A method study: using eye tracking as a supportive method in qualitative usability testing with people with and without autism

Moa Bogren
Södertörn University
Huddinge, Sweden
19mobo@suni.se

ABSTRACT
This paper describes a qualitative study with the goal of investigating if eye tracking is a relevant method when performing usability tests with respondents with autism. The study explores this through performing usability tests with two target groups, five respondents with autism and four without. The usability test is executed on the respondents' own smartphones and both screen and eye tracking is recorded. To get a deeper insight into the respondent reasoning throughout the interaction a retrospective recall report was added after the test was performed. The paper describes the procedure of the test but focuses on methodology findings. Eye tracking allowed us to not rely upon the respondents' capacity for verbal expression, the eye tracking video in itself was able to generate usability issues and insights without the respondents' verbal feedback. This shows a method suitable when testing with target groups who are not comfortable or able to verbally communicate and that opens up doors for performing usability tests with target groups which otherwise might not be included.

Author Keywords
Eye tracking; autism; universal design; inclusive design; retrospective recall report; concurrent think-aloud;

ACM Classification Keywords
Human-centered computing → Accessibility → Empirical studies in accessibility, eye tracking, autism

INTRODUCTION
Universal design strives for design optimised towards an audience as broad as possible [16, 17]. The concept of universal design is sometimes called inclusive design, design for all, accessible design and is clustered in the paradigm of User sensitive inclusive design [20]. In this paper we will refer to it as universal design. Universal design has been criticised, not for its purpose but for the reason that creating design suitable for "everyone" is not possible and therefore the concept of universal design cannot succeed, and designers should instead focus their attention upon more personalized solutions [16]. Other scientists working in the field argue that: *"Just as usability includes some appeal to universality, universal design acknowledges difference. Universal design is not about "one size fits all" in a narrow sense but about flexible and inclusive design*" [31]. Whatever you may call it, the common factor is the demand for more tangible guidelines and frameworks on how to succeed with it [5, 6, 14, 16, 20, 26].

Many of the studies existing in the realm of universal design focus on the end product being as accessible as possible, and very few focus on investigating and exploring methods suitable when executing usability tests with respondents with different disabilities such as autism. Autism diagnoses have increased over the last years and today 1–2 percent of the Swedish population has been diagnosed with autism [34]. The autism diagnosis has also increased worldwide [8]. The design community is in need of more knowledge about neurodiversity such as autism and other cognitive disabilities and how that affects interacting with digital interfaces. This lack of knowledge was alerted already a decade ago [25].

Autism or Autism Spectrum Disorder (ASD) affects attention, where people with autism can find it hard to focus on several things at the same time, which might result in challenges when interacting online on webpages [8]. Earlier studies also show that people with autism, especially children, might prefer interacting with computers rather than people [28] In the WCAG 2.1 [33] autism is included under the umbrella term cognitive disabilities [8]. Cognitive disabilities include cognitive neurological disorders or learning disorders such as dyslexia, dyscalculia, autism spectrum disorder, aphasia and Down syndrome, to some extent these symptoms are also age-related [25]. Cognitive disabilities are known to be the least discussed in WCAG 2.1 but also in other literature. Eraslan et al. [8] argues that few empirical studies exist that investigate how respondents with autism interact with search pages, the author investigates the interaction through eye tracking just as in this study but makes few reflections if eye tracking as a method is suitable when testing with respondents with autism [8].

According to the "eye-mind hypothesis" there is a strong correlation between what we're looking at and what we're thinking about, meaning that observing a respondent's gaze gives a great insight to what is on the respondent's mind [3, 8]. The hypothesis also claims that the time it takes for the respondent to process a specific object is the same amount of time the respondent's gaze is fixated upon the object [8].

The "eye-mind hypothesis" gives a strong foundation for using eye tracking methods with a target group who might not feel entirely comfortable with giving verbal feedback. The focus for this study is to investigate eye tracking as a potential method which favours respondents who are not comfortable with verbal expression such as people with autism. The paper will describe the process of performing qualitative usability tests with two target groups, one with autism and another without (neurotypical). The insights on the method are directed towards design practitioners and researchers working in the field of production design towards end users, and therefore several bullet points highlighting insights on how to execute an eye tracking study with respondents with and without autism will be presented at the end of the paper.

BACKGROUND AND RELEVANT WORK

Eye tracking

The use of eye tracking devices and systems allows for recording eye movement [1] and can be used as a tool to measure and analyse the user's gaze and attention [6, 30]. Eye tracking today is widely used to optimize e-commerce for better sales [30] and optimising layout and design elements [8]. Eye tracking has also been used in studies investigating how different types of context affect the users, such as dynamic content versus static content, and content with different complexity [8].

Autism

Except for having difficulties in social communication and interaction, people with autism often focus on details rather than the whole [8, 12]. They can also have difficulties in processing "context-relevant" information and making sense of it [21]. There are different levels of autism, being on the low-functioning level of the spectrum can cause severe learning difficulties while people on the high end of the spectrum can be highly able and have normal to high intelligence [8, 29]. Being on the high end of the spectrum is what is called high-functional autism which before was called Asperger syndrome [8]. People with autism can be sensitive to smell, lights, textures, sounds and colours [8] which was something that we got to experience in this study when we got feedback on the facility for the test.

Retrospective reports and think-aloud

The "Concurrent think-aloud" (CTA) methodology is one of the most widely used methodologies in usability research and was founded by Ericsson and Simon [9]. The CTA methodology is grounded in human cognition, the user can verbally communicate and reflect upon what is happening in the interaction based on fetching the information from short term memory [3, 6]. The methodology has its flaws, the most argued flaw is that the user can edit the verbal response and it doesn't become a valid representation of the true user's experience [3, 7]. Other flaws like interrupting the respondent can cause a negative impact on the respondent's task performance and also distract the respondent's attention and concentration [15]. Think-aloud is still widely used in usability research, and is even called the "single most valuable usability engineering method" [11] but is often needed to be combined with observational analysis, this to both take advantage of what the user is saying but also make conclusions based on behaviour observed during a test [11, 19].

To overcome the obstacle of interrupting the respondent during the session which can be created with the CTA protocol, the respondent's feedback can be triggered by visual cues after the performance of the usability test is done. This is called stimulated recall, stimulated recall takes the advantage of visual cues [24] and in this study we use videos of the respondent's gaze as visual cues to trigger the respondent to remember and recall what they were thinking about during the interaction. A version of this is retrospective think-aloud protocol (RTA). RTA meaning that the user will recall from memory after the test is done, this can of course also lead to biases in the form of fabrication and recall errors [3, 7]. Adding eye tracking to this adds another dimension based upon the "eye-mind hypothesis", that there exists a strong correlation between what a person is looking at and what the person is thinking about [8]. But this methodology like the others accounted for is not without flaws, eye tracking is very good at showing where the user is looking but not why. So to add the why into the where, the Retrospective Reports and the CTA methodology can be combined, to let the user verbally explain more about the interaction which took place during the test. Adding these two methods together is a common way to qualify an otherwise quite quantitative study [3, 7]. Earlier studies show that the combination of retrospective reports with the aid of eye tracking and think-aloud protocols caught more usability issues and more quality comments than using CTA [3, 11]. On the other hand a study made by Sanne Elling, Leo Lentz, and Menno de Jong [7] shows that there were no differences in usability issues found using CTA and RTA. This study does not aim to investigate the number of usability issues found and does not compare the two methods, it takes advantage of the RTA protocol with the aim of exploring the methodology while executing usability tests with respondents with and without autism. The method of RTA may also be called post-task testing, retrospective protocol, retrospective report and think after [15].

Methods used in universal design

Even though most studies within universal design have come to the conclusion that more easy to understand frameworks and guidelines are necessary to make inclusive and accessible design a natural part of everyday design processes, some studies focus on the methodology itself. The studies which have more of a theoretical focus are also the ones that have been focusing on inclusive and accessible design methods rather than accessible design solutions. These studies highlight the importance of making it easier and more natural for design practitioners to include respondents with cognitive disabilities [4, 5, 14, 16, 20, 26].

Methods in focus in earlier studies are remote usability testing [27] and participatory design [5, 13, 14], and how they are suitable methods both for the benefit of the end result (the design artefact) but also for the involvement of the respondent with cognitive disability. In this study neither remote usability testing or participatory design will be the method of evaluation and instead the focus will be on using eye tracking technology and RTA to explore and discuss why these methods might be suitable when testing with respondents with cognitive disabilities.

METHODOLOGY

The focus of this study was to investigate if eye tracking and RTA methods are suitable methods when testing with people with autism. This was done by executing usability tests with two target groups, one with autistic, the other with neurotypical respondents and focusing the analysis upon the outcome of the method for the two groups and how the respondents experienced the sessions.

Research questions

Is eye tracking and RTA a suitable method when executing qualitative usability tests with people with autism?

This paper also describes the experience of planning and executing a study using eye tracking and RTA with respondents with autism, and what to think about for future researchers and designers.

SBAB

The study was executed together with SBAB. SBAB is a government owned bank and provides what could be considered a crucial societal service. SBAB's main service is mortgages, and they provide no physical services in bank offices, only on telephone and digitally, which of course puts more demands on accessibility on their existing channels.

SBAB is required to fulfill the European accessibility act [10] which has a due date in 2025, and sees this as an opportunity to both contribute to more research in the area of accessibility, to provide knowledge to the field but also as a chance to develop internal guidelines and frameworks to help their designers make accessibility a more natural part of the design and product development process. SBAB.se, their website, is one of the first things a potential customer sees, and becomes an important communication channel, therefore a task involving finding crucial information at SBAB.se will be the task at focus for this study.

Based on how the traffic is divided between devices and how the traffic has been evolving towards more mobile uses on SBAB.se (the page where the respondent got to solve their task) a decision was made to let the respondent perform the task on their own smartphones. Another argument for this was that the respondents would feel more comfortable interacting on their own phones rather than on a fixed computer which would have been necessary if the test was executed on a desktop.

Pilot tests

Two pilot tests were executed, none of them with autistic respondents, so it mostly became a technical rehearsal which was much needed. During the pilot tests it became evident that the eye tracking glasses could not be used together with regular glasses, which meant at the last minute we had to exclude people who could not execute the test without their regular glasses or wear contact lenses. During the pilot test we also saw problems with the quality of the recording so a decision was made to add secondary screen recording. This screen recording meant that the respondent had to use an application on their phone to be able to record their screen, adding another time consuming activity to an already tight test-schedule. The quality of the eye tracking recording was then no problem during the actual tests. There was also a problem with the connection between the glasses and the computer, it connects via WiFi, so to secure the connection during the actual test with the strained schedule an Ethernet cable was connected instead of relying upon the WiFi connection. During the pilot test the level of light on the respondent's screen became a problem, a screen with high brightness resulted in reflections in the recording which led to problems seeing what was on the screen. This was something which did not occur during the actual tests. These pilot tests were absolutely necessary to both minimize the risk of technical malfunctions during the tests but also as a rehearsal for the moderator.

Facility

The facility was of high importance because of the Covid-19 virus roaming the world at the time. Respondents with autism are also considered a risk group for Covid-19 so because this study had to be executed physically due to the eye tracking, the facility had to be secured in the best way possible. The facility, just like the equipment, was loaned out by a company called Conversionista which has their offices located in downtown Stockholm. They could also provide the necessary security and had the policy that everyone in their facility needed to Covid-19 test themselves before entering the premises. This also added logistic complications because the lab which performed the test was only available two days a week for three hours, this meant that we had to spread out the test over a longer period of time and do them early in the morning when the lab staff was there.

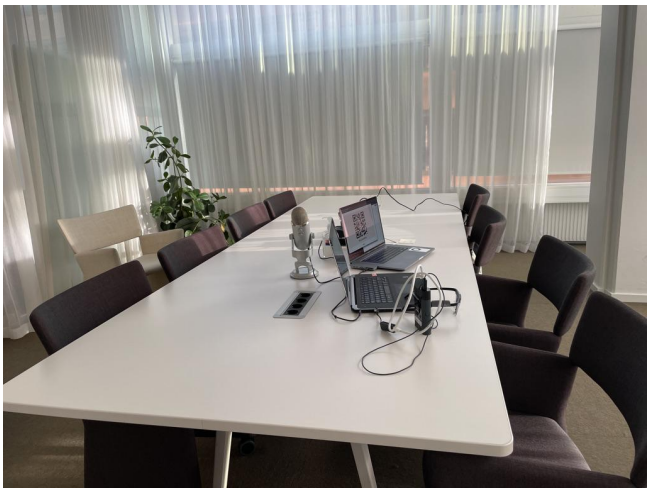

Figure 1. The room which was used two out of three days

Technique and material

The hardware used for this study was the Tobii Pro Glasses 2 [32]; a third edition of the Tobii Pro glasses is available on the market at the moment of the test but was not available for this study. This means that any limitations that were caused by the hardware might not be relevant when using the Tobii Pro Glasses 3. The Tobii Pro Glasses 2 consist of cameras, projectors and algorithms. The projectors shine an infrared light straight at the eyes, the camera then takes high resolution images of the user's eyes and creates a gaze pattern. Then machine learning and algorithms are used to determine the positions of the eyes and the gaze point.

Figure 2: The Tobii Glasses 2, wide angle HD scene camera, gyro and accelerometer, 4 eye cameras and microphone.

One of the main benefits using the Tobii Pro Glasses 2 is that tests can be performed in the field to represent a more realistic interaction. In this study the glasses were used in a fixed lab setting connected to a computer for the moderator to be able to observe the eye tracking and the interaction while it was going on. The benefits of being able to test out in the field will not be evaluated in this study.

Figure 3: Tobii Pro glasses 2, an adapter which stores the data, and the computer connected to it.

An additional application was added as a way to secure a good quality screen recording of the respondents' screens but ended up only as a comparison datasource to the eye tracking videos.

The software used during the study was the Tobii Controller, which is the recording software installed on the computer itself. Tobii Controller manages the set up of the project, the calibration of the glasses (calibration is needed to correctly measure the respondent's gaze) and acts as the video player during the RTA.

The analysis software used was Tobii Pro Lab, Tobii's most complex analysis program, optimised for quantitative eye tracking studies. The Tobii Pro Lab was used mostly for evaluating purposes. Tobii Pro Lab is quite expensive, so evaluating if it is relevant for a study of this scale is important knowledge for forthcoming studies when balancing investment and outcome.

The sound was recorded on a secondary computer through QuickTime Player.

Participants

Several Autism associations were contacted to recruit respondents but without luck, none of the associations were available to participate even though they seemed to believe the study was of relevance. In the end a total of ten respondents were recruited with the help from a recruitment agency. The target groups will be referred to as Group A and Group B, where Group A consists of respondents without autism and Group B consists of people with autism. Which respondent belonged to which group was not revealed to the moderator during the usability tests, the groups were first revealed after the test and which group was which was revealed after most of the analysis was done, this to reduce as much bias towards any of the target groups as possible.

The respondents were recruited with criteria such as spread in gender, age, financial situation and work. They were also

recruited with the criterion of having a mortgage since before, but it became evident that a criterion as narrow as that was not possible when also adding the criterion autism. Only one respondent from Group B had a mortgage since before, this created a knowledge gap between the two groups which became a liability when analysing some of the results from the study. When we found this out it was too late to recruit new respondents without mortgages for Group A. It took the recruitment agency eight weeks to recruit for Group B compared to the two weeks it normally takes. It was obvious that respondents with autism were not a user group they were used to nor did they have that in their database since before and it became an entirely new domain for them. Because of this no other criteria could be added to Group B, not on top of them having autism. So the level of autism could not be defined, this to increase the chance of recruiting enough respondents for the study. Complication in recruiting respondents with disabilities has been shown to be one of the biggest obstacles in earlier studies done [27] and it required both more time and financial resources than recruiting neurotypical respondents for this study. The two groups got the same incentive so the additional cost was because Group B was harder to recruit and would require more time invested for the recruitment agency.

Ten respondents were recruited, nine showed up, one respondent got sick during the last day of tests and had to cancel. Because the respondent had to schedule a Covid-19 test at the facility prior to the test it was not possible to book any stand-ins.

Procedure

To gain insights about the respondents earlier experience interacting with digital devices and services the respondents answered a survey prior to the test.

The nine tests were spread out over two weeks and three days, to be able to match when the Covid-19 testlab was available at the facility. The respondent first got to sign a consent form agreeing that SBAB could use the data to optimise their digital services and that the respondent could retract the data at any time.

After the respondent had signed the consent form the session started with a short pre-task interview, which only served the purpose of warming up the respondent for the usability test. In the interview the respondent answered questions about their living situation and earlier experience when it came to buying a home. After that the calibration of the eye tracking glasses was done, the respondent tested the glasses and focused their gaze on a small dot located on a card held up by the moderator. This had been a problem during the pilot test but worked almost flawlessly during the real tests, it caused complications only with one of the respondents but worked better after correcting the fit of the glasses. After that the respondents connected their phones to the screen recording software as a backup if the eye tracking recording did not provide screen recordings with quality good enough for analysis. Calibration and setting up the screen recording took more time than the usability test itself. The screen recording showed to be unnecessary because the eye tracking video did capture videos with good enough quality for analysis, but the screen recording was instead used to compare the benefit of using eye tracking instead of just relying upon screen recording. The respondent then got presented with a scenario "You're about to buy a home, and you're in the process of comparing interest rates between banks. Use SBAB.se (SBAB's webpage) to see what interest rate they can offer you" and got to use SBAB.se to solve their task. The completion time ranged from 39 seconds to 10 minutes. After that the respondent got to take part in the RTA, where the respondent and the moderator went through the eye tracking recording.

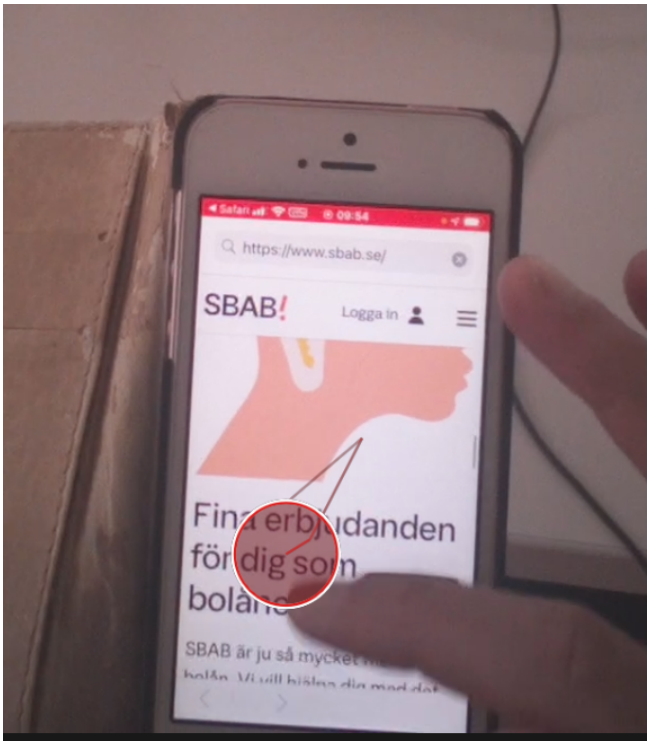

Figure 4: A snapshot from the eye tracking recording

The moderator would stop the recording when it indicated a point of interest such as when the respondent's gaze stopped for a longer period of time at an element or when the respondent moved their gaze rapidly between elements.

In the end of the session the respondent got to answer if they would be comfortable using this page again and rank their satisfaction with solving this task using SBAB.se. After the analysis of the data it became clear that the most important findings would be about the method and not on the differences between the two groups in gaze patterns. A survey was then sent out to the respondents to gain extra insight into how they had experienced the study. It would have been preferable to gather these insights through

qualitative interviews to get a deeper insight into the respondents' experience, but a survey with some open ended questions was the method of choice due to it being time-efficient and financially less expensive.

DATA ANALYSIS AND RESULTS

The data analysis was performed with both quantitative and qualitative methods. The conclusion and the foundation for this study relies on qualitative analysis but is supported by quantitative values. The metrics were based on common usability metrics such as completion time, number of tries and satisfaction rating [19]. The ones connected to eye tracking were based on what could be extracted from the Tobii software (the fixation metric) and phenomena and user interaction which were discovered along the way going through the eye tracking videos from a more qualitative perspective. This became somewhat an experimental process, adding and discarding metrics which did and did not add understanding into the respondent's gaze behaviour and usability problems.

Metric:	**Explanation:**	**Collected through:**
Completion Page	On which page on SBAB.se did they end the task.	Eye tracking video and Screen recording video
Completion Focus	User's focus on the completion page.	Eye tracking video and Screen recording video
Task time	Time it took to complete the task	
Tries	Number of tries/ways to find the "right" way.	Eye tracking video and Screen recording video
Content Focus	Preferred content focus on SBAB.se	Eye tracking video
GazeType	How did the respondent gaze on the webpage.	Eye tracking video
Fixations	Number of fixations	Tobii Pro Lab analysis software
Satisfaction rating	Satisfaction rating solving the task.	Thematic analysis and transcription
Would use it again	If they would use the page again.	Thematic analysis and transcription
Understanding of scenario	To what extent did the user understand the task they were given.	Thematic analysis and transcription

Table 1: Metrics and from which datasource they were extracted from.

Content focus

Through the eye tracking recordings we saw that only one respondent out of the nine focused on images rather than the text during the session. The images were also in the form of supportive illustrations rather than photographs. There was no indication that there was any difference in satisfaction rating or completion time for the respondent who tended to focus on images/icons rather than text. Earlier studies show that respondents with autism tend to focus on visual elements, headers and footers [8], we did not see any of these indications during these tests. The respondent who focused on images/icons belonged to Group A (the target group without autism).

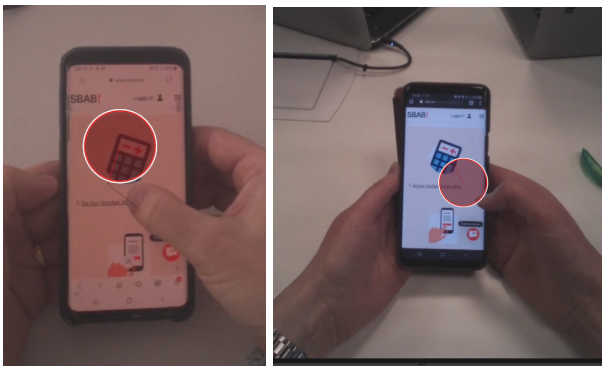

Figure 5. Respondent 1 (R1) from Group A focusing on image/icon; to the right R9 (group B) focusing on a text link rather than image/icon.

GazeType

A GazeType scale was created to indicate if the respondent was more prone to reading or if their gaze behaviour leaned more towards rapidly skimming the page. GazeType 1 indicates a respondent more prone to reading and the other end of the scale respondents more prone to rapidly skimming the page. A respondent in the lower part of the scale had more of an F-pattern tendency, meaning that they read the content from the far left to move to the right, down and do it all over again [22], and the respondents on the other end of the scale had more of a layer cake pattern, jumping from headline to headline [23].

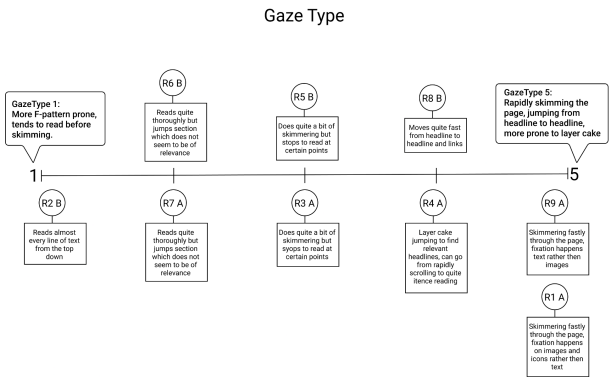

Figure 6. Respondents placed out on the GazeType scale, which represents what type of gaze pattern they had.

There was no correlation between where the respondent ended up on the GazeType scale and which group they belonged to, meaning that no conclusion can be drawn whether respondents with autism are more prone to rapidly skimming the page or reading. One thing worth mentioning is that R2, the respondent who ended up on the far left on the scale, was also the respondent which had the longest completion time (over 10 minutes, 6 minutes more than the second slowest time) and was the only respondent who did not complete the task, but still ranked 8 on the satisfaction ranking. Even though we could extract different GazePatterns through the eye tracking recordings it is hard to determine without further studies what impact the different gaze type behaviours have on design guidelines. There are also too many variables which can impact on why the respondents ended up on one side or the other on the GazeType scale, it can't be said that just because R2 ended up on the far left side of the scale in this study, that R2 would show the same behaviour on another webpage, solving another task, using another device or even in another situation solving the same task.

Fixations

The number of fixations the respondent made during the session was extracted from Tobii Pro Lab. Fixations mean gaze fixations, how many positions in total the respondent put their eyes on during one session. Tobii Pro Lab only runs on Windows 10 Pro or Enterprise, it runs only on Intel i5 6th generation or later processor, it requires 16GB RAM or more. Tobii Pro Lab is not required to qualitatively analyse the eye tracking data that can be made in Tobii Controller or the Tobii Manager, which is free to download from Tobii's webpage. Tobii Pro Lab is needed to make the more complex quantitative analysis. For this study it was used to extract the number of fixations every respondent made during one session. In this study the eye tracking recording started before the respondent started the actual task which meant that the early fixations happening were on the moderator's face while talking and not on the interface itself.

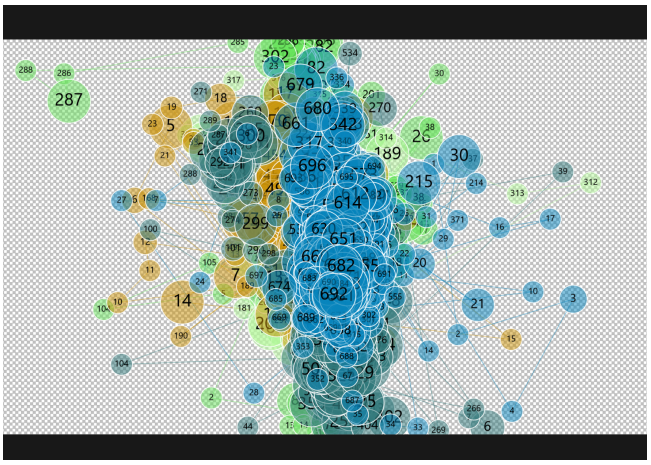

Figure 7. Aggregated Gazeplot visualisation of all the respondents belonging to group B over their entire sessions. The fixations which are outliers outside the main cluster have low index-numbers, which means they happened during the early stages of the session, when the moderator and the respondent still had a conversation.

If the eye tracking recording would have started when the respondent started solving the real task the number of fixations could have been extracted just by exporting a matrix of quantitative data. Even though that would have been done it is hard to make sure that the respondent does not ask any questions during the session or that something else catching the respondent's attention during the session messes up the number of fixations that actually is connected to the interaction itself. The number of fixations was instead extracted by looking at the eye tracking recordings and matching it with the Gaze Plots visualisation. The eye tracking videos in Tobii Pro Lab are time stamped and show an index number which indicates in which order they happened. So by identifying which index number the fixation had when the respondent started the interaction with SBAB.se the total amount of fixations could be extracted from every respondent. A *t*-test was done to see if there was any correlation between number of fixations and Groups, the *t*-test gave a *p*-value of 0.1582 so no indication that any target group ended up with more or fewer fixations. There were no patterns in satisfaction rating and number of fixations, nor could any patterns be found between number of fixations and what type of content the respondent was looking at. Overall no pattern could be found between the number of fixations and any of the metrics. So using Tobii Pro Lab to extract a number of fixations was in the end of no use to establish differences between the two target groups or in creating design guidelines. The only thing it resulted in is the insight that an eye tracking study of this size might not need to rely upon expensive software such as this and can instead use the free to download software.

Respondents' feedback on the method

The survey which was sent out was answered by 6 respondents out of the 9. The questions focused on how the respondent had experienced the study, from the facility to

the description provided by the researchers before the test. All of the respondents answering the survey answered that they could see themselves participating in another eye tracking study after this one. The respondents who answered the survey all thought the facility was “ok” or “great”. The only negative feedback we got related to the facility was from the respondent who had left after the Covid-19 test because of it being too noisy and loud (this respondent was replaced by another one). One of the respondents commented on the facility as “great” and mentioned how nice it was to be able to get a Covid-19 test and how that made them feel safe. The majority of the respondents thought the description had been “ok” but that it could have explained a little more in detail things like “ how to use the intercom” and where to go after the Covid-19 test. No one commented on wearing the glasses as something which made them feel uncomfortable nor watching their own interaction afterwards during the RTA session, something which has been problematic in earlier studies [3], 33% even answered that it was “exciting” and a “nice experience” to both wear the glasses and to watch their interaction on replay. Even though none of the respondents felt uncomfortable wearing the glasses or during the RTA session, 50% would have preferred thinking out loud during the task instead of reasoning afterwards while observing the interaction. We can only speculate about why the respondents seem to prefer the CTA method over RTA. It could be because of the experience in this specific study or earlier more positive experience with the CTA method which was mentioned by one of the respondents during the post interview.

The majority also answered that if they would have done the same task at home they would have preferred doing it on a laptop, which indicates a different behaviour than the traffic report mentioned earlier showing that the traffic over the last years has increased towards more mobile uses on SBAB.se. This indicates that the choice of doing it on their smartphone might not have been the most representative, one respondent also answered that they would never do it at home because it wasn’t relevant for them. What could also be speculated about is whether the respondent answered that they would prefer using a laptop while solving the task is based upon their experience solving the task on their mobiles during the test or that they actually thought they would prefer doing it on laptop for other reasons.

Usability problems/insights

Even though the best findings were not related to usability issues but about the methods itself we did find some rewarding usability insights too. They were not really usability problems; they were more insights into interaction patterns.

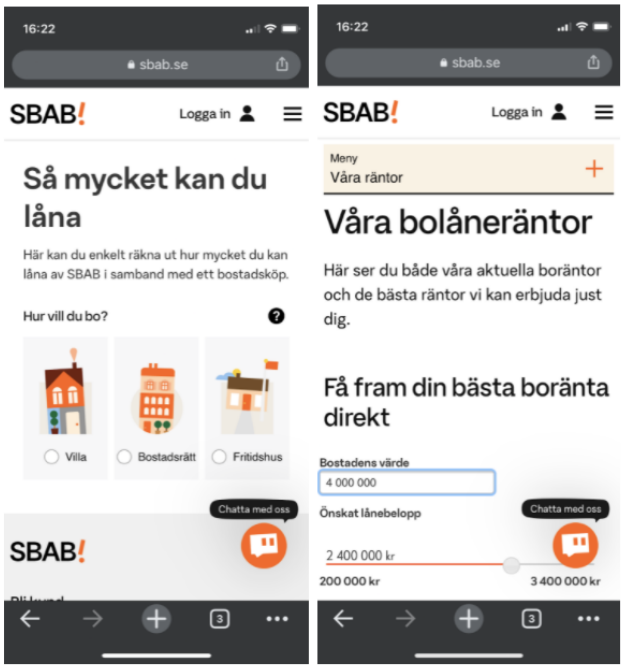

Figure 8. To the left “What can I borrow” and to the right “Our interest rates”, the two pages where the respondents considered they found the answers to the task.

The respondent found their answer either on the page “what can I borrow” or “Our interest rates”. It didn't matter on which page they solved the task; it had no effect on the overall satisfaction rate or completion time. But the respondents who solved the task on the “What can I borrow” tended to focus more on the monthly cost rather than the interest rate (which was the question), they believed they had solved the task and seemed completely satisfied with only focusing on the monthly cost. Maybe monthly cost is more important than just the interest rate alone. The respondents who ended up at “What can I borrow” and felt satisfied with monthly cost belonged to different target groups and none of them were respondents who had mentioned the scenarios to be hard to understand.

Headlines were the main focus point for many of the respondents but images were not. In the places icons/illustrations were together with text and text links the focus was almost every time on the text and the respondents barely gazed at the image/illustration. The respondents looked for words related to interest rate such as “counting”, “calculating” and so on, and they tended to not think that the images could provide them with that information and relied solely upon the headlines. Only one respondent mentioned the icons and said they were looking for an illustration of a calculator or a percent sign, that was the respondent’s indicator that they had gotten to the right place. What role icons, illustrations, and photographs have in finding the goal in the most effective way has to be

investigated further. But for this study images/illustrations or icons didn't play any important role in what way or how effectively the users found their goal. What role images versus text has might differ depending on the user-scenario. It might differ if the user uses SBAB.se to get an overall understanding for SBAB.se or, as in this study, to solve a very goal-driven task. This cannot be answered by this study but might be worth investigating further. Earlier eye tracking studies argue that people with autism focus more on visual elements [8] and that motivates further studies into the question what role do images/icons and illustrations have for making sense of the context on a webpage. Because most of the respondents focused on headlines, hierarchy becomes very important when clustering the content. These are no new findings, that is just a regular UX guideline, but it becomes validated when looking through the eyetracking videos that much more energy has to be put on creating hierarchical structures and copywriting to guide the users effectively towards their goal. This for both the target groups with or without autism.

Even though all of the respondents got the same scenario/task to solve, completion time differed from 39 seconds to over ten minutes. This could be the result of us not being able to recruit respondents with the same experience of buying a home or it could have something to do with the interface itself. Four out of the nine respondents solved the task on their first try and had an average completion time of 4,77min. One respondent solved the task trying two different ways and had a completion time of 3,28min. Two respondents did three tries and had an average time of 4,05min. This shows that making more tries doesn't necessarily mean a longer completion time. Completion also didn't have any effect in the satisfaction rating and neither did the number of tries, meaning that spending longer time solving the page or making more tries does not affect how satisfied the respondents felt about solving the task, it might come down to what content they came across and how well it is communicated and how well the functionality works to calculate the interest rate.

R9 (Group B with autism) mentioned that "I see things linearly, I believe I should do something in a specific order, I take the thing that comes first as the thing I should click on first". On the desktop these elements which R9 got stuck on are positioned next to each other and on mobile just one is visible after the other. This gives insight into the importance of thinking about which order clickable elements and communication end up on the screen. R9 said that buying a home is so complex that guiding is much needed and that the layout can help guide the users in which order they should do things. Even if a right order doesn't exist it might be helpful to create one. R9 was also one respondent who tended to very rapidly scroll the page up and down to help R9 to navigate and to feel comfortable with the page.

RTA and Mind eye hypothesis

The RTA sessions where the respondent got to reason about their interaction with the support of the eye tracking video after the task was executed gave insights into the respondents' minds but were also harder to execute than a regular CTA. Just like thematic analysis it does not become better than the themes found [19]. What became evident was that it was much harder to determine what gaze behaviours to act upon, and when to ask the respondent to reason about their behaviour, than in a CTA session where the moderator can act upon what the respondent is saying rather than solely based on gaze cues. The RTA session gave some insights about how the respondent had reasoned but very few insights which could not have been extracted just by looking at the eye tracking recording on its own.

With the mind eye hypothesis in mind, [7, 8] the eye tracking video can be analysed differently than the screen recording itself. Based only on the eye tracking recordings and not relying upon CTA or RTA, eye tracking recordings can provide an insight into the respondents' minds which a screen recording cannot.

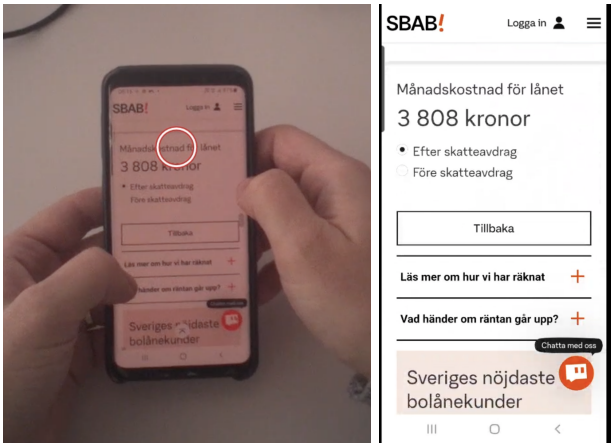

Figure 9. Here we see the eye tracking recording to the left and the screen recording to the right.

Looking at the image above (Figure 9), with the help from the image from the eye tracking we can draw the conclusion that the respondent is thinking and focusing on the monthly cost, based on the screen recording only it is impossible to determine where the focus points are for the respondent. The screen recording is dependent on a respondent who is capable of verbally expressing themselves during an CTA whilst the eye tracking video can be analysed without the respondents verbally expressing themselves during the session or after during an RTA. As mentioned earlier the problems with CTA have been that it might not represent a real interaction and that it can interrupt the user while solving the task [3, 7, 15] and for RTA the problem can be recall errors and fabrication of memory [3, 7]. Even though the RTA might make it possible to go even deeper into how the respondent reasons during the session, being able to only rely upon the eye tracking recording opens many doors

into testing with respondents who are not able or comfortable with verbally expressing themselves and at the same time also observing a more representative interaction.

DISCUSSION

Recruitment and demystifing cognitive disability

One thing which becomes clear is the problems of recruiting respondents for a study like this, including respondents with disabilities. Bohman and Anderson [2] put their focus on using universal design as a way to demystify cognitive disabilities, the authors argue that it is needed to progress in creating inclusive tools [2]. “cognitive disabilities must be demystified before any progress can be made toward developing tools that help make content accessible to people with cognitive disabilities” [2]. To be able to rely upon methods with user involvement, researchers and designers need to think about the entire process. By demystifying cognitive disabilities it might open up for more users with cognitive disabilities to take part in usability studies without feeling exposed.

Another factor which is mentioned by Pichiliani and Pizzolato [25] is the low demand from clients on creating inclusive and accessible digital products. Pichiliani and Pizzolato [25] refer to clients as stakeholders from an agency perspective and not internal stakeholders. Why this came to be is described 1) because of cultural barriers. 2) hard to consider people with cognitive disabilities as customers [25]. This indicates that the demand needs to come from designers and researchers independent of the clients being internal or external stakeholders.

Facility and surrounding

As mentioned earlier a lot of planning went into making sure the facility was Covid-19 safe but what was missed was the surroundings in the facility. As Erasalam et al. [8] mention, people with autism can be sensitive to bright lights, noises and get stressed in environments where a lot of things happen at the same time [8]. This was something which we got to experience when one of the respondents who didn’t show up had gotten to the facility, taken the Covid-19 test and then left due to unclear descriptions on where to go and due to it being too noisy with music and people. This is an important lesson even though most of the respondents answered that they felt comfortable with the facility. But not only the method needs to be evaluated, also the facility and the surrounding need to be adapted to make the respondent comfortable. It is interesting to think about how the respondent would have felt if they could have done it remotely, earlier studies show that remote usability testing is a suitable method when testing with respondents with cognitive disabilities [27], letting the respondent be in an environment where they feel at home and comfortable. At the moment there are logistic complications with letting the respondent do the test in their own home. The eye tracking equipment is quite expensive and takes some time to set up and to let the respondent do that on their own would be to put too much responsibility on the respondent and would create a huge liability when it comes to setting up the eye tracking equipment the right way. But maybe in the future it will be possible when regular smartphones and applications can allow the users to share eye tracking data straight to the researchers without the respondent leaving their home, and that might be a huge step in the right direction for inclusive methods and universal design.

Scaling an eye tracking study

From the beginning it felt like an eye tracking study would come with huge financial cost and require more planning than it is worth. For this study a one month Tobii Pro Lab license was bought (3750 SEK), in the end Tobii Pro Lab wasn’t a necessary software for a study of this small scale. From Tobii Pro Lab we extracted the number of fixations the respondents had during their sessions, but it didn’t contribute to any deeper understanding about the respondents' interaction and behaviour which the Tobii Controller (eye tracking videos) couldn’t generate. What gave the most value was the eye tracking recordings generated from the free to download Tobii controller and how the respondents' gaze indicated what was on their mind without the support of verbal feedback.

What has to be considered is the balance between what the method can contribute and how much work it will result in for the respondents and the moderator. An eye tracking study can be scaled to suit projects with different needs but some factors remain as constants even though you are executing a small or a big study. Almost the only fixed financial cost is the price of the hardware, that cannot be scaled based on the size of the project whilst recruitment of participants, the use of Tobii Pro Lab or not (the expensive analysis software) and cost of the facility can be scaled. Tobii presents very few prices on their webpage so without contacting Tobii or third party retailers it is hard to calculate on hardware cost.

But it is not only the financial investment that has to be considered; it is also about time spent for the moderator and the cognitive load for the respondent. In the survey done with the respondents of how they experienced the method the majority said that they prefer the CTA method before the RTA method. The RTA method reduces interruptions during the session and it also allows for more accurate representation about the respondents’ behaviour [15]. CTA is the more common method which might indicate that researchers have more knowledge of working with that method, which means it takes less time to get familiar with that method rather than RTA. An experience from this study and based on earlier experience is that it is also easier to know what to act upon based on speech rather than on gaze alone.

So there is no simple answer on what method to choose when but there are parameters which should be taken into consideration when choosing. The moderator’s prior experience with CTA, eye tracking and RTA (time to learn), what is the most suitable method for the respondents

(cognitive disability or not), the size of the project (will you be needing Tobii Pro Lab or will you be focusing more on qualitative analysis methods).

Another relevant perspective on this is to use CTA and eye tracking. To let the user think out loud while executing the task and wear the glasses at the same time. Letting the respondent choose the method they would prefer but at the same get the benefits of eye tracking but without the RTA. Using eye tracking as a support data source while executing a CTA test doesn't solely rely upon the respondent speaking the truth while thinking out loud (because the eye tracking can function as another source of truth) or rely upon the respondent's ability to verbally express themselves.

There is almost as much to think about planning and administering the study as in performing it. One lesson from this study is the more you plan ahead the less you have to worry about during the actual test. Take the time to pilot test, prepare the facility and make sure you don't add unnecessary time consuming elements which will take time from valuable session-time and things which can make the respondent exhausted. Some guidelines for future studies. .

What to think about before starting up:

- How do we get a hold of the equipment? Do we buy or can it be borrowed/rented from somewhere?
- Do we have time to understand the technique and the methods? Is there anyone we can ask for support?
- Do we have the facility needed to make the respondents comfortable, where the respondents can sit on their own while waiting for their turn?
- Always take time to do a pilot test if you have never done it before, get used to the technique, not getting enough or valid data is a waste of time.
- Make sure you decide before if you want the respondents to turn notifications off which can be alerted during the study from other applications. If you're interested in how they shift focus attention or if it's just going to be disturbing
- Think about devices, what device is most representative? If the respondents are not using their own computers/mobiles, make sure they feel as comfortable with the device as possible, some respondents might be more comfortable using a mouse rather than a trackpad, some might be more familiar with mac and some with PC and so on.
- Plan which metric you want to extract from the data. Will you be needing Tobii Pro Lab or can you rely upon the quantitative data which will be provided by the eye tracking recording itself?
- Make sure to send out information in good time prior to the study. One feedback we got is that people with autism need really concrete information on how to get to the lab and what will happen when they get there. You can't be too detailed.

What to think about during the test:

- Clear out the room, don't have any disturbing things which can catch the respondent's attention such as moving things, if there are windows you might want to cover them.
- Make sure the respondent's screen is visible through the eye tracking video if using eye tracking glasses. Brightness on the respondent's phone might need to be adjusted.
- Make sure the calibration works, let the respondent read a couple of lines on another webpage and see if it looks correct.
- Explain as much as possible about the technique for the respondent. What is eye tracking, why do you use it? and what will happen during the session.
- You might want to record sound separately so you don't need to rely upon the eye tracking being active during interviews and such.
- Think about if you will need a secondary screen recording. Problems with quality have occurred but in the end it only became a time consuming element and on top of the eye tracking video it could not generate insights into the respondent behaviour which the eye tracking video could not.
- Start the eye tracking recording when the respondent starts the tasks, it makes it easier to extract a number of fixations from Tobii Pro Lab if you're using that software.

After the test

- During the analysis make sure you know before what you're looking for during the time you spent looking through the eye tracking videos. Are you looking for usability issues, text and content, focus elements or just overall experience?
- Have in mind what parameters can affect the respondents gaze patterns. The scenario/task, the environment, or something else.

Limitations

Many of the issues highlighted in the discussion could be seen as limitations, such as the facility not being as accessible as we have wished for and resulting in a respondent leaving before performing the test, and the complication of recruiting respondents with the same background experience for the two target groups. However, these factors resulted in important lessons which could be carried on to future studies.

In earlier studies respondents have expressed discomfort wearing the eye tracking glasses [3]. Even though none of the respondents in this study expressed discomfort doesn't mean it can be overlooked in future studies.

In the end the biggest limitation is that very few results can be generalised outside of this study. As mentioned earlier in the result, just because a respondent showed a specific Gaze

Type behaviour in this study doesn't mean that that respondent would show the same behaviour in another context, on another webpage or solving another task. The usability issues/insights and the results related to gaze from the tests can be used to optimize SBAB.se, but shall not be taken as a fact and be implemented without criticism outside of this study and this context. But what could be said is that the insights regarding the method can be used as inspiration for future studies, into using eye tracking as a supportive tool for usability tests, but also as a foundation for future studies regarding accessible and inclusive methods in universal design.

CONCLUSION

Answering the problem formulation, is eye tracking a relevant supportive method while executing usability tests with respondents with and without autism? We did find usability issues and insights on how to design better for both of the target groups based on the eye tracking itself but what's really in favour of eye tracking is that it demands less of the respondents' verbal capacity (suitable for respondents with autism). This became noticeable through observing one of the respondents' tests. R2 was the respondent who ended up as a GazeType 1, very focused on details and reading every line of text. It did not come naturally for R2 to verbally communicate during the interview or during the RTA, but solely based upon the eye tracking video, insights into his interaction behaviour could still be noticable.

Even though we couldn't find any significant differences between the two target groups, conclusions can be drawn based upon methodology findings. Eye tracking is a highly relevant method when executing usability tests with respondents who do not feel comfortable or are able to communicate verbally. Being able to rely upon eye tracking solely opens up doors to test with respondents who otherwise might be excluded. But what has to be stated is that just because this study concludes that eye tracking is a suitable method when testing with respondents who are not able or does not feel comfortable with verbal feedback doesn't mean it's a fully inclusive method. This study does not explore testing this method with respondents with other disabilities, for example it would not work with respondents with visual impairments; it doesn't even work very well with people who rely on regular glasses. It is a method inclusive towards some target groups but at the same it excludes other target groups, and what that shows is that methods have to be adapted and cannot without criticism be inherited from one project to another [8], especially when it comes to performing studies with target groups with different disabilities. However, the study generates knowledge about eye tracking as a suitable method for one specific target group and hopefully that can add to the aggregated knowledge about which method is suitable for which target group in the field of universal design.

Another important conclusion is that eye tracking studies need to be evaluated and balanced between input and output. Financially, time wise, how comfortable it is for the respondent and what you actually get out of it. All studies can't rely upon methodology findings; it has to produce insights in the form of usability problems which can be addressed in a product development process—and that needs to be balanced in relation to time, cost and effort (moderator and respondent).

The one major conclusion is that—not only design needs to be accessible, processes and methods when involving users need to be accessible, inclusive and human. Eye tracking might be one of the methods on the road to succeed with that goal.

ACKNOWLEDGEMENTS

I would like to thank SBAB and Conversionista who helped with knowledge and equipment, and my supervisor at Södertörn University, Kai-Mikael Jää-Aro.

REFERENCES

1. Ali-Hasan, N.F., Harrington, E.J. and Richman, J.B. 2008. Best practices for eye tracking of television and video user experiences. Proceeding of the 1st international conference on Designing interactive user experiences for TV and video - uxtv '08 (Silicon Valley, California, USA, 2008), 5.
2. Bohman, P.R. and Anderson, S. 2005. A conceptual framework for accessibility tools to benefit users with cognitive disabilities. Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility (W4A) (New York, NY, USA, May 2005), 85–89.
3. Eger, N., Ball, L., Stevens, R. and Dodd, J. 2007. Cueing retrospective verbal reports in usability testing through eye-movement replay. (Jan. 2007), 129–137.
4. Eisapour, M., Cao, S. and Boger, J. 2018. Game Design for Users with Constraint: Exergame for Older Adults with Cognitive Impairment. The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings (Berlin Germany, Oct. 2018), 128–130.
5. Eisapour, M., Cao, S., Domenicucci, L. and Boger, J. 2018. Participatory Design of a Virtual Reality Exercise for People with Mild Cognitive Impairment. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC Canada, Apr. 2018), 1–9.
6. Elbabour, F., Alhadreti, O. and Mayhew, P. 2017. Eye tracking in retrospective think-aloud usability testing: is there added value? Journal of Usability Studies. 12, 3 (May 2017), 95–110.

7. Elling, S., Lentz, L. and de Jong, M. 2011. Retrospective think-aloud method: using eye movements as an extra cue for participants' verbalizations. Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11 (Vancouver, BC, Canada, 2011), 1161.
8. Eraslan, S., Yaneva, V., Yesilada, Y. and Harper, S. 2019. Web users with autism: eye tracking evidence for differences. Behaviour & Information Technology. 38, 7 (Jul. 2019), 678–700. DOI:https://doi.org/10.1080/0144929X.2018.1551933.
9. Ericsson, K.A. and Simon, H.A. 1984. Protocol analysis: Verbal reports as data. The MIT Press.
10. European Commission. European accessibility act: https://ec.europa.eu/social/main.jsp?catId=1202. Accessed: 2021-05-10.
11. Freeman, B. 2011. Triggered think-aloud protocol: using eye tracking to improve usability test moderation. Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11 (Vancouver, BC, Canada, 2011), 1171.
12. Frith, U. and Happé, F. 2005. Autism spectrum disorder. Current biology: CB. 15, 19 (Oct. 2005), R786-790. DOI:https://doi.org/10.1016/j.cub.2005.09.033.
13. Garzotto, F. and Gonella, R. 2011. Children's co-design and inclusive education. Proceedings of the 10th International Conference on Interaction Design and Children (New York, NY, USA, Jun. 2011), 260–263.
14. Gkatzidou, V., Pearson, E., Green, S. and Perrin, F.-O. 2011. Widgets to support disabled learners: a challenge to participatory inclusive design. Proceedings of the 23rd Australian Computer-Human Interaction Conference (New York, NY, USA, Nov. 2011), 130–139.
15. Guan, Z., Lee, S., Cuddihy, E. and Ramey, J. 2006. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06 (Montreal Quebec, Canada, 2006), 1253.
16. Keates, S., Clarkson, P.J., Harrison, L.-A. and Robinson, P. 2000. Towards a practical inclusive design approach. Proceedings on the 2000 conference on Universal Usability - CUU '00 (Arlington, Virginia, United States, 2000), 45–52.
17. Law, C.M., Yi, J.S., Choi, Y.S. and Jacko, J.A. 2006. Are disability-access guidelines designed for designers?: do they need to be? Proceedings of the 20th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artefacts and environments - OZCHI '06 (Sydney, Australia, 2006), 357.
18. Löwgren, J. 2002. The use qualities of digital designs. (Oct. 2002).
19. Marsh, S. 2018. User Research. A Practical Guide to Designing Better Products and Services.
20. Newell, A.F. and Gregor, P. 2000. "User sensitive inclusive design"--- in search of a new paradigm. Proceedings on the 2000 conference on Universal Usability - CUU '00 (Arlington, Virginia, United States, 2000), 39–44.
21. Ousley, O. and Cermak, T. 2014. Autism Spectrum Disorder: Defining Dimensions and Subgroups. Current Developmental Disorders Reports. 1, (Mar. 2014). DOI:https://doi.org/10.1007/s40474-013-0003-1.
22. Pernice, K. F-Shaped Pattern of Reading on the Web: Misunderstood, But Still Relevant (Even on Mobile): 2017. https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/. Accessed: 2021-04-29.
23. Pernice, K. The Layer-Cake Pattern of Scanning Content on the Web: 2019. https://www.nngroup.com/articles/layer-cake-pattern-scanning/. Accessed: 2021-04-29.
24. Pätsch, G., Mandl, T. and Womser-Hacker, C. 2014. Using sensor graphs to stimulate recall in retrospective think-aloud protocols. Proceedings of the 5th Information Interaction in Context Symposium (New York, NY, USA, Aug. 2014), 303–307.
25. Pichiliani, T.C.P.B. and Pizzolato, E.B. 2019. A survey on the awareness of brazilian web development community about cognitive accessibility. Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems (New York, NY, USA, Oct. 2019), 1–11.
26. Prior, S. 2010. HCI methods for including adults with disabilities in the design of CHAMPION. CHI '10 Extended Abstracts on Human Factors in Computing Systems (New York, NY, USA, Apr. 2010), 2891–2894.
27. Petrie, H. Hamilton, F. King, N. Pavan, P. Remote usability evaluations With disabled people | Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: https://dl.acm.org/doi/10.1145/1124772.1124942. Accessed: 2021-05-03.
28. Robins, B. and Dautenhahn, K. 2004. Interacting with robots: can we encourage social interaction skills in children with autism? ACM SIGACCESS Accessibility and Computing. 80 (Sep. 2004), 6–10. DOI:https://doi.org/10.1145/1055680.1055682.
29. Russell, A.J., Mataix-Cols, D., Anson, M. and Murphy, D.G.M. 2005. Obsessions and compulsions in Asperger syndrome and

high-functioning autism. The British Journal of Psychiatry. 186, 6 (Jun. 2005), 525–528. DOI:https://doi.org/10.1192/bjp.186.6.525.
30. Sari, J.N., Ferdiana, R., Santosa, P.I. and Nugroho, L.E. 2015. An eye tracking study: exploration customer behavior on web design. Proceedings of the International HCI and UX Conference in Indonesia (Bandung Indonesia, Apr. 2015), 69–72.
31. Sevilla, J., Herrera, G., Martínez, B. and Alcantud, F. 2007. Web accessibility for individuals with cognitive deficits: A comparative study between an existing commercial Web and its cognitively accessible equivalent. ACM Transactions on Computer-Human Interaction. 14, 3 (Sep. 2007), 12. DOI:https://doi.org/10.1145/1279700.1279702.
32. Tobii. Tobii Pro Glasses 2 wearable eye tracker: 2015. https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/. Accessed: 2021-05-10.
33. Web Content Accessibility Guidelines (WCAG) Overview: https://www.w3.org/WAI/standards-guidelines/wcag/. Accessed: 2021-05-24.
34. Zander, E. Det här är autism: 2018. https://www.autismforum.se/om-autism/det-har-ar-autism/. Accessed: 2021-05-10.