerecorded); 3.1.1 - Language of Page; 4.1.2 - Name, Role, Value; 4.1.3 - Status Messages. The guidance sections within these success criteria were referenced to better determine how to fulfill and test them in a mobile context.
6. **Report the results:** We employed the WCAG-EM report tool to generate evaluation reports. These reports were intended to be easily comprehensible, offering comprehensive insights into inaccessible elements, facilitating a more thorough analysis of the findings.

5.1.3 Data analysis. In our exploration of the evaluation process and its outcomes, we performed a critical analysis of the prescribed methodology. Building upon insights obtained from prior study involving accessibility experts, we focused this analysis on 1) the challenges of selecting screens, given the differences of the mobile context and the established adaptations, 2) assessing the predefined success criteria and exploring any potential ambiguities that might surface during this evaluation, and 3) any concerns related to the recommended reporting tool, WCAG-EM, and the complexity of presenting the results in a comprehensible manner, considering the diverse target audience, which comprises accessibility experts and monitoring agencies, and developers with varying degrees of accessibility knowledge.

5.2 Findings

Through manual evaluations of the mobile applications, it was possible to observe that the accessibility issues identified were consistent across various operating systems. Table 2 summarizes prevalent accessibility challenges across all eight applications, detailing associated success criteria and the frequency of each issue among the apps.

Across all applications, regardless of the operating system, we consistently observed issues. These included elements not being recognized as clickable by the screen reader, inadequate contrast in text and graphics, difficulties in resizing text or adjusting spacing, and keyboard inaccessible features. A more comprehensive evaluation is available in an external repository[5].

These evaluation results provide valuable insights into the extent of the issues we face and the most encountered problems in mobile applications. However, they also raise questions about the

[1] https://play.google.com/store/apps/details?id$=$pl.solidexplorer2&hl$=$en_US&pli$=$1
[2] https://validator.w3.org/#validate_by_upload
[3] https://play.google.com/store/apps/details?id$=$gmikhail.colorpicker&hl$=$pt_PT&gl$=$US
[4] https://apps.apple.com/gb/app/pixel-picker-image-color-picker/id930804327
[5] https://osf.io/q3wyt/?view_only$=$9432103c14574178925902c9b7247def

Table 2: Summary of Accessibility Issues Across Mobile Applications.

Evaluated criteria	Associated issues	iOS occurrence / Total	Android occurrence / Total	Occurrence / Total
1.3.1	Elements that aren't identified as clickable by the screen reader	4 / 4	4 / 4	8 / 8
1.1.1 1.4.5	Essential images or text images lacking alternative text	3 / 4	3 / 4	6 / 8
1.3.4	Failure of the application to adapt to landscape orientation	3 / 4	3 / 4	6 / 8
1.4.3 1.4.11	Insufficient contrast in textual and graphical elements	4 / 4	4 / 4	8 / 8
1.4.4 1.4.12	Inability to resize text or adjust spacing between letters, lines, or paragraphs	4 / 4	4 / 4	8 / 8
2.1.1 2.1.2	Keyboard inaccessible features	4 / 4	4 / 4	8 / 8
2.4.6	Lack of accessible labels or names announced by the screen reader	2 / 4	2 / 4	4 / 8
4.1.3	Changes in the application's content that go unannounced by the screen reader	1 / 4	1 / 4	2 / 8

commonly used methodology for such assessments. As previously mentioned, we utilized an adaptation of the WCAG-EM and the EN 301 549 methodologies, which aligns with the practices commonly employed by evaluators in their daily activities. Consequently, some considerations emerged regarding this process, especially during three main steps: Define screen sample, Evaluate the sample, Report the results.

5.2.1 Define screen sample. Both WCAG-EM and the EN 301 549 provide guidance on how to select the screen sample. This guidance is applicable to websites and mobile applications. While these instructions are meant to assist, some of the information provided may be somewhat ambiguous and not entirely authoritative, potentially leading to varied interpretations. Among these details, a few key points stand out. First, even the definition of a *screen* is less straightforward due to the absence of exposed URLs. Additionally, screens in mobile apps frequently feature overlapping elements, adding complexity to the distinction between individual screens. Furthermore, common screens on the web may differ from those on mobile platforms. For instance, screens like the sitemap or accessibility declaration are typically absent in mobile applications. This can pose challenges when selecting screens for evaluation in mobile applications. Another challenge arises from the subjectivity inherent in some of the points to be tested, potentially leaving the decision to the evaluator's discretion. For instance, consider the point addressing the inclusion of any other screens deemed relevant; the determination of relevance may vary from person to person. Even the criterion concerning randomly selected screens will depend on both the quantity of screens chosen and the evaluator's judgment. This circumstance may result in issues going unaddressed. While it is challenging to ensure the evaluation covers every potential problem, it is crucial to prioritize the assessment of the most significant screens that have a broader impact on users. In conclusion, all these factors introduce subjectivity in screen selection, potentially resulting in the omission of critical screens for evaluation or the oversight of significant flaws. This oversight can subsequently pose accessibility barriers for users.

5.2.2 Evaluate the sample. Evaluating certain success criteria for mobile applications can be challenging, as their application may not always be straightforward. The success criteria outlined in the WCAG were initially conceived for websites, and even in that context, their interpretation is not universally agreed upon [9]. The challenges encountered during the manual evaluations reinforce the feedback received during interviews, underscoring that applying these criteria to mobile applications introduces further uncertainties. Consequently, there is a need for adaptations tailored to mobile applications, including the incorporation of relevant examples and adjusted techniques, akin to the approach taken in Appt-EM, for improved clarity and effectiveness.

Among the evaluated criteria, some could benefit from a more straightforward and explicit process. For example, Success Criteria 1.2.1 and 1.2.3, focused on audio and video accessibility, poses significant challenges when adapting their descriptions and testing methods for mobile contexts. In our evaluation, we relied on the guidance and evaluation techniques outlined in Appt-EM. Another challenge specific to the mobile context is contrast evaluation. Success criteria 1.4.3 and 1.4.11 assess the color contrast of textual and graphical elements, respectively. To accomplish this, it is necessary to identify the color of elements on the mobile device. Several techniques can be employed for this purpose with the support of a color picker tool: projecting the mobile device screen onto a computer, taking a screenshot of the relevant display and opening it on the computer, or using the tool directly on the mobile device. Some of these approaches were also reported by interviewed evaluators. However, this entire process is time-consuming and can yield varying results depending on the technique used. Therefore, a more precise set of instructions for testing this criterion on mobile devices could assist in clarifying these concerns. The success criterion

3.1.1, pertaining to language definition, also presents challenges when it comes to its applicability in mobile applications. To assess this criterion, we referred to the description and evaluation guidelines provided in the Appt-EM. Finally, success criteria 4.1.2 and 4.1.3, which assess whether assistive technology users are aware of available actions and receive status messages, are somewhat unclear and leave room for ambiguity. To evaluate these criteria, we referred to the description and assessment methods provided in Appt-EM.

5.2.3 Report the results. In this step, reports are generated to document the accessibility issues identified during the application evaluations. As reported by the interviewed evaluators and as evident in the monitoring reports of our European member-states, there is currently no universal standard in place to facilitate the comprehension, comparison, and monitoring of results. Therefore, it would be advantageous to establish a universal format that could be adopted across the board. An example of such a format is the Evaluation and Report Language (EARL) [45], which is already utilized in some tools for result processing and reporting. Leveraging this example, it would be beneficial to develop a format tailored for monitoring reports that could also find utility in various tools. To report the evaluation results, accessibility evaluators and information gathered from monitoring reports typically indicate the use of the WCAG-EM tool or an Excel spreadsheet. In this study, the WCAG-EM reporting tool was employed. From a privacy standpoint, it is beneficial that this tool does not automatically save reports online. However, it would be advantageous if it offered this option for users who may choose to utilize it. There is a potential risk for users, such as accidentally closing the page or encountering device issues that result in page closure, causing the loss of work completed up to that point. The tool provides an option to save the report by exporting it as a JSON file, which can be accessed at any time. Additionally, it allows users to open previously saved reports. Nonetheless, it is worth noting that the tool, being relatively new, may exhibit occasional bugs in this specific functionality. These issues may manifest as difficulties in opening or incorrect handling of saved reports, potentially leading to user disruptions and the need to re-enter information.

5.3 Discussion

In this study, we conducted manual evaluations of eight mobile applications, including both Android and iOS versions, using the same methodology employed by the evaluators interviewed earlier. Our primary goal was to gain a comprehensive understanding of the nuances inherent in this evaluation process. To conduct manual evaluations, we devised and implemented a methodology based on WCAG-EM and EN 301 549, with a specific focus on mobile application requirements. We customized the WCAG-EM framework to align with the mobile platform context.

Our observations identified a **consistent presence of accessibility issues across diverse operating systems**. These issues encompassed elements that screen readers did not recognize as clickable, insufficient contrast in text and graphics, challenges associated with text resizing and spacing adjustments, and keyboard inaccessibility. Furthermore, certain concerns arose also during the design and execution of these evaluations. A notable challenge surfaced during the sample selection process, as **provided information exhibited ambiguity and lacked complete authority**, potentially leading to varying interpretations. For example, when determining the sample to be evaluated, it's recommended to include webpages that are relevant to the entire website, such as the homepage, sitemap, contact pages, and other pages typically linked from all other webpages. However, it's important to note that this same approach may not be transferable to the mobile context, as the common screens available in apps often differ from those found on websites. This layer of adaptation and subjectivity introduces complexities when choosing screens for mobile evaluations.

Another challenge emerged when evaluating specific success criterion, as **criteria originally intended for websites may not seamlessly apply to mobile applications**, introducing inherent uncertainties. Adaptations tailored to mobile applications, encompassing relevant examples and adjusted techniques, become imperative. This is evident, for instance, in the evaluation of color contrast (SC 1.4.3 and 1.4.11) in mobile apps. Various techniques exist for assessing color contrast in mobile apps, as highlighted during interviews. After testing three methods – capturing a mobile screenshot and using a computer-based color picker tool, projecting the mobile screen onto a computer and utilizing a computer-based color picker tool, and directly using a color picker on the mobile device – we found that the most accurate approach is using color pickers directly on mobile devices. This illustrates how a lack of guidance can lead to varying results.

Regarding the reporting of evaluation results, this step also requires further development. Currently, **there is no universally accepted format in place** to facilitate comprehension, comparison, and result monitoring. Additionally, a dearth of a robust, dependable, and universally embraced reporting tool compounds this issue.

6 STUDY 4: USER TESTING

Aligned with our previous study, we conducted usability testing with people with disabilities, informed by prior research [6, 12, 18, 28, 32, 38]. The rationale for this study's inclusion originates from the crucial role that user tests play in identifying accessibility challenges that may not be readily apparent through other means. This is often due to factors such as technological limitations or the need for real-time interaction for accurate identification. The primary objectives were twofold. First, as described in Section 2, a comprehensive methodology incorporates a range of techniques to uncover accessibility barriers. Thus, we aimed to (1) evaluate the efficacy of the manual evaluation conducted in the preceding phase in identifying accessibility issues. This involved investigating the accessibility barriers perceived by users with disabilities and comparing them with the outcomes obtained. Additionally, we aimed to (2) explore current challenges encountered in user testing processes within the framework of a comprehensive methodology. This study received approval from our University's Ethics Committee.

6.1 Method

6.1.1 Participants. A total of six individuals with diverse abilities were invited to participate in this study. These participants were recruited through the research team's network, and the selection criteria included being of legal age, using a mobile phone, and having a visual, hearing, motor, or cognitive impairment. Out of the six participants, three had visual impairments, two had motor impairments, and one had a hearing impairment. Two participants with visual impairments utilized screen readers: one used VoiceOver and

Table 3: Demographic data of participants of the user tests.

Participant	Age	Type of disability	Operative system used	Assistive technology or adaptative strategy
PU1	23	Visual	iOS	VoiceOver
PU2	26	Motor	Android	Screen mirroring
PU3	35	Visual	Android	Talkback
PU4	22	Visual	Android	Increased font size
PU5	54	Motor	Android	-
PU6	43	Hearing	iOS	-

the other used Talkback. The third participant used an increased font size configuration. Additionally, one participant with Cerebral Palsy used the mobile device's screen mirroring feature. Table 3 presents this information, including the participants' ages and the operating systems they used throughout the study.

6.1.2 Mobile applications and tasks. In this stage, we used the same four apps as in the manual evaluations, aiming for comparable results. The chosen tasks within the mobile applications were designed to include essential application functionalities, the possibility of encountering the previously mentioned errors, and interaction with different elements, such as buttons, forms, images, etc. Furthermore, a balance between interactive tasks and information retrieval was also pursued.

6.1.3 Procedure. Participants were recruited through the research team's social network. Noteworthy, our research and testing occurred in the same country, using locally developed apps in the participants' language to eliminate language barriers. Initially, they were contacted via email and provided with a concise study introduction. Following their agreement, a convenient schedule was arranged, and a thorough study outline was shared. Participants were also requested to confirm computer and mobile phone availability, along with pre-installing the Zoom application on both devices. This configuration allowed for the observation of participants' expressions and phone screens during task execution. Sessions were facilitated through Zoom, with participants' recording consent.

Tests began with a study and procedure overview. Any participant concerns were addressed, and consent for test recording was confirmed. The initial questions revolved around demographic aspects, including participants' impairments, age, and experience with mobile phones. Following the completion of tasks designated for each mobile application, participants were prompted to share their experiences. At the session's conclusion, after interacting with all mobile applications, participants were asked to elaborate on any specific issues they encountered and to provide feedback on their overall experience with the mobile applications. Lastly, participants were invited to share any additional comments and thoughts.

6.1.4 Data analysis. The data analysis from user tests consisted of three main stages. Initially, we conducted a quantitative examination of demographic information gathered from the questionnaire. Additionally, thematic analysis [10] was applied to the two open-ended questions. We then identified and classified barriers and errors experienced by participants during task execution, taking inspiration from prior research on accessibility barrier identification and categorization through user testing [8, 31], to enable a more objective analysis of the obtained data. These issues were categorized according to the following criteria: (i) problem: a concise description, (ii) source: detailing the triggering element or action, (iii) impact: explaining the consequences on interaction and task execution, (iv) severity: assessing the level of disruption (ranging from minor interaction shifts to task interruptions), (v) participant, (vi) assistive technology or adaptive strategy, (vii) frequency, and finally (viii) associated WCAG 2.1 success criteria.

This analysis included iterative discussions among researchers to achieve consensus on codes, themes, and barriers categorization.

6.2 Findings

Through an analysis of user tests, we gathered participant perceptions regarding the current state of accessibility provided by mobile applications, and the accessibility barriers that emerged during task execution. It is important to highlight that the main goal of this task is to gain insights into the user testing process itself, rather than to evaluate the accessibility levels of the chosen applications. Consequently, this section also discusses the challenges encountered during the designing, execution, and analysis of these sessions.

6.2.1 Accessibility barriers reported by participants. Prior to commencing task activities, participants were asked about their perceptions of mobile application accessibility and the challenges they faced in their daily activities. When inquired about their views on the suitability of current mobile applications for use with assistive technologies, the prevailing sentiment expressed was "Many don't have accessibility bugs" (PU1), "I'd say that most of the ones I've had contact with are at least minimally accessible" (PU3). Yet, when queried about the most frequent challenges encountered while using a mobile application, we observed a contrasting scenario.

Participants (PU4, PU5) reported encountering barriers due to the **small size of elements**, aggravated by limited support from the operating system's zoom feature. One participant (PU5) mentioned resorting to a mouse for specific tasks. Additionally, certain participants (PU2, PU3) highlighted **poor assistive technology support**, including difficulties, for instance, in entering security codes through an external keyboard (PU2), and issues such as application crashes and slow navigation (PU3). **Media accessibility** was also a concern raised by participants (PU4, PU6), such as the absence of zoom support for visual elements like stickers, images,

or GIFs (PU4), and the lack of subtitles in videos (PU6). Another well-known accessibility issue emerged in the response of one participant (PU1) - **Captchas**. However, in this instance, the participant faced challenges with the proposed alternative audio version. This participant cited difficulties with the speed at which Captchas are presented and the language in which they are delivered, which is not their first language, heightening the challenge. The **lack of labels** on buttons and other elements was also cited by one participant (PU3). Furthermore, one participant (PU6) expressed struggles in managing applications that necessitate **security calls**. Due to their hearing impairment, if no alternative is available, they consistently require assistance from others. Finally, one participant (PU6) summarized their perspective on accessibility: "The limitation will always be the responsibility of the application, not a constraint for me.".

6.2.2 Accessibility barriers identified during the sessions. Through an examination of the challenges encountered by participants, we identified a total of 10 barriers, of which 8 can be linked to WCAG criteria. The barrier most encountered (n=4) by participants relates to **1.3.1 - Info and Relationships**, and, in most cases, users did not complete the respective tasks. Additional barriers impeding task completion included those related to **1.4.4 - Resize text** (n=1) and **1.3.3 - Sensory Characteristics** (n=1). Furthermore, there were barriers that, while not preventing task completion, led to extended completion times: **2.5.5 - Target Size** (n=1) and **3.1.3 - Uncommon Terminology** (n=1). Furthermore, two barriers were identified but could not be directly associated with a success criterion. The first pertained to **information overload** (n=2), causing users to lose relevant information. The second involved a **lack of information about the screen the user is on while using a screen reader** (n=1), hindering the participant's task navigation. It's worth noting that the absence of a page title on mobile app screens can contribute to this issue, as it prevents users, especially those relying on screen readers, from quickly understanding the context and purpose of the current screen.

6.2.3 Comparing results. Comparing the results of manual tests with those from user tests, we observed that success criteria 1.3.1 - Info and Relationships, 1.3.3 - Sensory Characteristics, and 1.4.4 - Resize text were also identified in the manual analyses, but not perceived with the same frequency by the users. However, the other two criteria identified in the user tests (2.5.5 and 3.1.3) were not assessed in the manual tests, as they are not mandatory for evaluation according to EN 301 549. Both criteria pertain to AAA-level compliance, which is not covered by the standard.

6.2.4 Designing, executing, and analyzing sessions. User tests, like other techniques, can only identify a specific set of issues that may be present in a product or service. To gain a more comprehensive understanding of these issues, conducting tests with a larger pool of users than those included in this study would be necessary. It was observed, however, that certain problems, though covered by success criteria, may not be mandatory for evaluation in various contexts due to the required levels of compliance. This leads to unresolved issues. Furthermore, there are problems not addressed by any WCAG success criteria, even though they pose barriers for users. This analysis underscores the significance of user testing in a comprehensive accessibility assessment, revealing issues that might not be uncovered by other methods.

6.3 Discussion

This study aimed to evaluate the effectiveness of the current methodology employed by accessibility evaluators, as per the previous study, in identifying accessibility issues. We focused on scrutinizing the accessibility barriers encountered by individuals with various disabilities and compared these findings with the results of the manual evaluation. Six participants with diverse abilities, including blind, deaf, and motor-impaired users, participated in the user tests. To ensure a fair comparison, we employed the same mobile apps that underwent manual evaluations. Our analysis revealed that most challenges faced by participants could be attributed to WCAG success criteria. However, **two criteria, related to AAA-level compliance, remained unassessed and unreported** during manual evaluations since they are not mandatory according to EN 301 549 standards. While this study does not aim to provide a comprehensive accessibility assessment, it does underscore a significant limitation. For instance, Success Criterion 2.5.5 - Target Size, belonging to conformance level AAA and therefore often overlooked, becomes a fundamental requirement for mobile app users to effectively interact with and perceive content. In conclusion, this study highlights the necessity of reevaluating existing standards from a mobile perspective. It also underscores the pivotal role of involving users in accessibility assessments, as they can identify issues that other methods may overlook. Additionally, it's worth noting that, while not being legally mandated, most Member-States have not conducted these tests, as discussed in section 3.

7 DISCUSSION

In this study, we examine techniques and methodologies for evaluating the accessibility of mobile applications. We conducted four distinct studies, starting with an analysis of accessibility monitoring reports from EU Member States. We then interviewed accessibility experts with prior experience in mobile application assessments. Building on insights from these studies, we performed two accessibility evaluations: one involving manual assessment and the other incorporating tests with users who have disabilities. In this section, we will examine how the findings obtained through these studies can assist us in addressing our research questions concerning current practices, challenges, and opportunities.

7.1 What are the current practices used by evaluators in mobile accessibility, and how do they impact the outcomes of evaluations?

Our study investigated the current practices in mobile accessibility evaluations, revealing a **heavy reliance on manual evaluations**. Despite its widespread use, this method faces significant challenges due to the **lack of mobile-specific guidelines**. This results in varying approaches among evaluators, **affecting the consistency and reliability of the evaluations**. A closer look at the manual evaluation process uncovers further issues. Selecting representative samples is often hindered by unclear guidance, and **applying success criteria designed for websites to mobile applications** leads to additional complications. Moreover, the **limited scope of available support tools** for evaluators makes the process more cumbersome.

Automated evaluations, though less common, also face significant challenges. Our interviews indicate a **scarcity of tools specifically designed for mobile**, possibly explaining their limited use.

Our research revealed that **user testing is not widely employed**, likely due to the absence of legal mandates and the additional effort it entails. However, upon comparing the outcomes of manual evaluations with those integrating user testing, we uncovered specific **accessibility barriers within the user testing approach that had not been encountered in manual evaluations**. This observation underscores the pivotal role of user involvement in the assessment process. User perspectives are instrumental in identifying issues that might be missed by evaluators, thus significantly enhancing the overall assessment process and its outcomes.

Additionally, our study highlighted challenges in reporting evaluation results. While there is a suggested structure to follow, there is **limited clarity regarding the level of detail and the specific format for reporting these errors**. This issue is compounded by the **limited understanding of accessibility issues among developers**. The absence of a clear, standardized format for reporting errors and results not only hinders effective communication of the outcomes but also poses **challenges in addressing the identified accessibility issues**.

7.2 How can current methodologies for evaluating mobile accessibility be enhanced?

After conducting four studies and gathering diverse perspectives on mobile accessibility evaluations, it has become evident that this is a highly complex undertaking. The multitude of variables involved, including devices, assistive technologies, and operating systems, demands the provision of more comprehensive guidance for those engaged in such evaluations. In response to the three primary methods employed in this process - automated evaluations, manual evaluations, and user testing - we offer insights on how to enhance each of these approaches to establish optimal methods that yield reliable and comparable results.

7.2.1 Advancing automated evaluations. Advancing automated evaluations is a critical step in improving mobile accessibility assessments. We advocate for the integration of automated tools as a crucial aspect of this process. These tools offer evaluators and developers an initial assessment of accessibility issues that can be readily identified, significantly streamlining the evaluation process. For instance, a barrier frequently identified in our tests was the absence of labels on elements, which could be easily detectable through such tools. There's a need for **more comprehensive tools** that work in scenarios with limited source code access, cover various mobile platforms, and enable larger-scale testing. This broader approach ensures that mobile accessibility evaluations are more efficient, addressing challenges across various platforms and scenarios.

7.2.2 Improving manual evaluations. We identified three main issues related to the manual evaluation, all requiring additional guidance and support.

To improve the clarity of **sample selection** and reduce subjective choices that may omit crucial content, it is essential to establish precise guidelines for this step concerning mobile applications. These guidelines should prioritize content that users interact with most frequently, including key sections related to definitions, help, and support. This is crucial because users will refer to these pages when encountering unexpected issues while using the application. While similar concerns have been addressed in prior studies for websites [20, 47, 48], it is equally important to undertake comparable efforts for mobile applications.

Another critical step identified concerns the **evaluation of the sample**. Efforts must be made to establish precise guidelines and standards for assessing mobile applications. This includes the formulation of clearer and more objective success criteria, as well as the development of tailored testing techniques for mobile applications. Additionally, the existence of a common guide, detailing the testing procedures for each criterion, would be beneficial in preventing misinterpretations and ensuring evaluators are equipped with explicit directives. This uniform approach would promote consistency among evaluations. Furthermore, it is crucial to consider the levels of compliance required in different contexts. While the standard level of compliance is AA, this standard was established based on content delivered through websites. The same standard cannot be directly applied to the mobile context without further analysis. A notable example is Success Criteria 2.5.5 - Target Size, which ensures that elements in an application have a minimum size to ensure everyone can click them. This criterion belongs to the AAA compliance level, therefore it is often overlooked. Another criterion that warrants review due to the evolving landscape of social media platforms is Success Criteria 3.1.3 - Unusual Words, which poses a challenge in our studies and addresses a prevalent issue in various contexts [26].

Additionally, some success criteria require optimized testing information for the mobile context. The first case we highlight relates to color contrast evaluation (1.4.3 and 1.4.11). Multiple techniques for assessing color contrast in mobile applications, including variations used by evaluators during interviews, emphasize the importance of standardized guidance. Clear and precise testing procedures are essential for ensuring consistent results.

Finally, the last significant step requiring attention is the one responsible for providing instructions on how to **report the results**. We've noted that it tends to be overlooked due to the absence of official guidance, despite its importance. A common report format should be established, ensuring a standard and enhancing result comparability. Additionally, it is imperative to make concerted efforts to ensure that these reports are also designed to assist developers in identifying and resolving the reported issues.

7.3 Optimizing user tests

Regarding recommendations for enhancing the execution of user tests, the initial step would be to **emphasize the imperative nature of conducting user tests**. For comprehensive methodologies widely adopted, including those within legal frameworks, mere encouragement proves insufficient, as observed in the monitoring reports. Furthermore, to ensure effectiveness, it is crucial to provide detailed instructions on the proper execution of these tests. This includes details such as the procedure to be conducted, sample that

should be assessed, which and how many individuals should be involved, and specific assistive technologies to be included.

8 CONCLUSION

Mobile devices have become ubiquitous, seamlessly integrating into our daily routines. While affording greater independence and autonomy for people with disabilities, they also present unique challenges due to their specific features. This paper centers on methodologies for evaluating the accessibility of these devices. To this end, we conducted four studies. First, we analyzed reports detailing the outcomes of monitoring and enforcement activities in Europe. Subsequently, we conducted interviews with accessibility evaluators to gain further insights. Following, we undertook two additional studies to glean diverse perspectives through practical application. We performed a manual evaluation, mirroring the methods employed by the accessibility experts interviewed. The concluding study involved conducting user tests with people with disabilities to provide methodological insights, shedding light on current limitations and challenges.

It is important to mention that while this research focuses on accessibility assessment, providing accessibility entails a much broader process. This process commences with training and raising awareness among the personnel involved in designing and implementing these products. It further extends to ensuring that accessibility is integrated throughout the development cycle, culminating in the final product. Moreover, while guidelines and standards play a crucial role in upholding accessibility standardization, additional methods for evaluating accessibility are encouraged to explore different perspectives, exemplified by the work conducted by Ross et al. [32].

This work addresses challenges in assessing mobile accessibility within the broader context of widely adopted legal frameworks and current guidelines. We identified a lack of authoritative guidance for conducting such assessments, as well as a scarcity of automated tools to streamline the process. While the accessibility community emphasizes the importance of involving real users in assessments, the absence of formal requirements from legal entities and a dearth of guidance and procedures may contribute to the neglect of this critical task.

ACKNOWLEDGMENTS

This work was supported by FCT through the LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020). We would also like to thank to the participants of our study for their willingness to share their experiences and for their time and engagement.

REFERENCES

[1] Shadi Abou-Zahra. 2008. Web Accessibility Evaluation. In *Web Accessibility: A Foundation for Research*, Simon Harper and Yeliz Yesilada (eds.). Springer London, London, 79–106. https://doi.org/10.1007/978-1-84800-050-6_7

[2] Patricia Acosta-Vargas, Javier Guaña-Moya, Janio Jadán-Guerrero, Cleofé Alvites-Huamaní, and Luis Salvador-Ullauri. 2021. Towards Accessibility Assessment with a Combined Approach for Native Mobile Applications. In *Advances in Human Factors and System Interactions*, Isabel L. Nunes (ed.). Springer International Publishing, Cham, 234–241. https://doi.org/10.1007/978-3-030-79816-1_29

[3] Patricia Acosta-Vargas, Luis Salvador-Ullauri, Janio Jadán-Guerrero, César Guevara, Sandra Sanchez-Gordon, Tania Calle-Jimenez, Patricio Lara-Alvarez, Ana Medina, and Isabel L. Nunes. 2020. Accessibility Assessment in Mobile Applications for Android. In *Advances in Human Factors and Systems Interaction*, Isabel L. Nunes (ed.). Springer International Publishing, Cham, 279–288. https://doi.org/10.1007/978-3-030-20040-4_25

[4] Nancy Alajarmeh. 2022. The extent of mobile accessibility coverage in WCAG 2.1: sufficiency of success criteria and appropriateness of relevant conformance levels pertaining to accessibility problems encountered by users who are visually impaired. *Univ Access Inf Soc* 21, 2 (June 2022), 507–532. https://doi.org/10.1007/s10209-020-00785-w

[5] Ali S. Alotaibi, Paul T. Chiou, and William G.J. Halfond. 2022. Automated Detection of TalkBack Interactive Accessibility Failures in Android Applications. In *2022 IEEE Conference on Software Testing, Verification and Validation (ICST)*, April 2022, Valencia, Spain. IEEE, Valencia, Spain, 232–243. . https://doi.org/10.1109/ICST53961.2022.00033

[6] Claudine Auger, Emilie Leduc, Delphine Labbé, Cassioppée Guay, Brigitte Fillion, Carolina Bottari, and Bonnie Swaine. 2014. Mobile Applications for Participation at the Shopping Mall: Content Analysis and Usability for Persons with Physical Disabilities and Communication or Cognitive Limitations. *IJERPH* 11, 12 (December 2014), 12777–12794. https://doi.org/10.3390/ijerph111212777

[7] Marco Billi, Laura Burzagli, Tiziana Catarci, Giuseppe Santucci, Enrico Bertini, Francesco Gabbanini, and Enrico Palchetti. 2010. A unified methodology for the evaluation of accessibility and usability of mobile applications. *Universal Access in The Information Society* (2010). https://doi.org/10.1007/S10209-009-0180-1

[8] Giorgio Brajnik. 2006. Web Accessibility Testing: When the Method Is the Culprit. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang L. Zagler and Arthur I. Karshmer (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 156–163. https://doi.org/10.1007/11788713_24

[9] Giorgio Brajnik, Yeliz Yesilada, and Simon Harper. Testability and validity of WCAG 2.0: the expertise effect.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (January 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[11] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (July 2021), 328–352. https://doi.org/10.1080/14780887.2020.1769238

[12] Michael Crystian Nepomuceno Carvalho, Felipe Silva Dias, Aline Grazielle Silva Reis, and André Pimenta Freire. 2018. Accessibility and usability problems encountered on websites and applications in mobile devices by blind and normal-vision users. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, April 09, 2018, Pau France. ACM, Pau France, 2022–2029. . https://doi.org/10.1145/3167132.3167349

[13] CEN (European Committee for Standardisation), CENELEC (European Committee for Electrotechnical Standardisation), and ETSI (European Telecommunications Standards Institute). EN 301 549 V3 the harmonized European Standard for ICT Accessibility. *ETSI*. Retrieved September 8, 2023 from https://www.etsi.org/human-factors-accessibility/en-301-549-v3-the-harmonized-european-standard-for-ict-accessibility

[14] Wendy Chisholm, Gregg Vanderheiden, and Ian Jacobs. Web Content Accessibility Guidelines 1.0. Retrieved September 8, 2023 from https://www.w3.org/TR/WAI-WEBCONTENT/

[15] Raphael Clegg-Vinell, Christopher Bailey, and Voula Gkatzidou. 2014. Investigating the appropriateness and relevance of mobile web accessibility guidelines. In *Proceedings of the 11th Web for All Conference*, April 07, 2014, Seoul Korea. ACM, Seoul Korea, 1–4. . https://doi.org/10.1145/2596695.2596717

[16] Marcelo Medeiros Eler, Jose Miguel Rojas, Yan Ge, and Gordon Fraser. 2018. Automated Accessibility Testing of Mobile Apps. In *2018 IEEE 11th International Conference on Software Testing, Verification and Validation (ICST)*, April 2018, Vasteras. IEEE, Vasteras, 116–126. . https://doi.org/10.1109/ICST.2018.00021

[17] European Union. 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 (Text with EEA relevance). Retrieved July 6, 2023 from https://eur-lex.europa.eu/eli/dir/2016/2102/oj

[18] Simone Bacellar Leal Ferreira, Denis Silva Da Silveira, Eliane Pinheiro Capra, and Ariane Oliveira Ferreira. 2012. Protocols for Evaluation of Site Accessibility with the Participation of Blind Users. *Procedia Computer Science* 14, (2012), 47–55. https://doi.org/10.1016/j.procs.2012.10.006

[19] Tiago Guerreiro, Luís Carriço, and André Rodrigues. 2019. Mobile Web. In *Web Accessibility*, Yeliz Yesilada and Simon Harper (eds.). Springer London, London, 737–754. https://doi.org/10.1007/978-1-4471-7440-0_37

[20] Simon Harper, Anwar Ahmad Moon, Markel Vigo, Giorgio Brajnik, and Yeliz Yesilada. 2015. DOM block clustering for enhanced sampling and evaluation. In *Proceedings of the 12th International Web for All Conference*, May 2015, Florence Italy. ACM, Florence Italy, 1–10. . https://doi.org/10.1145/2745555.2746649

[21] IBM. Verify - automated – IBM Accessibility. Retrieved July 12, 2023 from https://www.ibm.com/able/toolkit/verify/www.ibm.com/able

[22] W3C Web Accessibility Initiative (WAI). WCAG 2 Overview. *Web Accessibility Initiative (WAI)*. Retrieved July 5, 2023 from https://www.w3.org/WAI/standards-guidelines/wcag/

[23] W3C Web Accessibility Initiative (WAI). WCAG-EM Overview: Website Accessibility Conformance Evaluation Methodology. *Web Accessibility Initiative (WAI)*. Retrieved July 13, 2023 from https://www.w3.org/WAI/test-evaluate/conformance/wcag-em/

[24] Ronald Jabangwe, Henry Edison, and Anh Nguyen Duc. 2018. Software engineering process models for mobile app development: A systematic literature review. *Journal of Systems and Software* 145, (November 2018), 98–111. https://doi.org/10.1016/j.jss.2018.08.028

[25] Padmaja Joshi and Saidarshan Bhagat. 2022. Effective Accessibility Testing Methodologies and Seamless Accessibility Integration in Mobile Applications. In *15th International Conference on Theory and Practice of Electronic Governance*, October 04, 2022, Guimarães Portugal. ACM, Guimarães Portugal, 449–455. . https://doi.org/10.1145/3560107.3560175

[26] Hae-Na Lee and Vikas Ashok. 2022. Impact of Out-of-Vocabulary Words on the Twitter Experience of Blind Users. In *CHI Conference on Human Factors in Computing Systems*, April 29, 2022, New Orleans LA USA. ACM, New Orleans LA USA, 1–20. . https://doi.org/10.1145/3491102.3501958

[27] Delvani Antônio Mateus, Carlos Alberto Silva, Arthur F. B. A. De Oliveira, Heitor Costa, and André Pimenta Freire. 2021. A Systematic Mapping of Accessibility Problems Encountered on Websites and Mobile Apps: A Comparison Between Automated Tests, Manual Inspections and User Evaluations. *JIS* 12, 1 (November 2021), 145–171. https://doi.org/10.5753/jis.2021.1778

[28] Delvani Antônio Mateus, Carlos Alberto Silva, Marcelo Medeiros Eler, and André Pimenta Freire. 2020. Accessibility of mobile applications: evaluation by users with visual impairment and by automated tools. In *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems*, October 26, 2020, Diamantina Brazil. ACM, Diamantina Brazil, 1–10. . https://doi.org/10.1145/3424953.3426633

[29] Kimberly A. Neuendorf. 2017. *The content analysis guidebook* (Second edition ed.). SAGE, Los Angeles.

[30] Eunju Park, Sungjun Han, Hogon Bae, Raekyung Kim, Seungjae Lee, Daejune Lim, and Hankyu Lim. 2019. Development of Automatic Evaluation Tool for Mobile Accessibility for Android Application. In *2019 International Conference on Systems of Collaboration Big Data, Internet of Things & Security (SysCoBIoTS)*, December 2019, Casablanca, Morocco. IEEE, Casablanca, Morocco, 1–6. . https://doi.org/10.1109/SysCoBIoTS48768.2019.9028034

[31] Letícia Seixas Pereira and Dominique Archambault. 2018. Correlating Navigation Barriers on Web 2.0 with Accessibility Guidelines. In *Computers Helping People with Special Needs*, Klaus Miesenberger and Georgios Kouroupetroglou (eds.). Springer International Publishing, Cham, 13–21. https://doi.org/10.1007/978-3-319-94277-3_3

[32] Erin Radcliffe, Ben Lippincott, Raeda Anderson, and Mike Jones. 2021. A Pilot Evaluation of mHealth App Accessibility for Three Top-Rated Weight Management Apps by People with Disabilities. *IJERPH* 18, 7 (April 2021), 3669. https://doi.org/10.3390/ijerph18073669

[33] Navid Salehnamadi, Abdulaziz Alshayban, Jun-Wei Lin, Iftekhar Ahmed, Stacy M. Branham, and S. Malek. 2021. *Latte: Use-Case and Assistive-Service Driven Automated Accessibility Testing Framework for Android*. https://doi.org/10.1145/3411764.3445455

[34] Navid Salehnamadi, Ziyao He, and Sam Malek. 2023. Assistive-Technology Aided Manual Accessibility Testing in Mobile Apps, Powered by Record-and-Replay. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023, Hamburg Germany. ACM, Hamburg Germany, 1–20. . https://doi.org/10.1145/3544548.3580679

[35] Navid Salehnamadi, Forough Mehralian, and S. Malek. 2022. *Groundhog: An Automated Accessibility Crawler for Mobile Apps*. https://doi.org/10.1145/3551349.3556905

[36] C. Siebra, W. Correia, Marcelo Penha, Jefté Macedo, J. Quintino, Marcelo Anjos, Fabiana Florentin, F. Q. Silva, and André L. M. Santos. 2018. *An analysis on tools for accessibility evaluation in mobile applications*. https://doi.org/10.1145/3266237.3266238

[37] Camila Silva, Marcelo Medeiros Eler, and Gordon Fraser. 2018. A survey on the tool support for the automatic evaluation of mobile accessibility. In *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, June 20, 2018, Thessaloniki Greece. ACM, Thessaloniki Greece, 286–293. . https://doi.org/10.1145/3218585.3218673

[38] Cláudia Ferreira da Silva, Simone B. Leal Ferreira, and Carolina Sacramento. 2018. Mobile Application Accessibility in the Context of Visually Impaired Users. In *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems*, October 22, 2018, Belém Brazil. ACM, Belém Brazil, 1–10. . https://doi.org/10.1145/3274192.3274224

[39] Jeanne Spellman, Jim Allan, and Shawn Lawton Henry. User Agent Accessibility Guidelines (UAAG) Overview. *Web Accessibility Initiative (WAI)*. Retrieved September 8, 2023 from https://www.w3.org/WAI/standards-guidelines/uaag/

[40] W3C Web Accessibility Initiative (WAI). 2023. Web Accessibility Laws & Policies. *Web Accessibility Initiative (WAI)*. Retrieved September 8, 2023 from https://www.w3.org/WAI/policies/

[41] W3C Web Accessibility Initiative (WAI). WCAG2ICT Overview. *Web Accessibility Initiative (WAI)*. Retrieved July 11, 2023 from https://www.w3.org/WAI/standards-guidelines/wcag/non-web-ict/

[42] W3C Web Accessibility Initiative (WAI). What's New in WCAG 2.1. *Web Accessibility Initiative (WAI)*. Retrieved July 11, 2023 from https://www.w3.org/WAI/standards-guidelines/wcag/new-in-21/

[43] W3C Web Accessibility Initiative (WAI). What's New in WCAG 2.2 Draft. *Web Accessibility Initiative (WAI)*. Retrieved July 11, 2023 from https://www.w3.org/WAI/standards-guidelines/wcag/new-in-22/

[44] W3C Web Accessibility Initiative (WAI). Overview | WCAG-EM Report Tool. *Web Accessibility Initiative (WAI)*. Retrieved September 8, 2023 from https://www.w3.org/WAI/eval/report-tool/

[45] W3C Web Accessibility Initiative (WAI). Evaluation and Report Language (EARL) Overview. *Web Accessibility Initiative (WAI)*. Retrieved September 8, 2023 from https://www.w3.org/WAI/standards-guidelines/earl/

[46] WebAIM. WebAIM: Contrast Checker. *WebAIM web accessibility in mind*. Retrieved September 8, 2023 from https://webaim.org/resources/contrastchecker/

[47] Zhi Yu, Jiajun Bu, Chao Shen, Wei Wang, Lianjun Dai, Qin Zhou, and Chuanwu Zhao. 2020. A Multi-site Collaborative Sampling for Web Accessibility Evaluation. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Roberto Manduchi, Mario Covarrubias Rodriguez and Petr Peňáz (eds.). Springer International Publishing, Cham, 329–335. https://doi.org/10.1007/978-3-030-58796-3_39

[48] Meng-Ni Zhang, Can Wang, Jia-Jun Bu, Zhi Yu, Yu Zhou, and Chun Chen. 2015. A sampling method based on URL clustering for fast web accessibility evaluation. *Front. Inf. Technol. Electron. Eng.* 16, 6 (June 2015), 449–456.

[49] 2022. Web Accessibility Directive - Monitoring reports | Shaping Europe's digital future. Retrieved July 7, 2023 from https://digital-strategy.ec.europa.eu/en/library/web-accessibility-directive-monitoring-reports

[50] 2023. Mobile accessibility checklist - Accessibility | MDN. Retrieved July 11, 2023 from https://developer.mozilla.org/en-US/docs/Web/Accessibility/Mobile_accessibility_checklist

[51] 2023. Appt Evaluation Methodology (Appt-EM). *Appt*. Retrieved July 13, 2023 from https://appt.org/en/guidelines/appt-em

[52] Convention on the Rights of Persons with Disabilities (CRPD) | Division for Inclusive Social Development (DISD). Retrieved July 5, 2023 from https://social.desa.un.org/issues/disability/crpd/convention-on-the-rights-of-persons-with-disabilities-crpd

[53] Web Accessibility Directive — Standards and harmonisation | Shaping Europe's digital future. Retrieved September 11, 2023 from https://digital-strategy.ec.europa.eu/en/policies/web-accessibility-directive-standards-and-harmonisation

[54] Android Lint Overview - Android Studio Project Site. Retrieved July 12, 2023 from http://tools.android.com/lint/overview

[55] Accessibility Scanner – Apps on Google Play. Retrieved July 12, 2023 from https://play.google.com/store/apps/details?id$=$com.google.android.apps.accessibility.auditor&hl$=$en_GB

[56] Accessibility Programming Guide for OS X: Testing for Accessibility on OS X. Retrieved July 12, 2023 from https://developer.apple.com/library/archive/documentation/Accessibility/Conceptual/AccessibilityMacOSX/OSXAXTestingApps.html#//apple_ref/doc/uid/TP40001078-CH210-TPXREF101

[57] Section508.gov. Retrieved July 11, 2023 from https://www.section508.gov/

[58] Mobile Web Best Practices 1.0. Retrieved July 11, 2023 from https://www.w3.org/TR/mobile-bp/

[59] Mobile guidelines. *Funka*. Retrieved July 11, 2023 from https://www.funka.com/en/research-and-innovation/archive---research-projects/mobile-guidelines/

[60] Mobile Accessibility Guidelines - Accessibility for Products - BBC. Retrieved July 11, 2023 from https://www.bbc.co.uk/accessibility/forproducts/guides/mobile/

[61] Section 508 Trusted Tester Conformance Test Process Version 5 | Homeland Security. Retrieved July 13, 2023 from https://www.dhs.gov/trusted-tester

[62] IBM Equal Access Toolkit – IBM Accessibility. Retrieved July 13, 2023 from https://www.ibm.com/able/toolkit/verify/www.ibm.com/able/toolkit

[63] Test your app's accessibility | App quality | Android Developers. Retrieved July 13, 2023 from https://developer.android.com/guide/topics/ui/accessibility/testing

[64] Accessibility. *Apple Developer Documentation*. Retrieved July 13, 2023 from https://developer.apple.com/design/human-interface-guidelines/accessibility

APPENDICES

A.1 INFORMATION ABOUT APPLICATIONS USED DURING THE TESTS

Table 4: Details about App1, a public sector app for train travel information, including its main features and identified accessibility issues.

Mobile application	App1
Domain	Public services
Description	This mobile application provides users with information on both long and short-distance train journeys. It offers access to timetables, ticket prices, ticket purchasing, and journey planning.
Main features	Discover ticket prices. Plan trips. Access train timetables and routes.
Main accessibility issues identified	Limited color contrast and unnotified pop-up menus or alerts for screen reader users. The mobile app necessitates text field completion and exploration to discover certain functionalities, potentially causing issues for keyboard and screen reader users. A time counter for ticket purchases lacks user-adjustable settings. Error identification and correction guidance are absent in login or sign-in text fields.

Table 5: Details about App2, a social network app, including its main features and identified accessibility issues.

Mobile application	App2
Domain	Social networks
Description	This mobile application enables users to stay updated on others' news and updates, facilitating easy communication and content sharing with their chosen contacts.
Main features	Search for themes or users. View content and news on any topic. Publish content. Comment or like content. Send messages to other users.
Main accessibility issues identified	Images without alt-text, auto-playing videos, and videos lacking subtitles. Small interface elements, such as buttons or boxes, that are difficult to click on.

Table 6: Details about App3, a public entity app for municipality information, including its main features and identified accessibility issues.

Mobile application	App3
Domain	Public entities
Description	This mobile application offers municipality information, including news, upcoming events, and their history.
Main features	Discover municipal information and news. Explore upcoming events. Report incidents.
Main accessibility issues identified	Excessive descriptive text lacking clear headings and small font size. Navigation issues when using a keyboard and difficulties in zooming on elements. Lack of intuitive guidance for users to discover functionalities in the mobile application. Absence of clear instructions in form text fields, leading to potential user confusion when encountering errors.

Table 7: Details about App4, an app providing information and guidance for a historical site, including its main features and identified accessibility issues.

Mobile application	**App4**
Domain	Museums and cultural entities
Description	This mobile application serves as a guide to a cultural and historical site, offering historical information.
Main features	Acquire historical information about the site.
Main accessibility issues identified	Automatic audio playback without user notification. Navigation issues when using a keyboard.