Chapter 11
End-User Evaluations

Sukru Eraslan and Chris Bailey

Abstract The past few years have seen tremendous development in web technologies. A range of websites and mobile applications have been developed to support a variety of online activities. The ubiquitous nature and increasing complexity of technology mean that ensuring accessibility remains challenging. Accessibility evaluation refers to the process of examining a product and establishing the extent to which it supports accessibility through the identification of potential barriers. While accessibility guidelines can guide the development process and automated evaluation tools can assist in measuring conformance, they do not guarantee that products will be accessible in a live context. The most reliable way to evaluate the accessibility of a product is to conduct a study with representative users interacting with the product. This chapter outlines a range of methods which can be used to ensure that a product is designed to meet the requirements and specific needs of users, from the ideation phase to the design and iterative development. The strengths and weaknesses of each method are described, as well as the primary considerations to ensure that the results of a study are reliable and valid, and also participants are treated ethically. This chapter concludes with a discussion of the field as well as an examination of future trends such as how data from user studies can be used to influence the design of future accessibility guidelines to improve their efficacy.

11.1 Introduction

Websites should be designed in a way so that they are accessible to users in the target population. When users access websites on devices with small screens, they should be able to complete their tasks. Similarly, when visually disabled users access

S. Eraslan (✉)
Middle East Technical University, Northern Cyprus Campus, 99738 Kalkanlı,
Güzelyurt, Mersin 10, Turkey
e-mail: seraslan@metu.edu.tr

C. Bailey
Enabling Insights Ltd., London, UK
e-mail: chris@enablinginsights.co.uk

Y. Yesilada and S. Harper (eds.), *Web Accessibility*, Human–Computer
Interaction Series, https://doi.org/10.1007/978-1-4471-7440-0_11

websites with their screen readers, they should not be distracted by unnecessary clutter which can cause a failure of task completion. Accessibility guidelines and automated evaluation tools provide guidance on how to develop usable and accessible websites, but unfortunately they do not guarantee that websites will be accessible to all users in the target population in a live context. Since user evaluations can identify usability and accessibility problems which are not discovered by conformance evaluation, they are crucial for designing usable and accessible websites (Henry 2018). Without end-user evaluations, researchers cannot ensure that all functionality of websites is accessible to all users in the target population.

End-user evaluations can be conducted at different stages of website development. For example, researchers can conduct a user evaluation during website development to identify user requirements for the final version of the website and investigate possible problems that the users can experience. When the website is finalised, another user evaluation can be conducted to ensure that there are no problems in accessing and using the website. If any problem is detected, then the issue should be resolved before releasing the website. Iterative development of a website would allow the detection of accessibility issues in the development stage and minimise problems in the final version of the website.

There are many methods available which can be used for end-user evaluations, including observations, questionnaires, interviews, eye tracking, etc. When researchers conduct an end-user evaluation for a particular website, they usually prepare a set of tasks and they observe how users interact with the website while performing these tasks. An interview, questionnaire or both can then be used to further investigate their overall experience with the website.

A representative sample of the target population is crucial for end-user evaluations. If the sample does not represent the target population, then the results of the evaluation will not be reliable. External factors which can affect the results should also be controlled. However, over control of these factors may cause a problem in representing a real-life situation, and again the results may not be reliable. Data from user evaluations should be analysed carefully as the incorrect interpretation of the data can cause other problems. When researchers conduct a user evaluation, they are also responsible for safeguarding the general welfare of their participants.

The remainder of this chapter first gives the overview of commonly used evaluation methods and explains what should be taken into consideration for designing an effective end-user evaluation. It then discusses the strengths and limitations of end-user evaluations and provides some future directions. Finally, it gives the authors' opinions of the field and provides concluding remarks.

11.2 Overview

There are a range of user-centred design methods that can be conducted when performing a user study. The most important consideration is to select the most appropriate method(s) to support the goal of the research. If background research is conducted to shape the development of a new product, then qualitative methods such as interviews and focus groups could be the most appropriate. If the goal is to elicit feedback or measure performance on an existing or prototype product, then observational studies or user studies following the think aloud protocol may yield the most effective results. Conducting research in accessibility can present significant challenges in recruiting suitable participants. It is important to factor this in the research timeline and consider if remote studies are possible to include participants who may experience difficulties travelling on-site for a laboratory-based study.

An overview of several commonly used research and evaluation methods are provided in Sect. 11.2.1. The key factors of designing an effective study including sampling of participants, internal and external validity, ethical treatment of participants and data analysis are covered in Sect. 11.2.2.

11.2.1 Commonly Used Evaluation Methods

Evaluation methods in user studies generally collect either quantitative or qualitative data. Some methods allow the collection of both quantitative and qualitative data for measuring the user's performance in terms of success, speed and satisfaction (Leavitt and Shneiderman 2006). In all cases, users will perform representative tasks on the interface to achieve a previously defined goal.

Quantitative evaluations are concerned with the collection and analysis of measurable numeric performance data that are obtained from users' interaction with a product. The collection and analysis of numerical data can describe, explain, predict or control variables and phenomena of interest (Gay et al. 2009). As these results are numeric, the results can be analysed using a range of techniques for statistical analysis (Dix et al. 2004). Quantitative evaluations are especially useful for benchmarking, that is, measuring and comparing users' performance on an interface over time. While generally not considered the best practice in the context of market research, there are benefits to recruiting the same participants for accessibility benchmarking studies as the participants' familiarity with the interface reduces the learning curve and can provide more useful results. Previously reported issues can be resolved in subsequent iterations of the design of the product.

Qualitative evaluations are more focused on gaining non-numeric information. These evaluations are conducted to gain an insight into users' existing experiences, their expectations and impressions of an interface, identify elements which cause negative user experience and potentially explore design solutions. Data from such evaluations is subjective as it is influenced by factors such as familiarity with the technology being tested, but the data can subsequently be coded to establish patterns and trends in users' opinions or users' feedback can be used to enhance the overall user experience of the product by removing barriers to the users' interaction.

The rest of this section provides an overview of some of the commonly used quantitative and qualitative evaluation methods, which are summarised in Table 11.1.

Table 11.1 The commonly used evaluation methods

Method	Data gained	Common uses
Performance measures	Quantitative	Collect numerical data to establish how well interface supports users
Logging user actions	Quantitative	Gather longitudinal data about a user's interaction with a product
Questionnaires	Quantitative/Qualitative	Collect information about users and their preferences
Observation	Qualitative	Obtain information on interaction in a live context
Interviews	Qualitative	Collect users' knowledge, thoughts, feelings and attitudes towards a product
Think aloud	Qualitative	Obtain users' thoughts and opinions about a product for identifying positive and negative aspects of their interaction
Eye tracking	Quantitative	Understand users' visual paths on page
Crowdsourcing	Quantitative/Qualitative	Collect numerical and non-numerical data from a large number of users at remote locations

11.2.1.1 Performance Measures

To investigate how well a user can interact with a digital product such as a website, it is necessary to decide which attributes of performance you wish to investigate and then define metrics with which to measure them (Brajnik 2006). If we look at the field of usability, there are five attributes of human–computer interaction that could be investigated in a user study (Nielsen 2003):

- Learnability: How easy is it for users to accomplish basic tasks the first time they encounter the design?
- Efficiency: Once users have learned the design, how quickly can they perform tasks?
- Memorability: When users return to the design after a period of not using it, how easily can they re-establish proficiency?
- Errors: How many errors do users make, how severe are these errors, and how easily can they recover from the errors?
- Satisfaction: How pleasant is it to use the design?

International standards can also provide guidance. ISO-9241-11 defines usability in terms of effectiveness, efficiency and satisfaction in a specified context of use (see the 'Usability, Universal Usability, and Design Patterns' chapter). The intention was to emphasise that usability or accessibility is an outcome of interaction rather than a property of a product and it is now widely accepted (Bevan et al. 2015). If using these standards as a benchmark, then attributes and subsequent metrics can be defined as follows:

- Effectiveness: The accuracy and completeness with which users achieve specified goals. It can be measured as the extent to which participants complete the defined task, expressed as the completion rate.
- Efficiency: The resources expended in relation to the accuracy and completeness with which users achieve specified goals. It can be measured in the time taken to complete the defined task, expressed as the time taken from the start to end time of the task.
- Satisfaction: The comfort and acceptability of use. User satisfaction is measured through standardised satisfaction questionnaires (SUS) which can be administered after each task and/or after the usability test session.

Experiments to measure user performance with an interactive system can be conducted throughout the product development life cycle; from the initial stages by testing paper prototypes and throughout the design process including interactive phases of development. Performance measures need to be tailored to reflect the stage of development being tested. For example, it would be appropriate to measure the satisfaction of a user's interaction with paper prototypes, but not efficiency.

When considering the specific context of accessibility research, examples of performance measures that can be obtained during user evaluations, but are not limited to, are as follows:

- The number of users who complete a task successfully.
- The time taken to complete a task.
- The number of errors a user makes while completing a task (such as selecting an incorrect link).
- The frequency that users can recover from such errors.
- The number of accessibility barriers that the user encounters.
- The number of observations of user frustration.

It should be noted that while some measures can be measured quantitatively, others can be measured qualitatively or by a combination of both. While task completion rate is purely a quantitative measure, the number of observations of user frustration can be measured quantitatively, but additional qualitative data is required to understand the reason behind the frustration. Measures may also be tailored to suit the user group being investigated. For example, visually disabled users who use screen readers generally take significantly longer to complete tasks on a website than sighted users (Borodin et al. 2010). Therefore, when conducting a study with visually disabled users, emphasis may be placed on measures such as effectiveness and satisfaction, over others such as efficiency.

In some domains, such as industry, when testing the accessibility of a product, the emphasis is placed on detecting and investigating solutions to accessibility barriers in a product. The use of research to detect barriers, rather than performance, is emphasised by practitioners (Brajnik 2006; Clegg-Vinell et al. 2014). By using a combination of quantitative and qualitative measures when investigating accessibility barriers encountered by a user, rich and useful data can be gained to remove these barriers and enhance the product.

11.2.1.2 Logging User Actions

Logging user actions is a quantitative research method which allows researchers to capture a continuous stream of data in real time as tasks are performed. Therefore, it is capable of providing valuable insights about users' interactions with websites. By using this method, researchers can capture large amounts of data from multiple users over a long period where log files are typically recorded by web servers and client logs (Nielsen 2004; Burton and Walther 2001). This method can be used to generate inferences about website design, to test prototypes of websites or their modifications over time and to test theoretical hypotheses about the effects of different design variables on web user behaviour (Burton and Walther 2001).

The logging process can occur with users in their natural environments without a researcher being present and therefore less invasive than qualitative evaluation techniques. While users are initially aware that they are being observed, over time the process becomes invisible and users often forget that logging is taking place. However, while users can often behave as though logging is not occurring, the evaluator should always inform users of what actions will be captured and why. Failure to do so raises serious ethical issues, and in some countries covertly capturing user data is illegal. The primary advantage of using logging for evaluations is that analytical data collection is typically built into the server or hosting providers software. This can produce records of server activity that can be analysed to describe user behaviour within the website. A typical application of web server logging is to enhance navigation for the user by establishing common paths through a website. An analysis of the logs can reveal the navigation path users take during their browsing sessions.

Metrics that can be obtained from logging user actions can be grouped into two categories; session-based and user-based metrics. Session-based metrics can be used to measure the average number of page views per session, the average duration of the session and the first (entry) and last (exit) pages of the session. User-based metrics can measure the number and frequency of visits, total time spent on the site, the retention rate (number of users who came back after their first visit) and conversion rate (the proportion of users who completed expected outcomes) (Kuniavsky 2003). These metrics can be correlated to the performance measures described in Sect. 11.2.1.1. Session-based metrics can be used to gain insight into the efficiency of users' interaction and identify pages which may present accessibility issues such as the exit page. User-based metrics can also be used to gain an insight into efficiency, but also give an insight into possible satisfaction with the site (frequent and repeated visits) as well as effectiveness (depending on the conversion rate).

One limitation of this method is that it may not provide sufficient insight into reasons behind users' breakdown in their interactions with a website—for example, failure to complete the checkout process of an e-commerce site. Quantitative data will detail the number of users who did not complete the interaction, but the reasons behind this breakdown would need to be explored with further research. Evaluators can understand what users did when interacting with a website, but cannot necessarily understand why they did it, and if they achieved their goal in a satisfactory manner (Nielsen 2004).

11.2.1.3 Questionnaires

One of the most reliable quantitative research methods to investigate and collect information about users and their opinions is to distribute a questionnaire to the target user group. A questionnaire or survey is a set of questions that when distributed to the target user group creates a structured way to ask a large number of users to describe themselves, their needs and their preferences. When done correctly, they produce a high degree of certainty about the user profile in areas such as demographics which cannot be so easily obtained when using qualitative methods. Special care should be taken over the design of the questionnaire to ensure that the questions are clearly understood and not leading, and also the results are accurate and provide sufficient detail to collect the desired information. These issues can arise due to the lack of direct contact with the participants, and therefore the questionnaire should be trialled beforehand (Kuniavsky 2003).

The questionnaire itself can consist of closed questions, open questions or a combination of the two. Closed questions typically pose a question and provide a predetermined limited set of responses which are accompanied by a numerical scale. Closed questions are heavily influenced by the commonly used System Usability Scale (SUS) used in usability testing. As an example, a question could be posed as ‘I was able to complete the form without any difficulties’, with the following predetermined responses: (1) Strongly Disagree, (2) Disagree, (3) Neither agree nor disagree, (4) Agree and (5) Strongly Agree. In this case, a score of five would indicate the form is accessible, while a score of one would indicate it is not. While such a question would gain an insight into the overall accessibility of a form, it would not provide detail around any accessibility barriers or features that prevented—or supported—form completion. If this detail was required, it would be necessary to complement it with an open question.

Open questions are phrased to allow the respondent to answer in free text and provide more detail. The data can subsequently be coded to establish any patterns and trends. Again, using a form to provide context, an open question could be phrased as ‘Please describe your experience when using the form on the website?’. Such a wording avoids making assumptions about the user being able to complete the form and allows the user to express their opinions. For example, a screen reader user could respond that required form fields were not announced to them and this response provides specific insight into possible accessibility problems.

Questionnaires can be delivered electronically to a large sample of users or can be completed manually on paper in conjunction with observational or think aloud research sessions. The former is useful when starting research into the accessibility of a product as it can provide useful demographic information and influence the design of observational and think aloud research sessions. For example, responses to open-ended questions could provide insight into specific areas or components of the product that require further investigation and testing. If large-scale questionnaires are used, then an incentive for participation should be offered. If used in conjunction with other research methods, they can provide a written record of responses directly from the user and provide information that can be used for benchmarking the accessibility of a product if further research is conducted.

11.2.1.4 Observation

Ethnography and observations are often used as part of a contextual inquiry and are used to gain a better understanding of users in their natural environments. Observational research methods involve observing users in the place where they would normally use the product (e.g. work, home, etc.) to gather data about who that target users are, what tasks and goals they have related to an existing product (or proposed enhancements) and the context in which they work to accomplish their goals. The outputs from this qualitative research can lead to the development of user profiles, personas (archetype users), scenarios and task descriptions on which the design team can base design decisions on empirical evidence throughout the development life cycle (Rubin and Chisnell 2008).

Observation of users while operating an interface provides real-time interaction information and can be performed by curing both the pre-design and post-design stages of a project. Observations conducted during the pre-design phase can be used to identify required enhancements to a newer version of a product or to establish user requirements for a new product. When conducted during the post-design stage, evaluations are used to identify task, performance and environmental factors which could negatively impact their experience. For example, evaluations can be conducted to ensure a product meets users' expectation and to identify required enhancements by gaining data on accessibility and usability barriers.

Observations conducted in a live context can be unstructured and discrete, meaning the researcher will take detailed notes but rarely interfere with the research setting ensuring the authenticity of the work. Alternatively, the researchers can intervene when they observe a breakdown in a user's interaction to determine the cause of the problem, evaluate the impact and provide an insight into a possible solution. Care must be taken to ensure that an inexperienced researcher does not influence the user's interaction and behaviour. Similarly, collecting a video and audio recording of the session can be useful when analysing the session, with the proviso that users who are aware they are being recorded may not behave exactly as they would in a regular context as they may feel uncomfortable. Due to the nature of observational studies being conducted in a live context, the research design should be well planned, the objectives should be clear and the researcher should be suitably trained and experienced. For further details on observational research design, the issues that need to be considered, see (Leedy and Ormerod 2016).

11.2.1.5 Interviews

Interviews are a useful method to investigate users' needs, preferences and desires in a detailed way. They can be used to deepen understanding of users and discover how individuals feel about a product, including why and how they use it. Interviews allow the researcher to gain an understanding about why users hold certain opinions, beliefs and attitudes towards a product. To ensure the interviewer does not influence the results, it is important to conduct non-directed interviewing to make sure the interview process does not lead or bias the answers. It ensures that the participants' true thoughts, feelings and experiences are gained without being filtered through the preconceptions of the interviewer (Kuniavsky 2003).

Interviews can be used to explain general data obtained from large-scale questionnaires or surveys, adding more depth and understanding to previously gained data due to the potential to ask very specific questions. Any uncertainties or ambiguity in a response can be quickly clarified. Due to the one-to-one nature, interviews can be an effective method to explore sensitive topics which people may not feel comfortable discussing in groups and they can allow a greater level of rapport to be built with participants. Interviews offer some flexibility in that they can be structured, semi-structured or ad hoc. Structured interviews are best used if accurate comparisons between users' responses are required. The larger sample size is reached if they are conducted over the telephone or the Internet by several interviewers. Semi-structured interviews allow the interviewers more flexibility to explore topics and themes which emerge during the interview and can offer a more informal approach. Ad hoc interviews are best used when performing guerrilla research or when performing a contextual inquiry and they are the most flexible method. Interview questions must be phrased to be open ended to more easily elicit useful responses from participants. The wording of the questions should not be loaded or worded in a way that could influence the response. Questions with complex answers should not be posted as binary questions, for example, instead of asking 'Is X feature of a screen reader useful to you?', ask 'Can you tell me how you use a screen reader?'. Finally, the interviewers should be wary of asking people to predict their future needs or assume that they will be able to answer every question.

There are also several participant behaviours that can influence the results which the interviewer needs to be aware of. Participants may not always say what they truly believe; this could be as they feel they want to avoid conflict; they may say yes when they mean no. Indicators of this include hesitation when responding or inconsistency with previous answers. More subtle cues may be noted by observing body language, such as someone shaking their head no when answering positively to a question. Participants may also give a different answer to the question asked. This could be because they misheard or did not understand the question, or might have the agenda they wish to discuss. It is important to listen carefully, as the information may still be relevant in the context of the interview script. It may be necessary to repeat the question by using the different wording of phrasing, and persistence may be required to obtain the required data (Kuniavsky 2003).

Group interviews with several participants can be conducted in the form of focus groups. Focus groups are an established user-centred design research method which captures shared experiences in a live situation. They are generally used in the early stages of a research investigation to gain qualitative feedback which can shape the development of a digital product. Focus group interviews involve a group of six to eight people who come from similar social and cultural backgrounds or who have similar experiences or concerns. They gather together to discuss a specific issue with the help of a moderator in a setting where participants feel comfortable enough to engage in a dynamic discussion for 1 or 2 hours. Focus groups do not aim to reach a consensus on the discussed issues. Rather, focus groups 'encourage a range of responses which provide a greater understanding of the attitudes, behaviour, opinions or perceptions of participants on the research issues' (Hennick 2007). Focus groups enable participants to share and discuss their thoughts and attitudes and provide a balanced understanding of their experience and perceptions for analysis. It is also possible to investigate the appropriateness of potential design solutions for new products. Due to their flexibility, multiple topics or themes can be explored in a session and they can be used for formative assessment with the use of visual probes, such as design mock-ups. The facilitator must ensure they manage the discussion to ensure that all participants are involved as one drawback of focus groups is that the discussion can be dominated by one or two individual participants (Breen 2006). While focus groups are an established user-centred design method, their roots in market research have led to a discussion on their suitability for HCI research. Great care should be taken in the research design to reflect the task-based investigation required for focus groups in the field of usability—and by extension—accessibility investigations (Rosenbaum et al. 2002).

Evaluating and analysing the results of interviews have the disadvantage of being time-consuming and being difficult to analyse. However, an experienced researcher can encourage the participant to provide an in-depth discussion that yields rich data to reveal strong conclusions when correctly analysed and coded (Jay et al. 2008).

11.2.1.6 Think Aloud

The think aloud protocol was originally developed to understand users' cognitive process as it encourages them to comment on their actions out loud when performing tasks on a system (Lewis 1982). The think aloud protocol is commonly used during user-testing sessions as it can provide a detailed insight into users' thoughts, feelings and actions during their interactions with a product. Using current or intended users of the product as participants in the think aloud protocol provides a closer view of how users use the product and reveals practical problems related to task performance (Holzinger 2005). It does not only highlight issues with the product that are detected during a live context of use but it can also provide suggestions for possible design solutions (Rubin and Chisnell 2008).

Think aloud testing sessions can be held on a one-to-one basis, with one facilitator working with one participant in a laboratory setting, or there may be an additional observer present. The observer may be in the laboratory with the facilitator and participant or they may be viewing the session in a separate observation room. Testing sessions may also be conducted remotely by using remote screen sharing software. In that case, a video–audio connection is required to view and hear the participant. If a researcher wants to observe the interactions of participants with a particular website remotely, then dedicated software such as Morea[1] can be used to assist with post-session analysis. Morea is a software application that allows researchers to observe and record users' interactions with a product remotely. It uses a timeline metaphor to allow researchers to place markers when observing a session which can then be reviewed and used for more in-depth analysis.

The key aspect of the think aloud protocol is that participants are encouraged to verbalise their thoughts, feelings, opinions, expectations and frustrations around their interactive experience in the form of a 'running commentary'. Participants are asked to explain their actions, such as their justification for following a certain navigation path on a website or what motivated their recent action. Crucially, this allows researchers to see and understand the cognitive processes associated with task completion and identify any barriers (Yen and Bakken 2009).

Asking participants to think aloud during their sessions also reveals important clues about how they are thinking about the product or system they are using and whether the way it works matches up with the way it was designed. In effect, does it match the participants' mental model? Participants may filter their verbalisation, so they may consciously or unconsciously leave things out as they talk. Likewise, it is impossible for a participant to articulate everything that is going through their mind during the session. Thinking aloud can also help participants think through the design problem and form ideas for recovering. One important reason to avoid asking participants to think aloud is when you measure time on tasks. Thinking aloud slows performance significantly (Rubin and Chisnell 2008).

11.2.1.7 Eye Tracking

Eye tracking has widely been used to understand how users interact with web pages for enhancing the design and usability of web pages (Ehmke and Wilson 2007; Yesilada et al. 2013; Eraslan et al. 2013). While users are reading web pages, their eyes become relatively stable at certain points called fixations and the sequences of these fixations show their scanpaths (Poole and Ball 2005). By tracking eye movements of users on web pages, we can discover which elements are used and which paths are followed by users. Table 11.2 shows some popular eye-tracking metrics along with their common interpretations, see more in Ehmke and Wilson (2007). For example, if users are asked to search for a specific item on a particular web page and they make many fixations and follow unnecessarily long paths to complete their tasks, then their searching behaviours tend to be considered as inefficient on that page for the given task.

[1] https://www.techsmith.com/morae.html.

Table 11.2 Some popular eye-tracking metrics along with their common interpretations

Metric	Interpretation
Fixation duration	Longer fixation duration, more difficult to extract information or more engaging
Number of fixations overall	Higher number of fixations overall, less efficient searching
Number of fixations on a particular element	Higher number of fixations on an element, more noticeable or important element
Scanpath length	Longer scanpath, less efficient searching
Transitions between areas	More transitions between elements, more uncertainty in searching

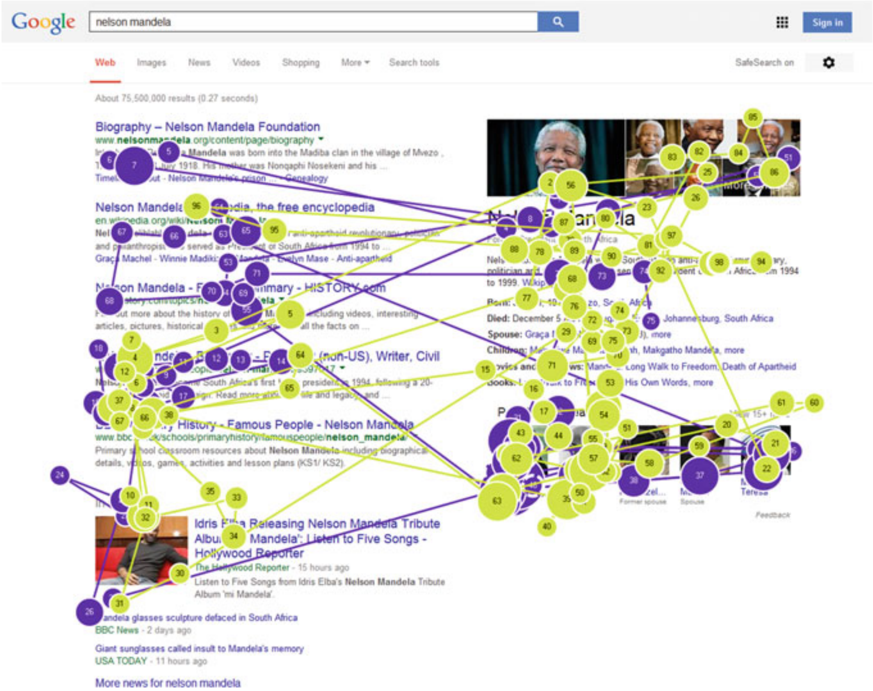
Google
nelson mandela
Sign in
Web
Images
News
Videos
Shopping
More ▾
Search tools
SafeSearch on
About 75,500,000 results (0.27 seconds)
Biography – Nelson Mandela Foundation
Nelson Mandela – Wikipedia, the free encyclopedia
Nelson Mandela - Facts & Summary - HISTORY.com
History - Famous People - Nelson Mandela
Idris Elba Releasing Nelson Mandela Tribute Album 'Mandela': Listen to Five Songs - Hollywood Reporter
Mandela glasses sculpture defaced in South Africa
BBC News - 2 days ago
Giant sunglasses called insult to Mandela's memory
USA TODAY - 11 hours ago
More news for nelson mandela
Died: December 5
Spouse:
Children:
Books:
Feedback

Fig. 11.1 A scanpath visualisation with a gaze plot on the search results page on the Google website (Eraslan et al. 2016a, c, 2017c)

Eye-tracking data can be visualised by using different ways (Blascheck et al. 2017). Figure 11.1 shows an example of a scanpath visualisation with a gaze plot on the search results page on the Google website. The circles illustrate the fixations where the larger circles illustrate the longer fixations. When multiple scanpaths are visualised on the same page with gaze plots, they will overlap and become difficult to analyse. In addition to visualisation techniques, some other techniques have also been

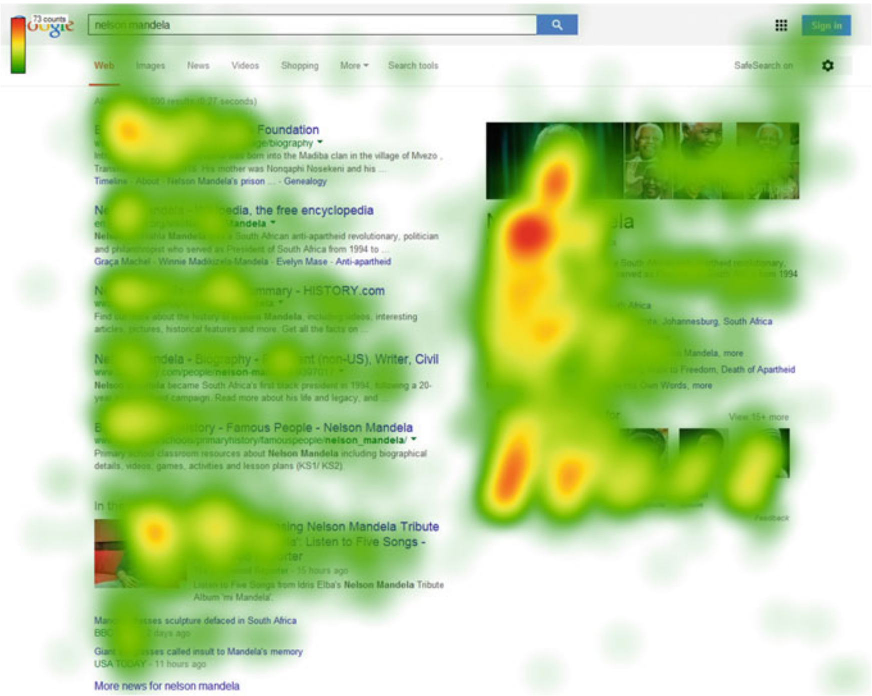
73 counts
nelson mandela
Sign in
Web
Images
News
Videos
Shopping
More
Search tools
SafeSearch on
Foundation
edia, the free encyclopedia
mmary - HISTORY.com
nt (non-US), Writer, Civil
story - Famous People - Nelson Mandela
ing Nelson Mandela Tribute
More news for nelson mandela

Fig. 11.2 A heat map on the search results page on the Google website (Eraslan et al. 2016a, c, 2017c)

proposed and used in the literature to analyse scanpaths for different purposes, such as computing a similarity score between two scanpaths, determining transition probabilities between elements, detecting patterns in a set of scanpaths and identifying a common scanpath of multiple users (Eraslan et al. 2015).

Fixations can also be aggregated based on certain features, such as duration or count, to generate a heat map. Figure 11.2 shows an example of a heat map on the search results page on the Google website. Heat maps consist of different colours where the red colour usually illustrates the most commonly used elements, whereas the green colour illustrates the rarely used elements. However, these maps do not illustrate sequential information.

Even though eye tracking gives valuable insights about how users interact with web pages, it does not tell why certain elements are fixated by users. Besides this, there can be some involuntary eye movements which are made by users without any specific objective. Furthermore, eye-tracking studies can be costly due to the expensive equipment and can be time-consuming as eye-tracking sessions cannot be conducted in parallel when there is only one eye tracker.

11.2.1.8 Crowdsourcing

Oxford English Dictionary[2] defines crowdsourcing as ‘the practice of obtaining information or input into a task or project by enlisting the services of a large number of people, either paid or unpaid, typically via the Internet’. As researchers usually experience difficulties in finding participants for their user studies (Eraslan et al. 2016a, 2017c), they can use this approach to access a wider and diverse set of people from remote locations and ask them to assess a particular software product (Sherief et al. 2014). In particular, they can ask participants to perform certain tasks on a particular website in their environments and fill in a questionnaire in regards to their overall experience (see Sect. 11.2.1.3). The questionnaire can involve a set of questions which aim to identify what kinds of problems can be encountered. The actions of the participants can also be logged to be analysed (see Sect. 11.2.1.2).

Eraslan et al. (2018) recently propose a methodology for creating a corpus of eye-tracking data on web pages by using crowdsourcing. This corpus can allow researchers to access the eye-tracking data collected on different kinds of web pages, thus providing an opportunity to investigate how people interact with different kinds of web pages, such as web pages with various levels of visual complexity. The analysis of eye-tracking data on different kinds of web pages provides valuable insights for possible problems for other similar web pages.

Even though crowdsourcing allows to collect data from a large number of users, the analysis of the collected data is crucial. There can be many factors which can affect users’ evaluations, such as expertise, disabilities, experience, etc. The reliability of the collected data should also be checked as this method can reveal unreliable data due to low control. For example, if a questionnaire is conducted to evaluate a website, some participants may fill in the questionnaire randomly. When crowdsourcing is used, researchers should ensure that they obtain sufficient data from users to investigate the reliability of the collected data and analyse the data by considering different factors.

11.2.2 Designing an Effective Study

User studies have their own characteristics, research questions, hypotheses and methodologies, but they should be designed effectively in order to achieve reliable results. These studies should be conducted with a representative sample of the target population and the sample size can vary due to the heterogeneity of the population (Sect. 11.2.2.1). Researchers should also consider how to control external factors which can affect dependent and independent variables (i.e. internal validity) and how to have generalizable results (i.e. external validity) (Sect. 11.2.2.2). When researchers plan their quantitative and/or qualitative data analysis beforehand, they can directly start to analyse the data once the data collection stage is completed (Sect. 11.2.2.3).

[2] https://en.oxforddictionaries.com/definition/crowdsourcing.

Since users are included in these studies, ethical issues should be taken into consideration to assure the rights and welfare of the participants (see Sect. 11.2.2.4). All of these issues are briefly explained below.

11.2.2.1 Sampling

When researchers want to carry out an end-user evaluation for a particular product, they need to select their participants from their target population. Specifically, when researchers want to evaluate a website designed for a specific field, their participants should have knowledge in this field. Generally speaking, if a user evaluation is conducted with more participants, then its results will be more generalisable. However, the generalisability of the results does not depend on only the sample size but also depends on the representatives of the target population. In particular, visually disabled users access web pages with their screen readers that follow the source code of the pages (Yesilada et al. 2007). Therefore, these users tend to follow similar strategies when they interact with web pages, even though there can be some differences because of their screen readers. When researchers want to assess the accessibility of a particular website for visually disabled users, they can have a representative sample with a small number of participants. However, if the heterogeneity of the target population is high, then a larger sample size will be needed for a representative sample. Hence, researchers should first investigate and understand how the individuals of the target population are similar or different from each other to determine their sample size.

After a user evaluation study, researchers typically analyse their results by applying some statistical tests. The number of participants required for a specific statistical test to achieve a particular statistical power can be estimated based on statistical approaches. Studies with low statistical power tend to have a Type II error which is a failure to detect a significant difference (Gravetter and Wallnau 2008). G*Power is a software application which is designed to estimate the required sample size based on the study design.[3] For example, when the Mann–Whitney U Test is planned to be used to compare two unrelated groups based on a particular dependent variable, this software application can be used to estimate the required sample size to achieve specific statistical power. To achieve at least 95% statistical power, the required sample size is determined as 92 with this application when the other parameters are set to their default values (Effect size d: 0.5, α err prob: 0.05, Allocation Ratio: N2/N1).

11.2.2.2 Validity

While an end-user evaluation is being performed, there can be some confounding variables which are outside factors possibly affecting both dependent and independent variables (Eysenck 2005). Assume that researchers want to investigate whether

[3] http://www.gpower.hhu.de/en.html.

a particular task on a specific web page can be completed within a given duration by two separate user groups which have different levels of computer knowledge. If they conduct their studies with mainly male users, there can a problem for internal validity. Specifically, when two different groups of users will be compared based on a particular dependent variable, all the other variables should be the same to achieve the highest internal validity. Therefore, when researchers set up their studies, they should take confounding variables into consideration. However, it is difficult to achieve very high internal validity because of external factors, such as individual preferences, knowledge, familiarity, etc. For example, there are different kinds of small-screen devices available. If a user evaluation is carried out with a specific small-screen device, the internal validity can be decreased as some of these users can be more familiar with the device in comparison with others. The users can be allowed to use their devices, but in that case the individual settings of the devices could also negatively affect the internal validity.

On the other side, if all the confounding variables are eliminated, then the evaluation design will not represent the real world, and this situation will negatively affect the external validity which is related to the generalisability of the findings. When the sample is not representative of the target population, the external validity will also be negatively affected. Both the real-world conditions and the representatives of the sample are crucial for achieving high external validity. Thus, when researchers design a user evaluation, they should consider both internal validity and external validity based on their objectives.

11.2.2.3 Data Analysis

When researchers design a user evaluation, they also need to make a plan on how they will analyse the data that will be collected in the evaluation. The data should be carefully analysed with appropriate methods and techniques to interpret the data accurately. Incorrect interpretation of data can result in unsuccessful evaluation even though its data collection stage does not have any problem. Data analysis mainly depends on the type of the data collected which can be quantitative or qualitative.

There are many statistical methods that can be used for analysis of quantitative data such as the time required by the user to complete a task, the number of incorrect links choices, etc. For example, when researchers want to investigate whether people with autism can complete a specific task as efficiently as people without autism, they can use the independent T-Test or its non-parametric alternative Mann–Whitney U Test to investigate if these two groups are significantly different from each other based on a particular dependent variable, such as the time required by the user to complete a task (Pallant 2007; Eraslan et al. 2017a). There are some assumptions that should be satisfied to apply the T-Test. For example, the values of the dependent variable for both groups should be normally distributed. If these assumptions are not met, then the Mann–Whitney U Test should be used. When the sample size is small, some sampling methods can also be used to eliminate the effects of the individuals. For example, some subsamples can be randomly created by using the bootstrapping technique where a specific participant is not seen in more than one subsample (Hesterberg 2015).

When a user evaluation provides some qualitative data such as manuscripts from interviews, observations and think aloud studies, researchers can use different techniques to analyse them to discover patterns and relationships between these patterns. Specifically, they can use an iterative approach to first divide larger units into smaller units and then sort and categorise them to draw an overall conclusion.

11.2.2.4 Ethical Issues

Ethical issues should be considered when users are involved in studies. Even though web accessibility evaluation does not threaten physical welfare of participants, researchers are responsible for safeguarding general welfare of their participants. At the beginning of the study, the participants should be informed about the study to understand the main objectives of the study, how it will be conducted, how long it will take and their rights. They will then need to sign a consent form. It is unethical, and also illegal in some countries, to capture data from users without their consent.

Researchers typically experience difficulties in finding participants for their studies (Eraslan et al. 2016a, 2017c). People typically participate in these studies because of their interest, their willingness to help the researchers and/or small gifts which are given after the studies. Therefore, researchers should avoid asking them to complete very complicated questions and keeping them in a laboratory for a long period as they may become tired and stressed, and then decide to withdraw from the study.

Researchers should avoid collecting sentinel information from their participants. Web pages with sensitive information are typically not suitable for user evaluations. For example, a social media account of a particular user is not appropriate as it is likely to include sentinel information and users may not want to share their sentinel information. Similarly, personal email pages should also not be used in a user evaluation. Researchers should take care of sentinel data. In particular, electronic sentinel data collected from the participants should be stored securely on a computer, and written information should be stored in a locked drawer. Researchers should also not pass them to any third party.

When the analysis of the data is reported, the data should be anonymised. It means that participant names and any other information that may identify the participants should not be provided, thus no one can recognise who the data belongs to. To maintain the anonymity, a code can be given to each participant, and the participants can be referred with their codes.

If people with disabilities are invited to a laboratory for a user evaluation study, the laboratory should be accessible for them and well equipped to ensure that they feel comfortable. For example, if participants have some physical disabilities, the laboratory should be set up properly, and therefore they can easily move around.

Many universities have an ethics committee to check whether a particular study is suitable in terms of ethical issues and approve the study to proceed. The committee can also check the study case by case whether it follows the ethical rules.

11.3 Discussion

One of the most important aspects of the research design is to ensure that the hypotheses and research questions are clearly identified and defined. Without this, the goal of the research may be unclear, and the subsequent research design may be flawed. The research design defines what methods will be used, how the study will be conducted, what equipment will be required, what data will be collected, how it will be obtained and then subsequently analysed. Well-defined research questions are required to understand what participants will be recruited, how many will participate, what information is required from them, such as what tasks they will perform. A literature review of existing research in the field can provide guidance on how these should be defined.

The research design should include both independent and dependent variables. An independent variable is a variable which is changed or controlled in a scientific experiment to test the effects on the dependent variable, whereas a dependent variable is a variable which is being tested and measured in the research study. The dependent variable is 'dependent' on the independent variable. As the experimenter changes the independent variable, the effect on the dependent variable is observed and recorded. While purely exploratory studies may not have clear hypotheses and can deliver valid results, there should still be high-level research question(s) that are being investigated.

The tasks that participants are required to complete during the evaluation should be chosen appropriately to answer the research questions or test the hypotheses. If the goal of the research is to progress or reinvestigate existing research, then these can be drawn from a review of existing literature. Conversely, new tasks may need to be defined depending on the context of the investigation. The tasks should be designed to be typical of what target users would complete in a live context. Although it may not be necessary to test all elements of an interface, the tasks should cover the core functionality of the product or interface being evaluated. The preparation of the tasks should be influenced by the overall hypothesis and research questions.

One of the possible limitations of user studies, especially in the domain of accessibility research, is to ensure that the results are generalisable. Participants with disabilities have such a broad spectrum of capabilities and needs which makes difficult to draw definite conclusions. Taking screen reader users as an example, the preferred choice of assistive technology and the varying level of competence and experience can greatly influence the results. Assume that highly competent and experienced screen reader users and novice screen reader users are asked to interact with the same interface and the think aloud protocol is used. Experienced participants are likely to report fewer issues in comparison with novice participants due to their knowledge of how to 'work around' possible problems with the interface or the fact that they do not consider them to be 'issues' as such. Some of these limitations can be addressed by ensuring the number of participants is appropriate to ensure that statistical significance can be achieved. A purely quantitative study will require a higher number of participants. Recruiting large numbers of suitable participants can be challenging, but it depends on the methods used. It is perhaps easier to recruit a large number of participants for a survey than it is for an observational study. A study designed to obtain qualitative data can be conducted with a smaller number of participants. If the study design is appropriate, then the results can have a high level of validity.

Regardless of the goal of the research, a participant screener should be produced to ensure that the participants recruited for the study are appropriate. A screener defines the characteristics of those who will be recruited to participate in the study. Again, if we use the example of screen reader users, certain parameters around the type of assistive used, the frequency of use, competency and previous exposure to the platform being tested may be defined. This way ensures that the most appropriate participants are recruited for the study.

The environment being tested is also an important consideration. For example, a 'live' website could provide the most realistic context of use, but if a longitudinal study is being conducted, the changes in the interface may influence the results as new variables may be introduced. As far as possible, the interface being tested should be consistent throughout the study. Developing a stable testing environment for the research can provide a solution to this potential problem.

Before commencing a live study, a pilot session should be conducted under similar conditions to the live study by using the same protocol and materials. The pilot study could be considered a 'rehearsal' for the real study. Its purpose is to identify any methodological flaws in the study design before engaging with participants in the real study. It allows any issues to be resolved without compromising the results of the study. In particular, a pilot study for an interview would allow checking the phrasing and wording of questions to ensure that they can be clearly understood and are not leading. A pilot study for think aloud user-testing sessions would allow testing the instructions given to participants to ensure that they are feasible and can be completed, and checking that the number of instructions is appropriate for the length of the session.

If the research design—and subsequent investigation—is found to be ineffective, then the rigour of the research could be questioned due to the lack of reliability and validity. Reliability refers to the extent to which the results of the investigation (or a series of investigations) are consistent and can therefore be applied to the larger population. Validity consists of two components which are external validity and internal validity. External validity refers to the extent to which the conditions of the study represent what would happen in a live context. In contrast, internal validity refers to the extent that the study measured or observed the appropriate criteria to meet the aims of the research, as defined by the hypotheses and research questions. As previously discussed, reliability can be addressed by ensuring that a screener is used to recruit appropriate participants and the sample size is sufficient. External validity can be assured by ensuring that the participants conduct tasks which are typical of what they would conduct in a live context (the 'real world') and the variables are controlled. Again, as previously discussed, internal validity can be assured by ensuring that the hypotheses and research questions are clearly defined and the research design is sufficiently implemented to support these. Conducting a pilot study can assist in achieving these goals.

The goal of any user study is that it should be reproducible and can be validated by an external party. It is important to ensure that the research design (such as hypotheses and research questions) is well defined and the materials used during the study (such as a discussion guide) are produced and made available so that the study can be repeated. If we consider the fact that multiple iterations of the design of an interface are to be tested for benchmarking purposes, the conditions of the experiment may need to be replicated to measure either an improvement in user performance or decrease in effectiveness. If the results of the study in an established field contradict with other studies, then other practitioners or the original researchers may wish to examine the datasets or indeed repeat the study to ensure that the reliability and validity of the investigation can be verified.

11.4 Future Directions

People can now use a range of different technologies to interact with websites. In particular, people with motor impairments can use eye-tracking technology to interact with web pages (Menges et al. 2017). However, when the elements of web pages are very close to each other, users can experience some problems in fixating and selecting these elements by using their eyes due to the accuracy rate of eye trackers. All of these kinds of technologies and their limitations should also be taken into consideration during website development. When an end-user evaluation is conducted for a particular website, there should be a set of tasks which can test whether the website is accessible by using different technologies.

There have been some recent studies to predict whether a particular user has autism or dyslexia by using their eye-tracking data (Rello and Ballesteros 2015; Yaneva et al. 2018) After the successful prediction, websites can be automatically adapted or transcoded to be more accessible for these users by meeting their needs. Different approaches are available to transcode web pages such as page rearrangement, simplification and alternative text insertion (Asakawa and Takagi 2008). Therefore, these approaches should first be investigated with a series of user studies to determine the most suitable transcoding approach(es). In these studies, a sufficient number of participants should be recruited from the target population and they should be asked to perform various tasks on the original version and the transcoded versions of web pages for comparison purposes.

Although web accessibility guidelines are beneficial for web designers to develop accessible and usable websites, some of these guidelines, especially the ones for people with autism, have not been validated by using an empirical study with the relevant user groups (Eraslan et al. 2017a; Yaneva et al. 2018). Empirical validation with users would strengthen the reliability of these guidelines.

Different metrics have recently been proposed to analyse how users interact with web pages (Eraslan et al. 2014; Eraslan and Yesilada 2015). Specifically, Eraslan et al. (2016b) have recently proposed an algorithm called Scanpath Trend Analysis (STA) which analyses eye movements of multiple users on a particular web page and discovers the most commonly followed path on the page in terms of its visual elements as the trending path (Eraslan et al. 2016d, 2017b). This path provides a better understanding of how users interact with web pages in general, and it can be used for different purposes. For example, it can be used to transcode web pages to make trending elements more accessible (Yesilada et al. 2013) or it can be used to investigate whether users follow the expected path for a particular goal (Albanesi et al. 2011). If there are two different groups of users in the target population, the STA algorithm can be applied to the samples of these two groups separately and their results can be compared to investigate how these groups are similar to each other. In this case, researchers can recognise whether a particular website is used similarly by different groups of users.

11.5 Authors' Opinion of the Field

End-user evaluations are required to determine the true picture of the accessibility of the interface being tested. They should be considered a required supplement to other evaluation methods, such as conformance review against accessibility guidelines, as research has already shown the limitations of accessibility guidelines. In recognition of the need for formal accessibility guidelines to evolve to be more user centred, following the recent update of version 2.0 to 2.1 of the Web Content Accessibility Guidelines (WCAG), the Accessibility Guidelines Working Group of the W3C is developing another revision to the formal accessibility guidelines which follows a research-focused and user-centred design methodology with the aim of producing more relevant and appropriate guidelines (for more details about the guidelines, see the 'Standards, Guidelines and Trends' chapter). Testing with users will be required to define such guidelines. The aim of any user study could be to identify any—or all—of the following:

- Understand the accessibility requirements of the users and gauge their opinions;
- Identify accessibility issues in the interface for the users being investigated;
- Investigate the severity of the issues and
- Explore potential design solutions to any issues.

The last few years have seen a shift towards a human approach to accessibility, rather than one which recognises accessibility primarily as a binary technical requirement that a product either 'meets' or does 'not meet'; end-user evaluations will remain a core element of this. Accessibility may not be considered as an intrinsic characteristic of a digital resource, but it is determined by a range of complex political, social and other wider contextual factors (Cooper et al. 2012). The authors believe a true human-centred model of accessibility must consider the full range of their users' technical, operational and psychological requirements. This is one of the key principles of the Web Accessibility Code of Practice, which has been formalised into a British Standard (BS8878) based on a user-centred approach to accessibility, to provide products that are conformant with guidelines, usable and satisfying.

Accessibility should not be considered as a separate quality attribute in isolation. Accessibility, usability and user experience are all interdependent quality attributes of a product. They all need to be at an optimal level to ensure that the product can be used effectively.

We can consider the concept of accessibility as having three separate, but interdependent and overlapping components as given below (Bailey and Gkatzidou 2017):

- Technical accessibility: This component refers to the fundamental requirements for users to be able to access a product, services or physical environment. It includes conformance with accessibility guidelines and compatibility with assistive technologies. Therefore, this component represents the basic user needs.
- Operational accessibility: Once users have access, this component refers to how well they can use and operate the product or navigate the physical environment. It refers to attributes such as efficiency (e.g. can the users accomplish tasks in a reasonable time frame?), error rate and error recovery (e.g. how many errors do the users make? how well do the users recover from them?). This component also represents the extent to which the product or feature set meets the users' expectations.
- Psychological accessibility: Once users can access and use a product, services or premises, this component refers to aspects including but not limited to, how useful the users find its functionality or facilities, how appropriate they are for the users and how satisfying they are for the overall experience of the users. Therefore, this component represents the users' desires.

We consider the psychological element of accessibility to be of great importance. For some audiences, specifically older users, there may be no technical or operational barriers to accessibility when attempting to use a product; the barrier may be psychological and can be due to a general lack of confidence when using digital technology from the user. For example, users may have had a negative experience with an online banking service in the past and they may be unwilling to use, or may assume they cannot use, the service again today, despite the accessibility and usability being significantly enhanced during this time.

When considered together, the defined attributes of accessibility include those described in Yesilada et al. (2012). We emphasise that all components can be evaluated and measured using methods described in this chapter. For example:

- Guideline conformance reviews and testing with end users' assistive technologies can be used to measure technical accessibility.
- Observational or think aloud methods with users can be used to measure operational accessibility.
- Questionnaires and interviews designed to capture both quantitative and qualitative data can be used to measure psychological accessibility.

No single method can comprehensively measure all attributes or all of the aims described earlier. However, when a research study is carefully designed, it can be possible to obtain some insights.

Recent developments have suggested that future research could utilise crowdsourcing (Sect. 11.2.1.8) as a solution to remedy the issue of recruiting a statistically significant number and a sufficiently diverse range of participants to provide reliable and valid results (Kouroupetroglou and Koumpis 2014; Li et al. 2017; Song et al. 2018). We expect further research in this area to contribute significantly to this field.

11.6 Conclusions

This chapter highlights the need for end-user evaluations to provide accessible websites. It gives an overview of the commonly used evaluation methods. There is no unique method which is valid for all end-user evaluations. Some evaluations need to combine multiple methods, whereas some of them can be conducted with a particular method. In addition to the right selection of the methods, the study should be effectively designed, especially by choosing the representative sample from the target population, controlling external and internal factors appropriately, and carefully considering ethical issues.

References

Albanesi MG, Gatti R, Porta M, Ravarelli A (2011) Towards semi-automatic usability analysis through eye tracking. In: Proceedings of the 12th International Conference on Computer Systems and Technologies, ACM, New York, NY, USA, CompSysTech '11, pp 135–141. https://doi.org/10.1145/2023607.2023631

Asakawa C, Takagi H (2008) Transcoding. In: Harper S, Yesilada Y (eds) Web accessibility, Springer, London, a foundation for research, human computer interaction series, pp 231–260

Bailey C, Gkatzidou V (2017) Considerations for implementing a holistic organisational approach to accessibility. In: Proceedings of the 14th Web for All Conference on the Future of Accessible Work, ACM, New York, NY, USA, W4A '17, pp 7:1–7:4. https://doi.org/10.1145/3058555.3058571

Bevan N, Carter J, Harker S (2015) Iso 9241–11 revised: what have we learnt about usability since 1998? In: Kurosu M (ed) Human-computer interaction: design and evaluation. Springer International Publishing, Cham, pp 143–151

Blascheck T, Kurzhals K, Raschke M, Burch M, Weiskopf D, Ertl T (2017) Visualization of eye tracking data: a taxonomy and survey. Comput Graph Forum 36(8):260–284. https://doi.org/10.1111/cgf.13079

Borodin Y, Bigham JP, Dausch G, Ramakrishnan IV (2010) More than meets the eye: a survey of screen-reader browsing strategies. In: Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), ACM, New York, NY, USA, W4A '10, pp 13:1–13:10. https://doi.org/10.1145/1805986.1806005

Brajnik G (2006) Web accessibility testing: when the method is the culprit. In: Miesenberger K, Klaus J, Zagler WL, Karshmer AI (eds) Computers helping people with special needs. Springer, Berlin, pp 156–163

Breen RL (2006) A practical guide to focus-group research. J Geogr High Educ 30(3):463–475. https://doi.org/10.1080/03098260600927575

Burton MC, Walther JB (2001) The value of web log data in use-based design and testing. J Comput-Mediat Commun 6(3):JCMC635. https://doi.org/10.1111/j.1083-6101.2001.tb00121.x

Clegg-Vinell R, Bailey C, Gkatzidou V (2014) Investigating the appropriateness and relevance of mobile web accessibility guidelines. In: Proceedings of the 11th Web for All Conference, ACM, New York, NY, USA, W4A '14, pp 38:1–38:4. https://doi.org/10.1145/2596695.2596717

Cooper M, Sloan D, Kelly B, Lewthwaite S (2012) A challenge to web accessibility metrics and guidelines: putting people and processes first. In: Proceedings of the International Cross-disciplinary Conference on Web Accessibility, ACM, New York, NY, USA, W4A '12, pp 20:1–20:4. https://doi.org/10.1145/2207016.2207028

Dix A, Finlay J, Abowd G, Beale R (2004) Evaluation techniques. In: Human-computer interaction, 3rd edn. Pearson Prentice Hall, pp 318–363

Ehmke C, Wilson S (2007) Identifying web usability problems from eye-tracking data. In: Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - volume 1, British Computer Society, Swinton, UK, UK, BCS-HCI '07, pp 119–128

Eraslan S, Yesilada Y (2015) Patterns in eyetracking scanpaths and the affecting factors. J Web Eng 14(5–6):363–385

Eraslan S, Yesilada Y, Harper S (2013) Understanding eye tracking data for re-engineering web pages. In: Sheng QZ, Kjeldskov J (eds) Current trends in web engineering. Springer International Publishing, Cham, pp 345–349

Eraslan S, Yesilada Y, Harper S (2014) Identifying patterns in eyetracking scanpaths in terms of visual elements of web pages. In: Casteleyn S, Rossi G, Winckler M (eds) Web engineering. Springer International Publishing, Cham, pp 163–180

Eraslan S, Yesilada Y, Harper S (2015) Eye tracking scanpath analysis techniques on web pages: a survey, evaluation and comparison. J Eye Mov Res 9(1). https://bop.unibe.ch/JEMR/article/view/2430

Eraslan S, Yesilada Y, Harper S (2016a) Eye tracking scanpath analysis on web pages: how many users? In: Proceedings of the ninth biennial ACM symposium on eye tracking research & applications, ACM, New York, NY, USA, ETRA '16, pp 103–110. https://doi.org/10.1145/2857491.2857519

Eraslan S, Yesilada Y, Harper S (2016b) Scanpath trend analysis on web pages: clustering eye tracking scanpaths. ACM Trans Web 10(4):20:1–20:35. https://doi.org/10.1145/2970818

Eraslan S, Yesilada Y, Harper S (2016c) Trends in eye tracking scanpaths: segmentation effect? In: Proceedings of the 27th ACM Conference on Hypertext and Social Media, ACM, New York, NY, USA, HT '16, pp 15–25. https://doi.org/10.1145/2914586.2914591

Eraslan S, Yesilada Y, Harper S, Davies A (2016d) What is trending in eye tracking scanpaths on web pages? In: Spink A, Riedel G, Zhou L, Teekens L, Albatal R, Gurrin C (eds) Proceedings of the 10th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2016), Dublin City University, MB 2016, pp 341–343

Eraslan S, Yaneva V, Yesilada Y, Harper S (2017a) Do web users with autism experience barriers when searching for information within web pages? In: Proceedings of the 14th Web for All Conference on the Future of Accessible Work, ACM, New York, NY, USA, W4A '17, pp 20:1–20:4. https://doi.org/10.1145/3058555.3058566

Eraslan S, Yesilada Y, Harper S (2017b) Engineering web-based interactive systems: trend analysis in eye tracking scanpaths with a tolerance. In: Proceedings of the ACM SIGCHI symposium on engineering interactive computing systems, ACM, New York, NY, USA, EICS '17, pp 3–8. https://doi.org/10.1145/3102113.3102116

Eraslan S, Yesilada Y, Harper S (2017c) Less users more confidence: how AOis dont affect scanpath trend analysis. J Eye Mov Res 10(4). https://bop.unibe.ch/JEMR/article/view/3882

Eraslan S, Yesilada Y, Harper S (2018) Crowdsourcing a corpus of eye tracking data on web pages: a methodology. In: Grant R, Allen T, Spink A, Sullivan M (eds) Proceedings of the 11th International Conference on Methods and Techniques in Behavioral Research (Measuring Behavior 2018), Manchester Metropolitan University, MB2018, pp 267–273

Eysenck MW (2005) Psychology for AS level, 3rd edn. Psychology Press, Hove, East Sussex

Gay L, Mills G, Airasian P (2009) Educational research: competencies for analysis and applications, 9th edn. Prentice Hall, Upper Saddle River, New Jersey

Gravetter FJ, Wallnau LB (2008) Statistics for behavioral sciences, 8th edn. Wadsworth Publishing

Hennick M (2007) International focus group research: a handbook for the health and social sciences. Cambridge University Press, Cambridge

Henry SL (2018) Involving users in evaluating web accessibility. https://www.w3.org/WAI/test-evaluate/involving-users/. Accessed 15 Aug 2018

Hesterberg TC (2015) What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. Am Stat 69(4):371–386. https://doi.org/10.1080/00031305.2015.1089789, pMID:27019512

Holzinger A (2005) Usability engineering methods for software developers. Commun ACM 48(1):71–74. https://doi.org/10.1145/1039539.1039541

Jay C, Lunn D, Michailidou E (2008) End user evaluations. In: Harper S, Yesilada Y (eds) Web accessibility. A foundation for research, human computer interaction series. Springer, London, pp 107–126

Kouroupetroglou C, Koumpis A (2014) Challenges and solutions to crowdsourcing accessibility evaluations. https://www.w3.org/WAI/RD/2014/way-finding/paper5/. Accessed 9 July 2018

Kuniavsky M (2003) Observing the User Experience: A Practitioner's Guide to User Research (Morgan Kaufmann series in interactive technologies). Morgan Kaufmann Publishers Inc., San Francisco

Leavitt M, Shneiderman B (2006) Research-based web design and usability guidelines. Department of Health and Human Services, Washington DC, US

Leedy P, Ormerod J (2016) Practical research: planning and design, 11th edn. Pearson

Lewis C (1982) Using the think aloud method in cognitive interface design. IBM Research Report, RC–9265 (#40713), IBM Thomas J. Watson Research Center, Yorktown Heights, NY

Li L, Wang C, Song S, Yu Z, Zhou F, Bu J (2017) A task assignment strategy for crowdsourcing-based web accessibility evaluation system. In: Proceedings of the 14th Web for All Conference on the Future of Accessible Work, ACM, New York, NY, USA, W4A '17, pp 18:1–18:4. https://doi.org/10.1145/3058555.3058573

Menges R, Kumar C, Müller D, Sengupta K (2017) Gazetheweb: a gaze-controlled web browser. In: Proceedings of the 14th Web for All Conference on the Future of Accessible Work, ACM, New York, NY, USA, W4A '17, pp 25:1–25:2. https://doi.org/10.1145/3058555.3058582

Nielsen J (2003) Usability 101: introduction to usability. http://www.useit.com/alertbox/20030825.html. Accessed: 09 July 2018

Nielsen J (2004) Risks of quantitative studies. https://www.nngroup.com/articles/risks-of-quantitative-studies/. Accessed 01 July 2018

Pallant J (2007) SPSS survival manual: a step by step guide to data analysis using SPSS version 15, 4th edn. Open University Press/McGraw-Hill, Maidenhead

Poole A, Ball LJ (2005) Eye tracking in human-computer interaction and usability research: current status and future. In: Prospects, Chapter in C. Ghaoui (Ed.): encyclopedia of human-computer interaction. Idea Group Inc., Pennsylvania

Rello L, Ballesteros M (2015) Detecting readers with dyslexia using machine learning with eye tracking measures. In: Proceedings of the 12th Web for All Conference, ACM, New York, NY, USA, W4A '15, pp 16:1–16:8. https://doi.org/10.1145/2745555.2746644

Rosenbaum S, Cockton G, Coyne K, Muller M, Rauch T (2002) Focus groups in HCI: wealth of information or waste of resources? In: CHI '02 extended abstracts on human factors in computing systems, ACM, New York, NY, USA, CHI EA '02, pp 702–703. https://doi.org/10.1145/506443.506554

Rubin J, Chisnell D (2008) Handbook of usability testing: how to plan, design and conduct effective tests. Wiley, New York

Sherief N, Jiang N, Hosseini M, Phalp K, Ali R (2014) Crowdsourcing software evaluation. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software

Engineering, ACM, New York, NY, USA, EASE ’14, pp 19:1–19:4. https://doi.org/10.1145/2601248.2601300

Song S, Bu J, Wang Y, Yu Z, Artmeier A, Dai L, Wang C (2018) Web accessibility evaluation in a crowdsourcing-based system with expertise-based decision strategy. In: Proceedings of the internet of accessible things, ACM, New York, NY, USA, W4A ’18, pp 23:1–23:4. https://doi.org/10.1145/3192714.3192827

Yaneva V, Ha LA, Eraslan S, Yesilada Y, Mitkov R (2018) Detecting autism based on eye-tracking data from web searching tasks. In: Proceedings of the internet of accessible things, ACM, New York, NY, USA, W4A ’18, pp 16:1–16:10. https://doi.org/10.1145/3192714.3192819

Yen PY, Bakken S (2009) A comparison of usability evaluation methods: heuristic evaluation versus end-user think-aloud protocol–an example from a web-based communication tool for nurse scheduling. In: AMIA annual symposium proceedings, American Medical Informatics Association, vol 2009, p 714

Yesilada Y, Stevens R, Harper S, Goble C (2007) Evaluating DANTE: Semantic Transcoding for Visually Disabled Users. ACM Trans Comput-Hum Interact 14(3):14. https://doi.org/10.1145/1279700.1279704

Yesilada Y, Brajnik G, Vigo M, Harper S (2012) Understanding web accessibility and its drivers. In: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility, ACM, New York, NY, USA, W4A ’12, pp 19:1–19:9, https://doi.org/10.1145/2207016.2207027

Yesilada Y, Harper S, Eraslan S (2013) Experiential transcoding: An eyetracking approach. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, ACM, New York, NY, USA, W4A ’13, pp 30:1–30:4, https://doi.org/10.1145/2461121.2461134