Tell me: Am I eye candy?

Retrospective Think Aloud method combined with eye tracking devices for usability testing on e-governance websites

- Bachelor Thesis -

Commnunication Science

Matraku, A. (s2099705)

Supervisor: Karreman, J.

University of Twente, The Netherlands

21.07.2023

Introduction: Providing accurate information online allows improved decision-making. This is particularly true for online governmental websites. Data dashboards are used to visualize complex data with the intent to communicate information in an understandable matter for the end-user. Individual usability testing methods can be implemented to gain insight into the effectiveness, efficiency and satisfaction of the user. However, the combination of multiple methods together still has research gaps.

Aim: The aim of this study was to 1) investigate the usability of e-governance website and data dashboard and 2) the effectiveness of combing retrospective think-aloud and eye tracking technology together. For this study the website of Kennispunt Twente was used.

Method: Using a mixed method approach, this study combined retrospective think-aloud protocol and eye-tracking technology with the system usability scale to gain further insight into the usability of e-government websites and data dashboards.

Results: The findings from this study show that the e-government website and data dash have issues related to usability. The usability issues are mostly related to the understanding of information and navigation on the website. These findings are underlined by the below-average rating of the system usability score.

Conclusion: These results underline that e-government website have room for improvement relating to usability of the website. Moreover, the combination of usability testing methods provides valuable and concrete insights about aspects that can be improved.

Keywords: Eye tracking, retrospective think-aloud, usability, system usability scale, data dashboards, e-government

Introduction

Proving information online became the default of any organization. This is true for small retailer and branches of governments. The extent of the online presence varies from a social media channel to online websites aimed to provide extensive data bases of information. Governments channels that are online are called electronic government (e-government). Bataineh et al characterize e-government as governments that leverage information communication technology (ICT) to share government-related information or services government with citizens (2018). E-government websites further aim to engage citizens via ICT, which results in more empowered citizens (Roberts & Vatrapu, 2010). Offering these online services results increased transparency, cost reduction and/or more channels in which citizens can interact with governments. There are different kinds of communication with different stakeholders. Robertson and Vatrapu (2010) summarize the many types of stakeholder relation e-governance can take (see figure 1.1). This research is done in cooperation with Kennispunt Twente, a non-profit research agency located in Enschede, the Netherlands, which provides information to connect the municipality of the region Twente together by leveraging ICT. Which makes it a e-government website. For Kennispunt the Government to Business and Citizen to Government are relevant. Kennispunt offers information that is accessible to everyone. They also have data dashboards only accessible to a public servant of the Netherlands that contains private and sensitive information of the citizen of the municipalities.

There are different ways to communicate large amounts of information. From text-based to visual representations of them. Data dashboards are visual representation of data. Smith (2013) provides a small historical overview of data dashboards. They were used in 1990 because they used less computational memory, which was scarce at that moment in time. In the 2000s they allowed quick access to information and to monitor changes more effectively. This trend continued till today, resulting in business intelligence (BI) tools to emerge, which allows for data dashboard creation. Computational memory is of abundance in today technology, but data dashboard are still used for their advantage to communicate information effectively. Dasgupta and Kapadia (2022) define data dashboards as visual displays that synthesize the most relevant information in such a way that it is understood by the user at a glance. Data dashboards are created by data analysts for an end-user that operate independently from the analyst (Dasgupta &

Kapadia, 2022). Here lies a threat of user-requirement mismatches which results in bad usability of the data dashboards.

Table 1.1

Communication streams of e-governments by Robertson and Vatrapu (2010)

Type of communication stream	Explanation
Government to government	Back office, intra- and intergovernmental exchange, government networks, standards, expertise
Government to citizen	Provision of public information and transparency of information (both passive and active) about government workings and performance, electronic service delivery.
Government to business	Delivery of business services and information, e-procurement (tendering), sales of government-owned business-relevant information
Business to government	Filing of business registration information, taxes, regulatory information, etc.
Citizen to government	Citizen information provision, tax filing, citizen reporting, electronic voting (e-democracy), vehicle licensing

One would assume that organizations and system designers learn from more than 30 years of information about usability. Already in 1993, Nielsen mentioned that there was a plethora of methods to test usability (Sauer et al., 2020). They can be categories in user and non-user testing methods. All of them are still relevant nowadays, but with new inventions came also new methods for testing. Lyzara et al. (2019) created a list of different methods to test the usability with their advantages and disadvantages, which can be found in Appendix A. The author found

that most e-governments use non-user testing methods. In 2003 Abran et al. highlighted that software usability is expected by the end user. However, designing a system that takes the requirements into account was not frequently done and presents a challenge from a usability point of view. Still, in 2016 Sørum presents an example of programmers that design systems without taking the user requirements into account. The author stated that organizations should become aware of users' needs and expectations to identify relevant user requirements. Most authors from the last 30 years agree that user requirements are an important part in the designing process, however evidence suggests that the requirements of the user are not always taken into account. This potential threat lies in all activities in which designers and users are separated from each other. With the aim to improve the e-government website it is crucial to investigate the usability in the first place. Different methods can be used to make sense of usability.

Eye tracking technology

Eye tracking describes the disciple of tracking the focus of visual attention. For that different methods and techniques can be used. One of the most prominent ones is Pupil Center Cornea Reflection (PCCR). In this method infrared light is directed towards the cornea. The position of the cornea relative to the pupil is measured, which can identify the direction of the gaze (Pude et al., 2017; Cater & Luke, 2020). This allows to measure the fixation and movement of the eyes. The PCCR method can be measured in two ways, screen based, and non-screen based. Eye tracking technology is placed underneath or above a monitor. Overall, this method is less obstructive and used to investigate screen related tasks. More commonly known are eye-tracking glasses, which are worn by the participants. The glasses allow for non-screen-based experiments. In sum, both technologies have their place and time. The bar allows for an immersive experience while the glasses allow for non-screen-based experiments.

Think aloud protocols

Eye-tracking data are difficult to interpret without context. A long fixation could be interpreted as difficulty to understand the information or that the participant finds that area interesting (Elbabour et al., 2017). Therefore, eye-tracking data should be complemented with additional data. Lyzara et al. (2019) proposes to capture a wide array of data researcher should implement think-aloud methods. In the fast-evolving world of HCI think-aloud methods are still used to gain insights into user behavior and their interaction between software and websites (Van der

Haal et al., 2003). Jasps et al. (2004) argue that the cognitive process becomes clear based on the verbalization of the tester. In addtion to Van der Haal et al. (2003) the authors elaborate that exploratory design lets usability researchers gain more insight into the user behaviour with software's.

Overall, there are two different kinds of think-aloud methods: concurrent think-aloud (CTA) and retrospective think-aloud (RTA). Both methods have in common that the participant verbalizes their thoughts which might include opinions or feelings (Van den Haal et al., 2003). Further, they allow to create mental models of the participants though process (Berkoff, 2020). The difference between the two methods includes the time of verbalization. Depending on the verbalization, different memory types are used (Jaspers et al., 2004). In the CTA the participant verbalizes their thoughts while interacting with the website and/or performing a task Elling et al., (2012). If the participant is instructed to use the CTA method the working memory is used. This is in contrast to RTA, which instructs the participant to verbalize their though process after the task is completed. This requires the long-term memory of the participant. One limitation of CTA are cognitive blocks (Elling et al., 2012). Therefore, this research focuses on RTA. Individually, think aloud methods and eye-tracking technology have been investigated by usability researchers, however the combination of both methods still is novel. Therefore, this research is aim to answer the research question:

1) To what extent does the integration of eye tracking and retrospective think-aloud enhance the effectiveness of usability testing?

As mentioned before this research is done in cooperation with Kennispunt. One aspect of the cooperation consists of using one of their data dashboards to investigate if usability testing methods can be employed to gain insights about improvements. In exchange, the findings will be shared with Kennispunt. The second part of this research aimed at providing practical improvements for Kennispunt Twente website and data dashboard. Or in other words:

2) To what extent can the integration of eye tracking and retrospective think-aloud techniques enhance the usability of data dashboards on e-government (Kennispunt) websites?

Theoretical framework

Usability

Usability is defined as “the degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” (ISO, 2021). In other words, it investigates the ease of use of a system or product so the user can ultimately achieve a specific task. This research focuses on the system of e-government data dashboards rather than a product, therefore from here on the definition will be narrowed down to systems. Taking the context of e-government into account the definition can be tailored into: “the extent to which a website can be used by citizens to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified e-government service context.” (Verkijika & De Wet, 2018).

To create a full picture of usability, it is of essential importance to investigate the constructs: effectiveness, efficiency and satisfaction of use. Effectiveness is characterized by the level of success of users to achieve a specified task (Abran et al., 2003). Bevan et al. (2015) use the characteristics to formulate effectiveness. In synergy with the previous authors, Bataineh et al. (2018) added that to the user ability to fulfill a task it also takes into account the users success rate of finding information in the system. However, the negative consequences have not been highlighted. In addition to positive consequences, Bevan et al. (2016) state that the absence of negative consequences should be taken into account for the effectiveness of a product. This was adopted by Bataineh et al. (2018) in which error frequency not only contained the amount of error but also the consequences of them and how users dealt with them. Typical effectiveness is investigated by objective measurements task, completion rate and error rate (Sauer & Sonderegger, 2020).

Efficiency refers to the user's capacity to fulfil a specific task quickly (Abran et al., 2003). Bataineh et al. (2018) further adds that the user should be able to achieve the task without enduring frustration. Satisfaction refers to the user's enjoyment of using the system. Bevan et al (2015) formulates it as the users willingness to use the system. This is in contrast to the other authors whos definition overlaps with each other. Since satisfaction is a subjective measure of usability which is dependent on the user (Sauer & Sonderegger, 2020).

These metrics give usability evaluator methods to investigate usability of systems in a structured way. However, other factors also have been proven as relevant for investing in the usability of systems. Bataineh et al., (2018) includes learnability as a factor for usability. Learnability refers to ease of learning and remember the system. The author claims that learnability of a system influences the effectiveness of the system and therefore the error rates. Further, systems with high learnability enable user to address the specific tasks faster (Sagar & Saha, 2017). In sum, usability in this research is defined as: "the extent to which a website can be used by citizens to achieve specified goals with effectiveness, efficiency, learnability, and satisfaction in a specified e-government service context."

The effects of usability are more meta than just usability alone. Kotamraju and Van der Geest (2011) identified that usability is one essential factor for adoption rate of e-government services. It can influence the attitude towards to product and by proxy the organization as well (Wathen and Burkell, 2002 as cited by Huang and Benyoucef, 2014). Huang and Benyoucef (2014) investigated the usability of e-governance websites in a quantitative study. The author found out that usability correlated with trust about the information. Further, usability influences also user experience. To summarize, usability influence on e-governance websites level of trust, user experience and attitude towards government.

Eye-Tracking findings

Eye tracking became more popular by the increase in accessibility. In research context eye tracking has been used to investigate human interaction with technology, such as visual search or software testing (wang et al., 2019; Chandra et al., 2015). It provides a flood of data about visual information. The foundation of eye-tracking research lies in the eye-mind hypothesis, which proposes that visual attention indicates mental attention (Cater & Luke, 2020). Eye-tracking technology introduces a method in which mental processes can be grasp by visual attention (Wang et al., 2018). This method is beneficial for data dashboards studies as it allows to investigate decision-making processes, reading behaviour and information processing (Bera, 2014). There are also limitation to this method. To collect fruitful data the eye tracking technology needs to be calibrated correctly. Despite calibration it cannot be ensured that the data collected will be reliable for experiment with a long duration of time (Chandra et al., 2015). It is

not possible to gain insights of the deviation until after the experiment. To balance the threat of data loss other usability methods should be used.

Think-aloud thinking

The thoughts of someone else are unknown until expressed. Think aloud method are used for usability testing because they allow for user to share their thought and feelings about a product or system (Van den Haak et al., 2010). Retrospective think aloud (RTA) protocols produce insight into cognitive behaviour and cognitive processes of the user after the initial use of the technology (Cho et al., 2019; Young & Kitchin, 2020).

Combining eye-tracking and think aloud

Eye tracking and think-aloud usability testing methods have proven to be insightful to gain information about user behavior, mental processes and identifying usability issues (Cho et al., 2019) . However, combining different methods together has been shown to result in more advanced usability testing (Sørum, 2016). Sørum additionally combined eye-tracking and CTA with a post-study questionnaire. Elling et al. (2012) combined the CTA method with eye tracking. Both authors concluded that it is challenging for participants to verbalize their thoughts while working on tasks. In those situations, the eye-tracking recording was able to provide useful user behaviour insights. This demonstrates that eye tracking in combination with think-aloud methods can result in relevant findings for usability evaluation, however also that CTA results in discomfort in the participants. To prevent cognitive blocks from happening the RTA method is chosen. In the RTA situation participants can pause the recording of their gaze which allows them to verbalize their thoughts to free them of mental constraints. For this study a combination of these three approaches namely: eye tracking, think-aloud protocols and post-study questionnaire; will be done. This methodology offers unique insights about combining the different usability testing methods.

Method

In line to answer the research question, a usability study has been used. The following paragraph will aim to elaborate on the Participants, material, the measurement used and data analysis. Ethical approval was given by the ethics board of the University of Twente.

3.1 Research design

To answer the research questions a mixed-method approach was conducted. The study followed a sequential process, in which first the participant was introduced to the study, then eye-tracking technology was used to capture the gaze of the participants while solving tasks. After each task, the difficulty was assessed. Once all the tasks were solved, the participant was instructed to watch the recording of the gaze. While watching the recording, the participant was instructed to verbalize their thoughts using the retrospective think-aloud method. After collecting the qualitative data, the participant was instructed to fill out a questionnaire consisting of demographic and the system usability score (SUS).

3.2 Procedure

This study consists of three parts: solving tasks on the website, retrospective verbalizing their thought and filling out the surveys.

After entering the room, the participants were briefed about the goal of the study. Further they were informed that they will be audio recorded and their gaze will be recorded. Then they signed a consent form (Appendix C). During the whole process, the participants were encouraged to ask questions. The researcher was present during the complete duration of each experiment.

After signing the consent form the researcher calibrated the eye tracking device. Here a screening occurred to filter out participants that were unable to participate. The participants were instructed to take the role of a policy advisor. They were asked to provide information which can be found on the twente monitor social domain (TMSD) Kennispunt website, which was the starting website. After each tasks, the difficulty of the task was measured. The tasks will be further elaborated on in the task paragraph. The participants were instructed to remain silent while navigating through the website. See Appendix D for full instructions. After concluding the first part of the experiment, all recordings were stopped.

In the third part of the experiment, the participants are instructed to watch the recording of their navigating and verbalizing what they thought in the process using retrospective think-aloud. The recording, which captured the participants gaze, was played back at 0.5 speed, which gave the participant more time to verbalize their thoughts. The participants were always able to see what they saw, via the gaze path. In addition to that they were made aware to pause and continue the recording at any moment to collect their thoughts and express them. If the

participant remained silent for more than five seconds, the researcher reminded the participant to think out aloud. Once the recording ended a last prompt was made to add any further comments until the participant concludes that there is nothing to add any more.

After the verbalizing was concluded, the participants were instructed to fill out the survey using Qualtrics Mx, which consists of demographics (Appendix X) and the system usability scale. The experiment was concluded after the responses were saved, and any further questions were answered. No compensation was provided for participating in the study. The experiment took around 45min to 1 hour to complete.

3.3 Participants

In total 15 participants participated in the study. The age of the participant deviated from 19 to 26 with an age mean of M= 21.67 (SD=2.02). 4 of the participants were female while 11 were male. 14 of the participants identified themselves as Dutch while one identified themselves as British Canadian. The recruitment of the participants took place at the University of Twente via convivence sampling. One prerequisite to participate was the participant needs to be fluent in Dutch.

3.4 Material and measures

To conduct the experiment Tobii equipment and software was used to collect data. The data of the Tobii Pro Fusion- 250Hz was used to conduct the eye-tracking study. It is a screen-based eye tracker. The Tobii Pro Lab (x64) software version 1.217.49450 will be used to generate heatmaps. To make statements about the findings of the eye-tracking technology the reliability and validity has to be investigated. Before each trail the eye-tracking device was calibrated to measure the degree of accuracy. In this study a 9-point calibration process was selected. The calibration processes was considered sufficient once the accuracy was within or below a SD= 0.5. The validity of this study is 65%. This means that in the recording the participant only looked 65% of the duration of the experiment at the monitor. This is considered invalid. To be considered valid it should be at least 80%. However, this can be argued that the task was on a secondary monitor, intentionally placed away from the main monitor. Therefore, the low validity was caused by the placement of the secondary monitor instead of an incorrect calibration of the eye tracking technology. This was also underlined by the interviews in which the majority of the

participants expressed that they looked away to read the instructions again. The experiment was conducted on a monitor with a resolution of 1920 x 1080 pixels. The experiment was performed in a closed office in the Flexperiment, which are research rooms offered by the BMS-Lab. All software and hardware were provided by the BMS-Lab, the social science laboratory of the University of Twente.

3.4.1 SUS and task complexity

In this study different kinds of measure are being combined. The main measurement will be the qualitative measurement of the output of the participants and the results of the eye-tracking measurements. In addition to that, two surveys will also be filled out by each participant after the experiment.

Firstly, the usability is investigated by a scale measuring the usability of the website overall. In the years of usability research, many different scales have been invented, tested and validated. The SUS, Usability Metric for User Experience (UMUX) and the lite version (UMUX-lite) and Computer System Usability Questionnaire (CSUQ) are scales which have been used by researchers from a variety of fields to test a product. Lewis (2018) investigated the similarities between them, and it was concluded that all scales have a high correlation with their individual measurements.

Hence in the context of usability studies, it can be argued that the most prominent scales can be used interchangeably. For this study, the SUS is used. The SUS was introduced by Brook (1996). It is a 10 items (5 positive and 5 negative) post-study questionnaire which aims to measure perceived usability (Appendix B). It is measured in 5 points Likert-Scala reaching from strongly disagree which is equals 1 point to strongly agree which is equals 5 points.

It is important to note that despite the scale ranging from 0 to 100 it is not represented in percentages but in points. In the SUS a score of 68 is considered average and a score above or equal to 80 is considered good (Lewis and Sauro, 2018). A more elaborate grading scale can be found in Appendix B. Further, it is important to acknowledge the complexity of the task at hand. For that, the complexity is measured by a 5-point Likert scale running from very difficult to very easy.

3.4.2 Tasks

One aspect of the cooperation with Kennispunt consists of providing tasks to the researcher which can be used in the study. Kennispunt provided tasks which are familiar in nature to the request of policy advisors. A full list of the tasks can be found in Figure X. The tasks are designed to be solved with the data dashboard provided by Kennispunt. Kennispunt provides information for everyday citizens and policy advisors of different complexity. Overall, the tasks of the policy advisor follow a repeating structure. They are interested in a target group, the municipalities, the timeframe and categories. Therefore, the complexity of the task depends on the number of parameters included. The tasks were always presented in the same order. A task considered complex would be: How many citizens in the age range from 45 to 60 receive health benefits in Almelo from 2020 to 2023?

Pre-test:

To verify that the tasks are similar in nature to tasks that are asked to policy advisors a pre-test was performed at Kennispunt. The researchers at Kennispunt have been asked to verify that the tasks are similar to task from policy advisors. It was concluded that the task are similar, hence the task can be used in the context of this study.

Table 3.1+

Tasks for usability testing

Ondersteuning English: benefits	1.2 How many youth citizens in Tubbergen received benefits in the year 2021? Answer: 10.5% How many citizens in Tubbergen started and stopped their application in 2021? Answer: Started: 2183 Stopped 2828
Aanbieders English: providers	What has been the most popular healthcare provider for adults in Twente from 2017 till 2022?

Achtergronden English: background	Please find how many men at the age of 40 till 64 receive support from the government in. What is the biggest age group that receives support in 2020?
Overall	Please find the email address and telephone number to reach them.

3.5 Data analysis

Both qualitative and quantitative analysis were performed.

Qualitative.

The data of the Tobii Pro Fusion- 250Hz will be used to generate heatmap and gaze plots using the Tobii Pro Lab software version 1.217.49450 (x64). The heatmaps will indicate the most looked spots on each website. Cooke identified that 58% of eye tracking and thinking aloud consist of verbalizing the text that is read aloud(2010, as cited by Elling et al., 2012). Despite that eye tracking still generates insights about gaze tracking, which can be used to gain insight about search behavior and heat maps, which can be used to visualize areas that have been attracted the most attention.

Table 3

Codebook

Nr	Codes	Explanation
1	Understanding Subcode: Positive Negative	Refers to the understanding of information on the website
2	Navigation	Refers to the navigation with the

	Subcode: Interaction with website Positive Negative Structure Positive Negative	website/features of the website
3	Feedback Subcode: Improvement Suggestion	Refers to feedback made by the participant
4	Relevance Positive Negative	Refers to statements about the level of information on the website.
5	Visual Design Subcode: Positive Negative	Refer to statements about the visual design of the website
6	Eye Tracking specific feedback	The information is provided because of the implementation of eye tracking. / The eye tracking information was used to provide more information.

The RTA was transcribed using Amber Script, a transcription software, however, the researcher went through transcription and the audio recording to correct any mistakes that might occur. During this process, all data have been anonymized. This process results in a transcript which is similar to the recording. To analyze the transcribed RTA an inductive coding scheme was used, hence the code has been created based on the transcripts. For this the transcripts have been segmented into relevant parts, which are used to generate themes. Based on the theme's codes were identified and a codebook was generated.

In total 7 codes were used, which consist of two groups of codes. Code group 1 consists of codes 1 though 6 are used to investigate the usability of the TMSD website. Understanding, Navigation and Relevance are used to acquire data that relates to the usability of the website. Whereas visual design is implemented to acquire insights about the visual design and User experience of the website. Additionally, 'Feedback' is used to categorize information about potential improvements stated by the participant. Code group 2 consists of only code 6. Code 6 'Eye tracking specific feedback' is used to acquire information about how the combination of eye tracking and RTA prompts the participant to generate data which can be used for usability studies.

Before coding the whole dataset, the interrater reliability (IRR) was calculated by instructing a second coder to code 10% of the corpus. The 10% of the corpus were selected parts of the transcript which entailed a wide variety of codes. To ensure the validity of the codebook the IRR should have a value of 0.65. Then the IRR was calculated. The IRR in this study was 0.71, which is higher than 0.65, therefore the codebook can be considered validated. The final codebook can be found in Table 3.

Quantitative

To calculate any quantitative data IBM SPSS 27 was used. Before the analysis was done a cleaning of the CSV file was performed. Further some measurements were recoded.

To calculate the SUS this formula was used.

(((The sum of score of Item 1,3,5,7,9) – 5) + (25 – the sum of Items 2,4,6,8,10)) * 2.5.

In addition to the SUS also the time of completion (TOC) was calculated.

Eye tracking

Finically, the eye-tracking results are used to generate heat maps, which help to identify where participants look the most. The eye tracking device is used in combination with RTA to test weather. Combining these two methods generate more information than using each method in isolation.

Results

In this section first, the results form the retrospective think-aloud method will be discussed, after which the results from the quantitative data will be presented, and lastly, the results of the eye-tracking device will be showcased.

4.1 System usability scale

The SUS provides a quantitative metric which is used to identify the usability rating of the website. The Twente monitor social domain (TSMD) from Kennispunt Twente received a below-average rating of 67.66 (SD: 19.46), see in table 4.1. Based on the grading scale from Lewis and Sauro (2018) the Kennispunt website receives a grade of a 'C'. This means that 41 –59 % of the internet website usability is equal to or worse than Kennispunt, see table 4.2.

Table 4.1

System usability score of the TMSD

Participant numbers	SUS score	
	M	*SD*
15	67.66	19.46

Table 4.2

SUS grades distribution

Grading criteria can be found in appendix b

Grade	Frequency	Distribution of grades
a	4	26.7 %
b	6	40 %
c	2	13.3 %

d	0	0 %
f	3	20 %

4.2 Task difficulty and time of completion

The task difficulty was implemented to gather insights about the complexity of the task. Task 1 & 2: The average task difficulty is 3.6 SD= 1.12. Task 3: The average task difficulty is rated at 2.6 with an SD= 0.632. Task 4 & 5: The average task difficulty was 3.4 with an SD of 0.91. Task 6: The average task difficulty was 4.93 with an SD= 0.25. See table 4.3 for the task difficulty. The task difficulty is not in line with the completion rate of each task. Overall, almost all tasks have a completion rate of 70% or more, except for task 2, which has a completion rate of 43%.

This is in line with the remarks of the interview where most people expressed difficulty selecting the correct municipality in the data dashboard. Additionally, the expectation was task 4 and task 6 which had a completion rate of 100%, meaning that every participant was able to answer the question correctly. This result for task 4 is in contrast to the interview in which many participants expressed a negative sentiment towards the time customization option. The results of task 6 were in line with the interview, in which positive remarks about the structure were made. See table 4.4 for an overview of completion rate per tasks. During the experiment, there was a technical issue, which made it unfeasible to conduct the post-study interview with the RTA. And since the answers needed to be verbalized in the RTA the task completion data was deleted for that participant. The participant took 10.30 minutes on average to complete all tasks (SD=4.35), see table 4.5.

Table 4.3

Task difficulty

No of Tasks	Level of Difficulty	
	M	*SD*
Task 1 , 2	3.6	1.12
Task 3	2.60	0.63
Task 4, 5	3.4	0.91
Task 6	4.93	0.258

Table 4.4

Completion rate per tasks

Task number	Task successful	Task failed	Completion rate
1	10	4	71%
2	6	8	42%
3	13	1	93%
4	14	0	100%
5	12	2	85%
6	14	0	100%

Table 4.5

Time of completion given in minutes

Time of completion	
Mean	*SD*
10.30	4.35

4.3 Retrospective think-aloud

In this section the findings of the coded RTA will be presented. An overview of the frequencies of the codes can be found in table 4.6.

Table 4.6

Frequencies of codes

	Overall	Positive	Negative
Understanding	117	53	39
Interaction	62	22	39
Structure	20	13	1
Navigation	95	30	19
Feedback	61	8	8
Suggestion	43		

Relevance	54	14	32
Visual Design	30	14	7
Eye tracking was used	28		

4.3.1 Understanding

The code understanding is divided into positive and negative understanding. The participant mentioned the understanding of the information on the website 117 times. Out of which were 63 positive remarks. Most positive remarks about understanding refer to the titles of the three dashboards. Most positive comments related to inwooners met ondersteuning (English: Citizen with benefits). Most participants expressed that the title correctly summarizes what the dashboard entails. Aanbieders (English: Providers) was also mostly understood, however some participants were unsure about what providers the dashboard is about. P7 expressed that it could also relate to electricity providers, which highlighted that the information is not understood directly. Most problems about understanding information of the website relate to the achtergronden dashboard. One participant referred that they expected background related to ethnicity, while someone else expected information related to the background of Kennispunt. Disregarding the interpretation, it led to misunderstanding of the information. Most participants selected achtergronden last, after exploring the rest of the website. The participants indicated that it was more a process of elimination rather than clear and thoughtful understanding of the information.

General remarks revolved around the clarity of information. P05 said that information was displayed in a user-friendly way. Additionally, there were situations in which participants went on a website that were not helpful to accomplish the task and realized that directly. Therefore, the information on the website, whether helpful to accomplish the task or not, was being understood was such.

Further, are there instances of positive remarks about the display of information that makes it understandable. Here the interactive dashboards are mostly cited. One pattern that emerged was that some titles were not conclusive enough. The age group customization *domain* led to confusion among the participants. It was commented by the participants that they did not

understand what domain meant. Once they interacted with it and the selection for jeudg (English: youth) and WMO (wet maatschappelijke ondersteuning; English: Social Support Act) appeared, most participants understand that the domain referred to age group customization. Participants 13 said:

> *"Um, then I looked around a bit confused because I knew I needed it for adults and I only saw the word youth and the word I didn't really know. So then I looked around. If I could find something about adults somewhere. Um, and then at some point, I assumed that the WMO was probably the opposite of youth. So that had to be adults. So then hesitantly I clicked the WMO."- Participant 13*

This example highlights that it is possible to infer that WMO can be understood as elders, or at least as the opposite of youth. However, only by process of elimination and not due to the clarity of the information. Lastly, was the button 'themas' not understood clearly. It navigated the user back to the website to select one from the three dashboards. That purpose was unclear to the participants. They opted to choose to click on the logo or other methods, instead of thema's to bring them back to the homepage. The results of the code understanding identified specific aspects of the website cannot be understood clearly by the users and therefore need improvement. These are, but not limited to the titles of the dashboard and customization.

4.3.2 Navigation

The code navigation is divided into multiple subcodes; interaction which can be positive or negative, or structure which can be positive or negative. Navigation refers to the interaction with the website and features. Interaction refers to remarks by the participant about their interaction with the features (e.g. adjusting age groups, selecting the municipality) of the website, and the participant can do this correctly or incorrectly, or respectively positive or negative. Structure refers to the remarks about the layout of the website. The participants can make positive or negative statements about that.

4.3.3. Interaction with the website

The participants made 67 remarks about the interaction with the website. Out of these 22 were positive. The Kennispunt Twente provides the data dashboard with a Business intelligence (BI) tool. It provides further options to interact with the website. One of which is insights into the

data once the mouse hovers over certain elements on the website. Many participants were able to interact with that feature. P16 highlighted this interaction with the mouse and data dashboards are intuitive and reminded them of similar websites. Participant 16 said:

> *"Well, when you just with the mouse went over like the, uh, uh, statistics, it popped up immediately. And then it was really easy to see, like, amount of clients. And that's quite clear what it was. And it looked like the period and stuff looked like the web pages that I visited before." - Participant 16*

Another way to interact with the website was to customize and alter specific data to the user needs. Participant 10 highlighted that the data dashboard provides indication that the values can be customized, which made interaction intuitive.

There were in total 39 negative remarks. Despite previous remarks that the ease of use once the method to interact is discovered, many participants had trouble discovering that the data dashboard can be interacted with in the first place. The participants forgot in 5 of 15 cases to alter the period in the first task. Further, the interaction with the BI tool is not commonly understood by all participants. Four participants provided a mathematical walkthrough about how they would solve task 3, which asked for a total number, instead of using the integrated BI tools. This case stresses that the participants were unable to interact with the website.

Overall, it can be that interaction with the website received more negative than positive remarks. This was caused by not becoming aware of the features that the BI tools offer. However, once discovered the participants thought positively about them.

4.3.4. Structure

The participants made 20 remarks about the structure of the website. 13 positive remarks were made. Most structure remarks were positive but repeating. Participants mentioned that the structure of the Kennispunt website follows a similar structure to governmental websites. Additionally, the structure was commented as 'standard', which is meant as a positive remark.

However, participants also mentioned one negative aspect of the structure. Participant 4 suffers from bad eyesight. The text on the dashboards are mostly same size, therefore the website did not offer a visual structure for the participant. This is also the case for the description of the

dashboards which made it more difficult to make sense of the dashboard. P04 was unable to alter the inwooner dashboard correctly as ‘domain’ blurred right into the dashboard. Participant 4:

> *“It's a bit small and it's under the title here, the title of the charts. Putting it above there would make maybe more sense because then you're like maybe a bit bigger, like a um, yeah, bigger and above. The title here would make sense since you're then drawn to it, towards it instead of it being a part of the chart. Um, because I, I always noticed that when I like I see a title, it's like, okay, I can just immediately start looking for the information I need. I don't need to like, like look for drop down menus in the middle of my charts anymore.” - Participant 4*

Overall, the structure of the Kennispunt Twente website received positive remarks. There is no need to adjust the structure of the website. However, it would be advisable to alter the size of certain headings of the website to create a structure which would improve the navigation of the website.

4.3.5. Feedback

The code ‘feedback’ is divided into three categories. Feedback, improvements and suggestions. There have been 93 remarks which entailed either positive, negative or general feedback. The feedback was directed towards specific aspects of the website.

Many participants wanted to have the information, which is provided on the website after selecting the dashboard, on the first website. The three dashboard options also were commented on.

Inwooners met ondersteuning has a dashboard with four displays (see figure 4.1). The display in the top left, top right and bottom right corner are synchronized with each other. Meaning that if one municipality is selected, the same municipality is also highlighted. This is not the case for the bottom left interface. There, a selection has to be made again. This has been noted by the participants multiple times and caused confusion and resulted in errors in task completion.

Achtergronden received feedback for its name. It appears that the participants perceive that there is a mismatch between the name and what they expect. It has been noted that some participants understood it as the background of the organization rather than about the population.

The table at the middle right of the dashboard was drowned by the colorful bar charts (see figure 4.2). After debriefing the participants about the table many expressed their liking for it, as it presents information in a more structured way. However, many participants were not aware of the dashboard.

Further it was noted that the colour coding on this website is inconsistent. The bar chart which visualized the total number of individuals that receive benefits is coded in grey and red to represent women and men. In the line chart at the bottom of the page, which visualizes the indicates naar leveringsvorm in Twente (English: indication by form of delivery in Twente) is colored in red and grey represent persoongebonden budget and zorg in natura respectively, which are two ways benefits are distributed. One participant was not aware of that despite the legend which describes the colour code. Therefore, it can be argued that participants did not become aware of the legend.

Aanbieders also suffered from the understanding issue like achtergronden. As described in understanding the participant with dyslexia it was read as ‘aarbeiders’(English: workers). Furthermore, the task was designed in such a way that the participant needed to take different years into account (see figure 4.3). Many were overwhelmed by the tasks and therefore gave feedback related to the way the dashboard is designed. Many criticized the year customization option, which only provided to select one year. There were no further ways to interact with it.

Figure 4.1

Screenshot of the inwoners met ondersteuning data dashboard

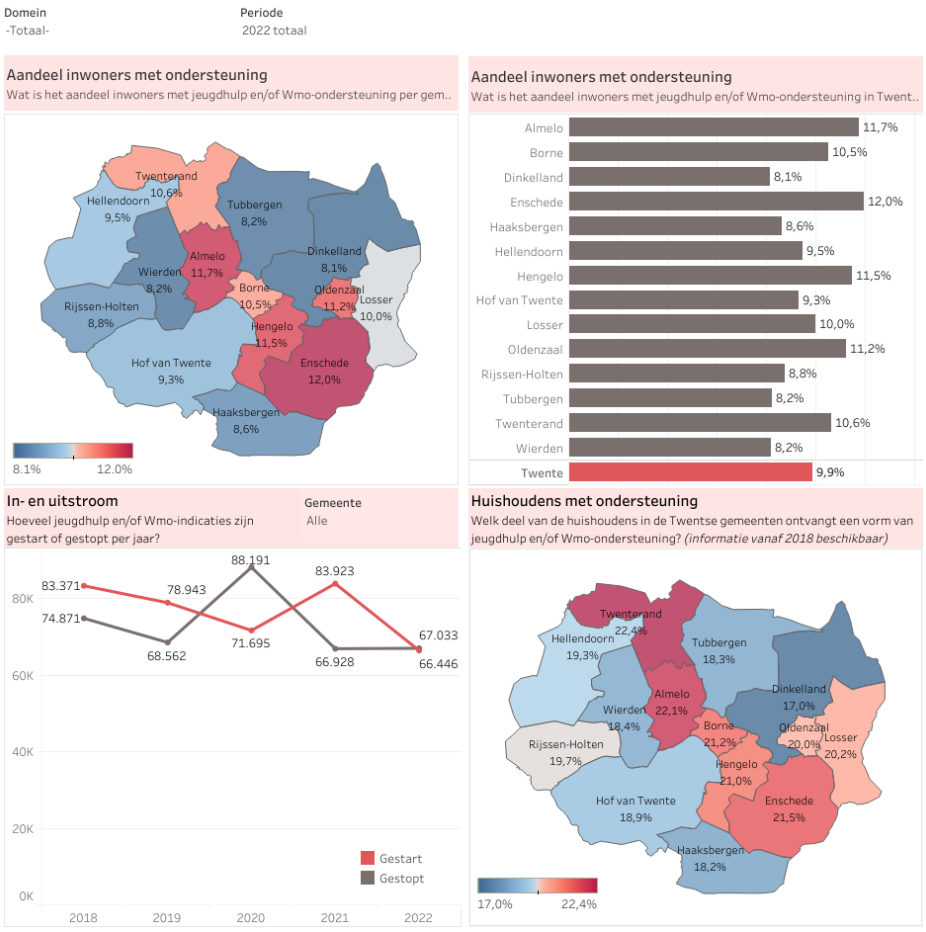


Figure 4.2

Screenshot of the achtergronden data dashboard

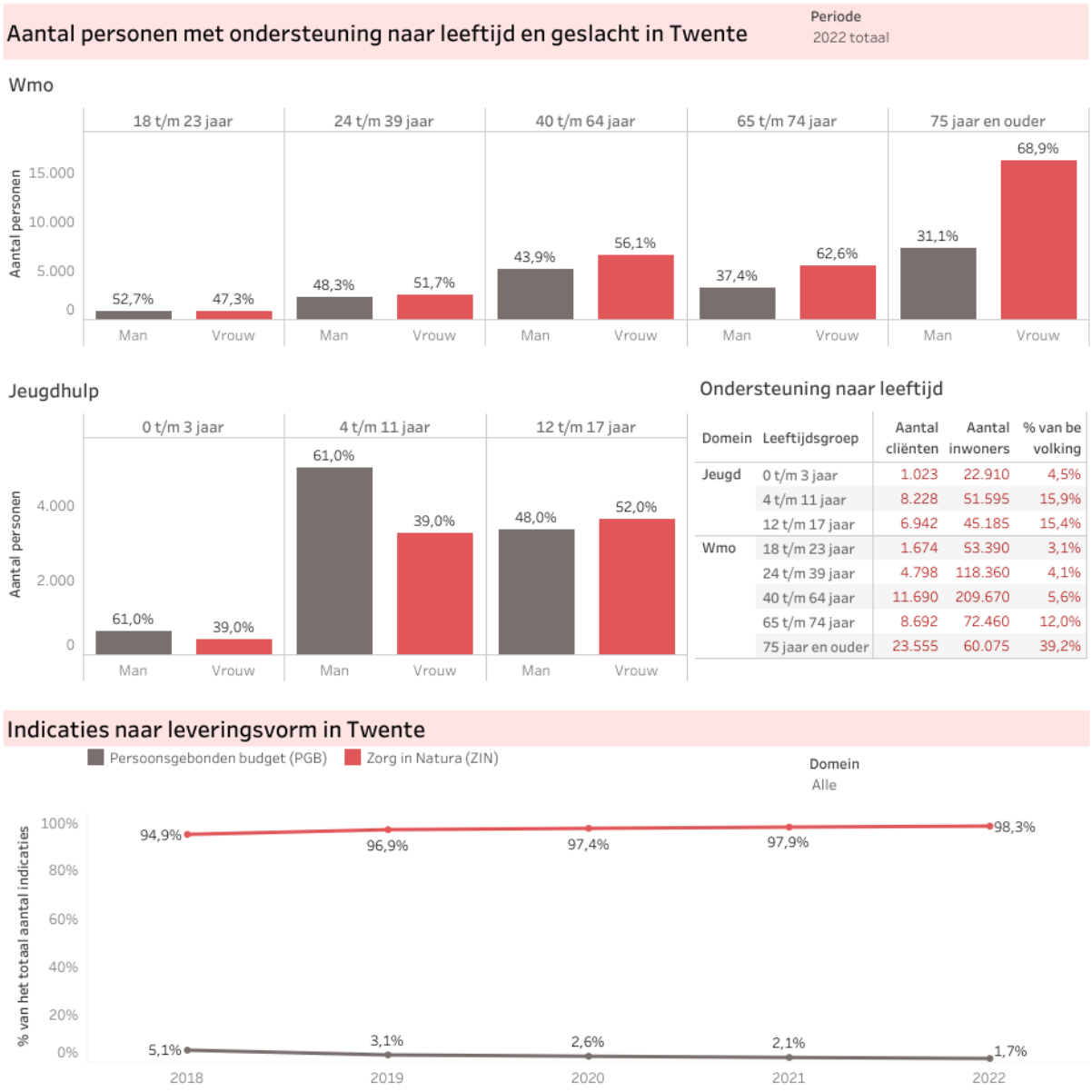
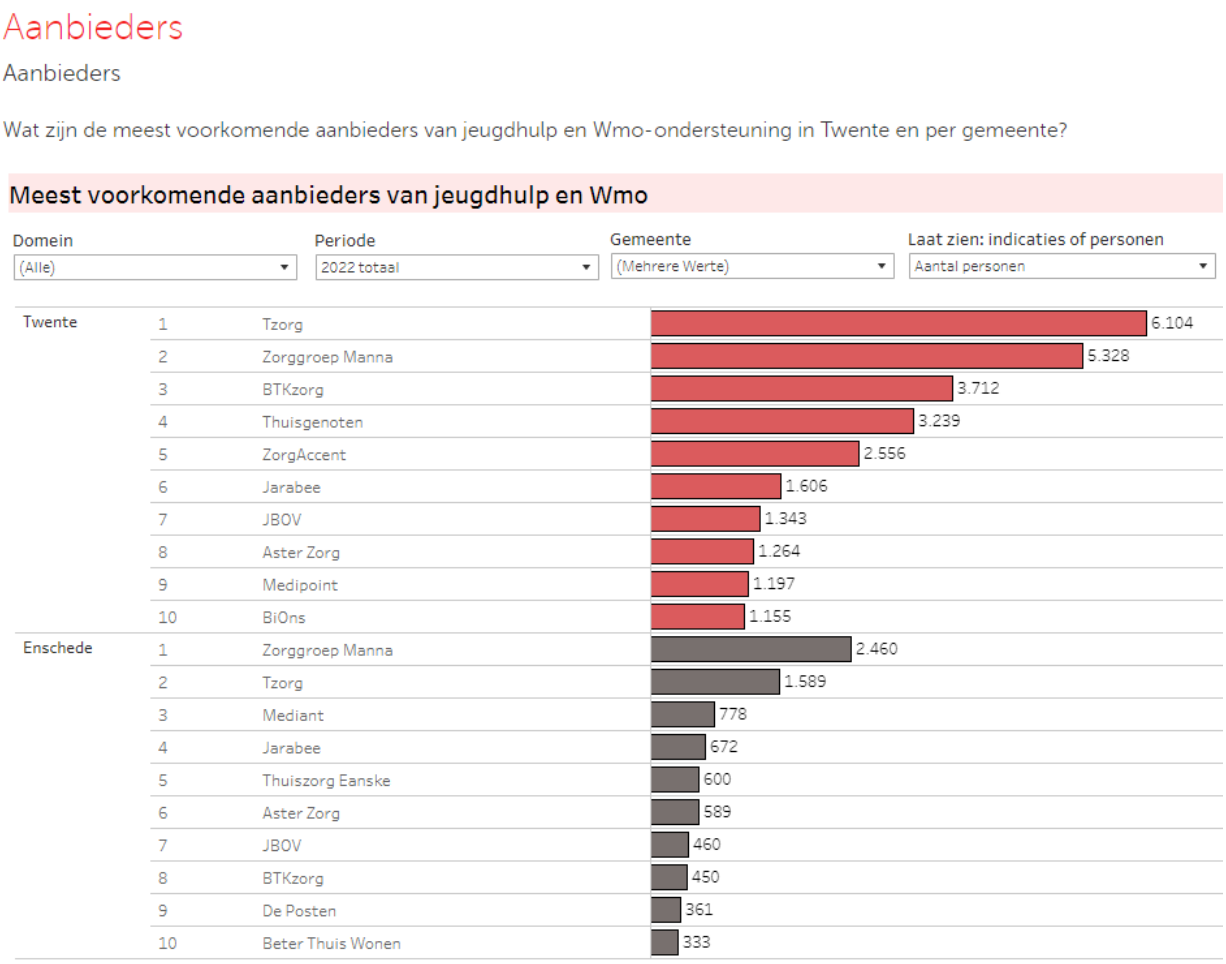


Figure 4.3

Screenshot from aanbieders data dashboard



During the experiment the participants gave suggestions which are aimed at improving the usability of the website. Overall, 47 comments were made that refer to suggestions. Overall, the participants would improve different parts of the dashboards. One participant highlighted that there is no path that indicates navigation process and would like the features as it would help to identify the current location.

There were two ways expressed to address the aanbieders problem about the year customization by the participants. One way was to provide a line chart instead of a bar chart that provides information about individual health care providers over a set of years. The second way was that the dashboard allows a year span and being able to select the year of interest. The information then would be provided in a cumulative chart. To address the colorblindness issue two participants commented to include accessibility features which would made the website accessible to a wider audience.

The different methods to interact with the website need to be discovered by each user on

their own. To guide the users on the website many participants stated that they would like information about what the constatation options are and how to interact with them.

4.3.6 Relevance

Overall, the relevance has been addressed 56 times. The code relevance refers to the level of detail. The level of detail can be either positive or negative. Many comments referred to the themas website and about the level of detail of each dashboard. The dashboards are titled with the dashboard title. The participants had no insights into what kind of information the dashboard provides. This problem was addressed by Kennispunt by introducing a small description before accessing the dashboard to provide information. The description helped the participants to identify the information on the dashboard.

Further are the titles percept differently. Statements about inwooners met ondersteuning are positive, as the title itself was descriptive enough for the participants to anticipate what the dashboard was about. This was not the case for achtergrond. All remarks about achtergronden were negative. In the process of solving the task, which is answerable to that dashboard, participants inspected the achtergrond dashboard last as they were unsure what the dashboard is about. As mentioned in understading the user used a process of elimniation. Therefore, it can be argued that the level of detail is not precise enough.

The dashboards offer a high level of adjustability. Many participants positively noted the customization options, however, there is no information about what these options are and how they can be used to manipulate the data. The titles are not conclusive enough. Many stated that the domain is not fitting and WMO, as stated in understanding, was not understood by the participants. There are two ways to manipulate data. There are dropdown menus which allow for selection, and once in the dashboard itself, in which highlighted values can be in or excluded. The in- and exclusion menus are not understood by the user and therefore also never used, which indicates lack of detail and awareness.

4.3.7. Visual Design

The visual design has been addressed 20 times It refers to statements about the visual design of the Kennispunt website. Participant 9 mentioned the Kennispunt Twente website shares a similar visual design to other Dutch Government websites. Furthermore, there were 7 instances in which participants mentioned that the website looks esthetically pleasing. These comments refer to

different parts of the website. Some highlighted the BI data visualization, while others stated the colour scheme of the website. Participant 3 highlighted that the red and grey colour coding is in line with the colour scheme of the region Twente, which also is red and grey. Furthermore, it was positively noticed that the colour scheme was used to differentiate genders from each other.

However, participants also noticed negative things related to the visual design of the website. One singled-out case was P04 who had bad eyesight. They commented that the features of the websites are small and therefore have decreased readability. The colour scheme, despite being in line with Twente, received negative comments. One participant was colorblind therefore colored data visualizations were unidentifiable for them. Furthermore, it was noted that the red and grey were hard to read as they were not fully saturated. Readability was also decreased on the data dashboards as they do not offer grids in the charts, which made it more difficult to read the data correctly.

4.4 Eye tracking

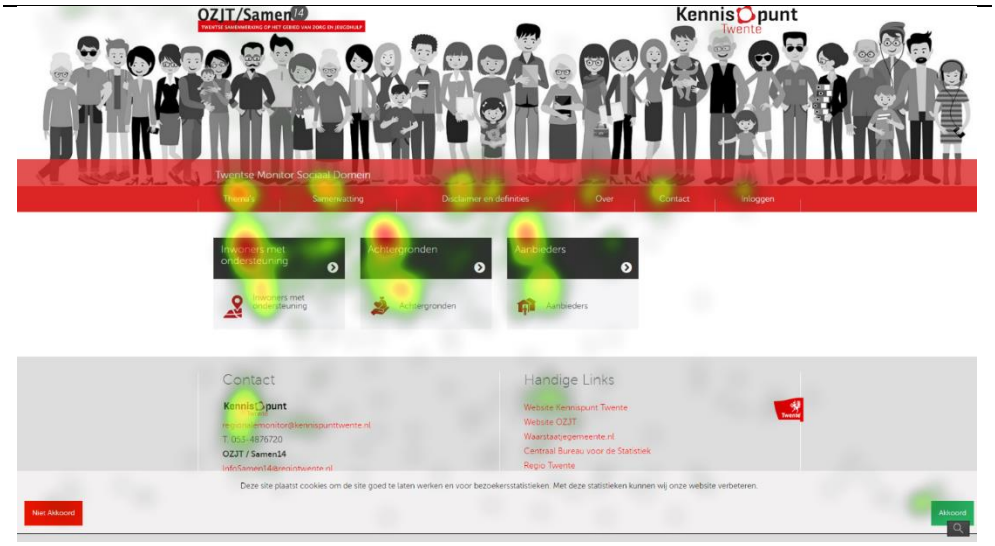
4.4.1 Home page

With the help of the eye tracking technology, it is possible to argue that most participants were able to identify the different methods to access the dashboard (see Figure 4.4). Moreover, did the participants recognize the menu bar, however based on the retrospective think-aloud it became apparent that the participants did not understand what to except from two buttons.

The heatmaps of the website that provides information about the dashboard looked similar to each other. All three textboxes were anchor points, as they were the most looked at (Appendix E). Despite that information, it should be noted that the text was not fully read by all participants. The achtergronden textbox had hotspots that only reached to the third line.

Figure 4.4

Heatmap of the homepage

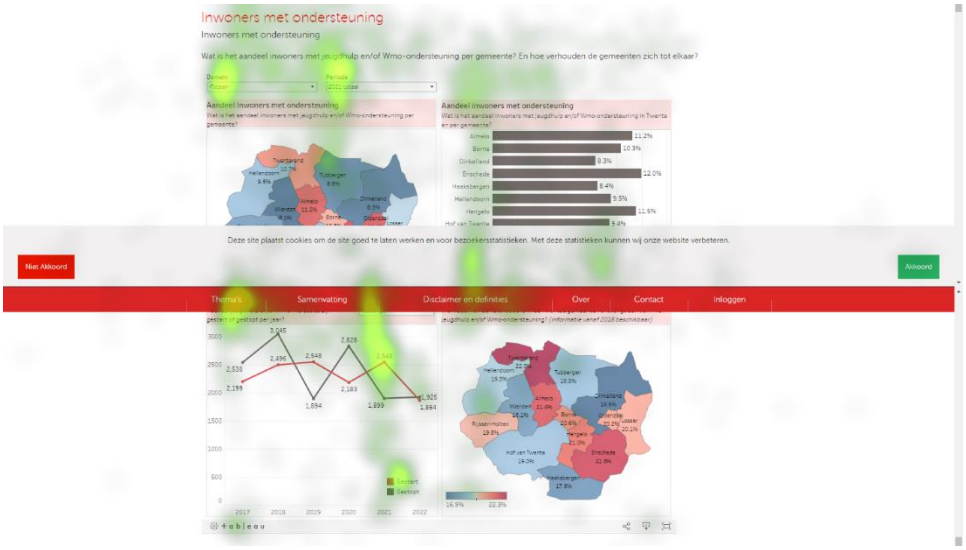


4.4.2. Inwooners met ondersteuning

In contrast to the interviews the customization elements of the dashboard were the most looked at (see Figure 4.5). In the bottom left corner of the BI-Tool, one can make out an anomaly. The dropdown menu to select the municipality folded itself out there. The anomaly can be explained as all participants were required to adjust the municipality and therefore took a certain amount of time to inspect the drop-down menu. Further there was little data generated on the bottom right corner of the dashboard as it communicated no information required to solve a task. However, there is a reading pattern on the title, therefore participants did acknowledge it.

Figure 4.5

Heatmap of the inwonner met ondersteuning website



4.4.3. Achtergronden

The achtergronden dashboard had hotspots at the customization option (see Figure 4.6). Furthermore, can it be noted that the participants recognize the table on the right side of the dashboard. However, in the interview it became apparent that the participants were not aware of the dashboard. After further investigation many participants stated that the table was perceived positive. Some participants even stated that they preferred the table compared to the bar chart. They argued that comparing values with each other became easier as the values were placed in closer proximity to each other.

Figure 4.6

Heatmap of the achergronden wesbite

4.4.4 Aanbieders

The heatmap of the aanbieders dashboard had an anchor point at the year customization option (see Figure 4.7). This can be explained by the fact that the participants needed to adjust the year 5 times to complete the tasks successfully. Furthermore, were the other customization options noted by the participants. It is noteworthy to mention that the default setting of the aanbieders dashboard consists of Twente and Enschede selected. All participants deselected Enschede, without being instructed to do so.

Figure 4.7

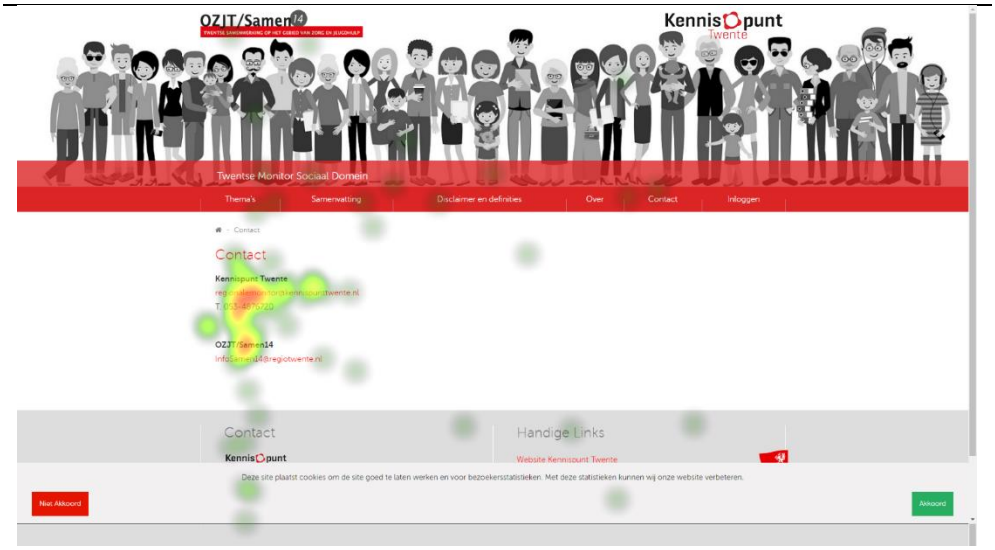
Heatmap of the aanbieders website

4.4.5 Contact page

To complete task 6 the participant could opt to visit the contact page or to scroll down to the bottom of the home page at which the information can also be found (see Figure 4.8). The contact page contains information on Kennispunt's email and telephone number. Anchor points were the email and telephone number. Most remarks were positive and dealt with that the website contains all the information that the user would expect on the page.

Figure 4.8

Heatmap of the contact page



4.4.6. Disclaimers, Over and Samenvatting

The disclaimers and definitions, about and summary pages, were text reliant websites that include different information which is aimed to elaborate on terms which are used on the website. Many participants visited the websites intending to find answers to the questions (see Figure 4.9; see Figure 4.10)). It is noteworthy that the heatmaps generate reading patterns that represent F-patterns. F-patterns emerge when the user reads the text fast and skims for information (Shrestha et al., 2007). In the interview, it was noted these pages contain too much information for it to be understood immediately.

Figure 4.9

Heatmap of the disclaimers page

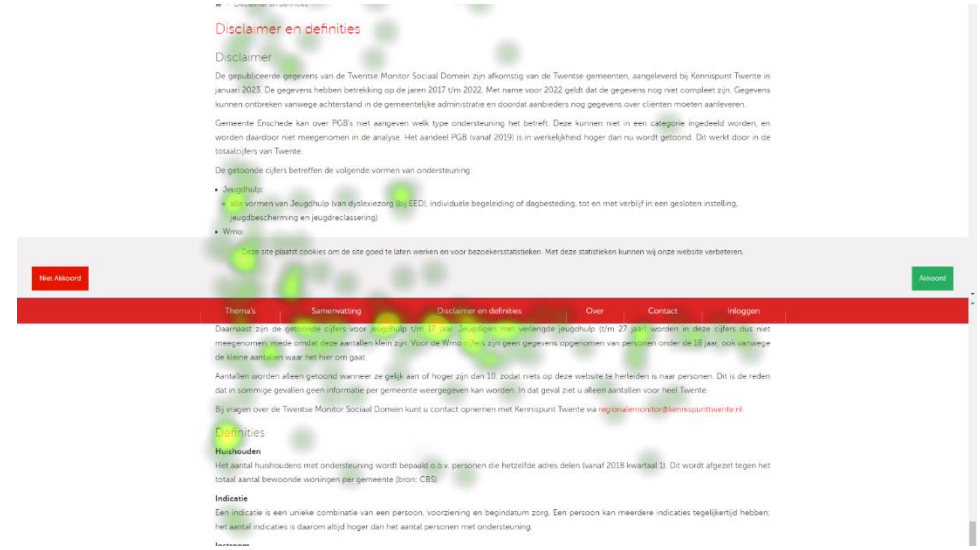


Figure 4.10

Heatmaps of the samenvatting page

All remaining heatmap of the website with the cumulative heat map can be appendix E.

Discussion

This study is aimed to answer two research questions. “To what extent does the integration of eye tracking and retrospective think-aloud enhance the effectiveness of usability testing?”

This study evaluated the usability testing method of combining retrospective think-aloud (RTA) and eye-tracking technology together by investigating the Twente monitor social domain (TMSD) website that used data dashboards of Kennispunt Twente an e-government website that is primarily by policy advisors.

The results from the system usability score (SUS) suggested that the TMSD dashboard from Kennispunt has below-average usability from Dutch university students. Imperative to mention is that the SUS only provides a value of usability (Pradini et al., 2019). A task-based user testing method was implemented. Eye-tracking technology in combination with the RTA was used to gather insights about the usability of the TMSD.

The second aim of this study was to investigate the integration of eye tracking and RTA in usability testing to identify usability issues. The interaction of eye tracking and RTA resulted in transcripts which helped to identify different issues throughout the website. In support of Cho et al., (2019) this study highlights that eye tracking and RTA can result in rich data from users. Overall, can it be concluded that the combination of RTA and eye tracking results in an accurate assessment of the usability of the website**.** The coding process resulted in an equal distribution of positive and negative remarks. This can be interpreted that the participant could not reach agreement about the usability of the website. It can be argued that the usability of the website can be classified as average and that it can be improved upon. These findings are underlined by the results of the SUS which rated Kennispunt Twente below average. This leads to the conclusion, that the findings of the usability testing methods produce similar results. The combination further produced results that gave fruitful and concrete insights about usability issues that the Kennispunt Twente website suffers from. Furthermore, are these findings recommendation for human centered design process.

Lastly the finding from this research will be explained with the created theoretical framework of usability. In the context of this research usability is defined as: “the extent to which a website can be used by citizens to achieve specified goals with effectiveness, efficiency, learnability, and satisfaction in a specified e-government service context.”.

The effectiveness of the Kennispunt Twente website relates to the error rate of the tasks. Overall, the tasks have a completion rate of 70%. Hence, one can assume that overall, the usability of the website is satisfactory. However, tasks 2 is an outliner with a completion rate of 43%. To successful accomplish task 2 the inwooner met ondersteuning data dashboard was needed.

Participants were not able to interact with the customization of data dashboard. Therefore, the aanbieders data dashboard has low effectiveness. This tasks also required the most customization of data from all the tasks. Task in which the data dashboard does not need to be adjusted have higher success rates.

Efficiency related to the time spend on solving the tasks. There is a high deviation of time to complete (TOC) the tasks in the sample. Therefore, from an efficiency point of view it can be concluded that the website suffers from usability issues which would explain the TOC deviation. Tasks difficulty also relates to TOC, as more complex task requires more time, however the task difficulty rating are do not lead to that conclusion.

Limitations and future recommendations

The answers to the tasks were verbalized in the post-study interview. Due to a technology issue, the recording of one participant was not watchable, it was not feasible to conduct the interview and therefore also not possible to gather insights if the participant achieved the task. To prevent such an issue from occurring again, future research should invite the participant to write down the answer to the task. This would still allow data collection despite technological issues and would not allow the participant to change answers. In this study, one participant expressed that did not remember one answer to one question. Writing down the answers serves as a memory aid for the participant.

The website that has been investigated was in Dutch, however, the verbalization was done in English, as the researcher is not fluent in Dutch. Switching between languages is not the golden standard in usability research. For future research, the researcher should make sure that the website and the verbalizing should be in the same language. Furthermore, there is a target group mismatched in this study. The participants in this study were Dutch university students, however, the main users of the Kennispunt dashboards are Dutch public servants. It would have been more fruitful to conduct this study with public servants instead. These are the more important stakeholders for Kennispunt. A study with them would be fruitful for Kennispunt.

In the RTA it would be advisable to record the screen. This way interaction with the participant is more communicative and can generate more data. In the current research design, the researcher has to synchronize their notes with the transcripts, which might lead to data loss. Furthermore, were there many instances in the transcript in which the participant verbalized

‘here’ or ‘this’ to communicate something about the dashboard. However, with the amount of data collected, it became more complex to correctly link those statements with the corresponding moment of the video recording, as the participant also was able to jump back and forth in the recording. By recoding the screen while doing the RTA these downfalls could be avoided, while also generating more data that is useful to analyse the website.

To gain most of the interviews it would be beneficial to create a prompt script. In the current experiment design, there was no prompt script. The script should result in more information about the functionality of RTA and eye-tracking devices, which would aim to answer the second research question. As mentioned before is the website only available in Dutch. The researcher is not fluent in Dutch. This created the need to use translation software for back-and-forth translation. Here lies potential issues in which words might be translated incorrectly.

Kennispunt Twente is just one of many different e-government that use data dashboards to visualize complex data. To be able to generalize findings to other e-government websites and data dashboards further research into other e-government websites is needed.

Future research should investigate other combinations of usability testing methods. The advantages and disadvantages of RTA, eye-tracking and SUS balanced each other out. In the field of usability testing other combinations of methods can generate valuable insights.

This study focused on usability. However, user experience is an important factor that influences adoption of data dashboards. Future research should focus on user experience of e-government website and data dashboard.

Implications

The implementation of the RTA and eye tracking combination was, from a usability testing point of view, successful. This study design yielded insightful data about the navigation on the website, the shortcomings and strong suits thereof. These methods however came with the disadvantage that it is time intense. Further, prices for eye-tracking technology are only decreasing with time, however, licensing the analysis software is costly. To capitalize the most out of this usability testing method it is advisable to implement it in the later stages of the design process. As mentioned in the introduction it is advantageous to keep user requirements into account and have a human in the center of the design process. Once that is included, it can be argued that is more effective, speaking about time and money, to implement shorter and cheaper

methods of usability testing. Lyzara et al. (2018) highlighted that most methods for e-governance websites are non-user testing methods like automated testing and heuristic evaluation, due to their time effectiveness, low cost and ease of evaluation. The six-dimension framework was tailored for heuristics evaluation for e-government websites (Verkijika & De Wet, 2018). However, once user testing methods are implemented Nielsen (1994) claims that 6 to 7 participants would be enough to identify 75% of usability issues on the website.

Conclusion

To provide information in an effective, efficient and satisfying way on e-government websites is challenging. E-government website provide data dashboards to communicate complex data with the aim to inform citizen. The combination of different usability testing methods such as qualitive eye-tracking technology and retrospective think-aloud method with quantitative system usability scale have the potential to generate user-based and user-centered recommendations.

In this case the combination of different usability testing methods was able to identify usability issues related to understanding of information and navigation on the website. Further were remarks by the participants valuable to generate recommendation to create additional features of on the website. Future research into the usability of e-government website and data dashboards should focus on the implementing user-centered design recommendations and evaluation them by a combined method approach.

Reference list

Bataineh, E., Mourad, B. M. A., & Kammoun, F. (2017). Usability analysis on Dubai e-government portal using eye tracking methodology. In *2017 Computing Conference (pp. 591-600)*. https://doi.org/10.1109/sai.2017.8252156

Bera, P. (2014). Do Distracting Dashboards Matter? Evidence from an Eye Tracking Study. In *Springer eBooks* (pp. 65–74). https://doi.org/10.1007/978-3-319-11373-9_6

Bevan, N., Carter, J., Earthy, J., Geis, T., & Harker, S. (2016). New ISO standards for usability, usability reports and usability measures. In *Lecture Notes in Computer Science* (pp. 268–278). https://doi.org/10.1007/978-3-319-39510-4_25

Bevan, N., Carter, J. M., & Harker, S. (2015). ISO 9241-11 revised: What have we learnt about usability since 1998? In *Lecture Notes in Computer Science* (pp. 143–151). https://doi.org/10.1007/978-3-319-20901-2_13

Brooke, J. H. (1996). SUS: a "Quick and Dirty" usability scale. In *CRC Press eBooks* (pp. 207–212). https://doi.org/10.1201/9781498710411-35

Carter, B. E., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Chandra, S., Sharma, G., Malhotra, S., Jha, D., & Mittal, A. (2015). *Eye tracking based human computer interaction: Applications and their uses*. *International Conference on Man and Machine Interfacing (MAMI)* (pp. 1-5). IEEE. https://doi.org/10.1109/mami.2015.7456615

Dasgupta, N., & Kapadia, F. (2022). The future of the Public Health Data Dashboard. *American Journal of Public Health*, *112*(6), 886–888. https://doi.org/10.2105/ajph.2022.306871

Donker-Kuijer, M. W., De Jong, M. D., & Lentz, L. (2010). Usable guidelines for usable websites? An analysis of five e-government heuristics. *Government Information Quarterly*, *27*(3), 254–263. https://doi.org/10.1016/j.giq.2010.02.006

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through Eye-Movement replay. In *BCS Learning & Development*. British Computer Society. https://doi.org/10.14236/ewic/hci2007.13

Elling, S., Lentz, L., & De Jong, M. D. (2011). *Retrospective think-aloud method*. https://doi.org/10.1145/1978942.1979116

Elling, S., Lentz, L., & De Jong, M. D. (2012). Combining concurrent Think-Aloud protocols and Eye-Tracking observations: an analysis of verbalizations and silences. *IEEE Transactions on Professional Communication*, *55*(3), 206–220. https://doi.org/10.1109/tpc.2012.2206190

Eye tracking in retrospective think-aloud usability testing: Is there added value? - UEA Digital Repository. (n.d.). https://ueaeprints.uea.ac.uk/id/eprint/64991

García, M., & Cano, S. (2022). Eye tracking to evaluate the user eXperience (UX): literature review. In *Springer eBooks* (pp. 134–145). https://doi.org/10.1007/978-3-031-05061-9_10

Henriksson, A. E., Yi, Y., Frost, B., & Middleton, M. R. (2007). Evaluation instrument for e-government websites. *Electronic Government, an International Journal*, *4*(2), 204. https://doi.org/10.1504/eg.2007.013984

Huang, Z. H., & Benyoucef, M. (2014). Usability and credibility of e-government websites. *Government Information Quarterly*, *31*(4), 584–595. https://doi.org/10.1016/j.giq.2014.07.002

Jaspers, M. W. M., Steen, T., Van Den Bos, C., & Geenen, M. M. (2004). The think aloud method: a guide to user interface design. *International Journal of Medical Informatics*, *73*(11–12), 781–795. https://doi.org/10.1016/j.ijmedinf.2004.08.003

Kotamraju, N. P., & Van Der Geest, T. (2012). The tension between user-centred design and e-government services. *Behaviour & Information Technology*, *31*(3), 261–273. https://doi.org/10.1080/0144929x.2011.563797

Lewis, J. D. (2018a). Measuring perceived usability: the CSUQ, SUS, and UMUX. *International Journal of Human-computer Interaction*, *34*(12), 1148–1156. https://doi.org/10.1080/10447318.2017.1418805

Lewis, J. D. (2018b). The system usability scale: past, present, and future. *International Journal of Human-computer Interaction*, *34*(7), 577–590. https://doi.org/10.1080/10447318.2018.1455307

Lewis, J. D., & Sauro, J. (2018). Item benchmarks for the system usability scale. *Journal of Usability Studies Archive*, *13*(3), 158–167. https://doi.org/10.5555/3294033.3294037

Lyzara, R., Purwandari, B., Zulfikar, M. A., Santoso, H. B., & Solichah, I. (2019). *E-Government usability evaluation*. https://doi.org/10.1145/3305160.3305178

Nielsen, J. (1994). Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-computer Studies*, *41*(3), 385–397. https://doi.org/10.1006/ijhc.1994.1065

Pradini, R. S., Kriswibowo, R., & Ramdani, F. (2019). *Usability evaluation on the SIPR website uses the system usability scale and net Promoter score*. https://doi.org/10.1109/siet48054.2019.8986098

Punde, P. A., Jadhav, M. E., & Manza, R. R. (2017). *A study of eye tracking technology and its applications*. https://doi.org/10.1109/icisim.2017.8122153

Robertson, S., & Vatrapu, R. (2010). Digital government. *Annual Review of Information Science and Technology*, *44*(1), 317–364. https://doi.org/10.1002/aris.2010.1440440115

Sagar, K., & Saha, A. (2017). A systematic review of software usability studies. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-017-0048-1

Sauer, J., & Sonderegger, A. (2022). Visual aesthetics and user experience: A multiple-session experiment. *International Journal of Human-Computer Studies*, *165*, 102837. https://doi.org/10.1016/j.ijhcs.2022.102837

Sauer, J., Sonderegger, A., & Schmutz, S. (2020). Usability, user experience and accessibility: towards an integrative model. *Ergonomics*, *63*(10), 1207–1220. https://doi.org/10.1080/00140139.2020.1774080

Shrestha, S., Lenz, K., Chaparro, B. S., & Owens, J. M. (2007). "F" Pattern Scanning of Text and Images in Web Pages. *Proceedings of the Human Factors and Ergonomics Society . . . Annual Meeting*, *51*(18), 1200–1204. https://doi.org/10.1177/154193120705101831

Smith, V. (2013). Data Dashboard as evaluation and research communication tool. *New Directions for Evaluation*, *2013*(140), 21–45. https://doi.org/10.1002/ev.20072

Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, *41*(3), 403–410. https://doi.org/10.1016/j.apergo.2009.09.002

Sørum, H. (2016). Design of Public Sector Websites: Findings from an Eye Tracking Study Emphasizing Visual Attention and Usability Metrics. In *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-319-44159-7_12

Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Van Den Haak, M., Junger, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. https://doi.org/10.1080/0044929031000

Van Nes, F., Abma, T. A., Jonsson, H., & Deeg, D. J. H. (2010). Language differences in qualitative research: is meaning lost in translation? *European Journal of Ageing*, *7*(4), 313–316. https://doi.org/10.1007/s10433-010-0168-y

Verkijika, S. F., & De Wet, L. (2018). A usability assessment of e-government websites in Sub-Saharan Africa. *International Journal of Information Management*, *39*, 20–29. https://doi.org/10.1016/j.ijinfomgt.2017.11.003

Wang, J., Antonenko, P. D., Celepkolu, M., Jimenez, Y., Fieldman, E., & Fieldman, A. (2018). Exploring relationships between eye tracking and traditional usability testing data. *International Journal of Human-computer Interaction*, *35*(6), 483–494. https://doi.org/10.1080/10447318.2018.1464776

Appendix A

Advantages and Challenges of e-Government Usability Evaluation

Method	Advantages	Challenges
Automated testing	- Cost-efficient - Automatic calculation	- Non-actual user-based testing - Too focus on accessibility - Limited insight
Performance Measurement	- High precision - Understand users's cognitive	- Difficult to get potential users to participate
SUS (System Usability Score)	- Quick - Simple - Need small respondents	- Limited insight
Think aloud	- Capturing wide range of cognitive	- Time-consuming
Heuristic evaluation	- Easy - Cost-Efficient - Quick - Usefull for early stage • Can be performed by	- Non-actual user-based testing

	a single inspector - Not only experts, but alco novices can particpate	
Focus group	- Deeper insight	- Time consuming
Interview	- Usefull for early stage implementation - Small group or individual - Encourgae capturing respondents thought	- Time consuming - Need a representative respondents
Questionnaire	- Actual user-based	- High cost - Time consuming - Need a representative respondents
User Feedback	- Actual user-based	- High cost - Time consuming
Field Observation	- Enable valuable information of the real work	- High cost - Time consuming

and social

ascpects

Appendix B

The system usability scale (SUS) questionnaire and grading scale

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Figure 7.2

Grading scale of SUS from Lewis and Sauro, 2018

Table 1. Curved Grading Scale for the SUS

Grade	SUS	Percentile range
A+	84.1 - 100	96 - 100
A	80.8 - 84.0	90 - 95
A-	78.9 - 80.7	85 - 89
B+	77.2 - 78.8	80 - 84
B	74.1 - 77.1	70 - 79
B-	72.6 - 74.0	65 - 69
C+	71.1 - 72.5	60 - 64
C	65.0 - 71.0	41 - 59
C-	62.7 - 64.9	35 - 40
D	51.7 - 62.6	15 - 34
F	0 - 51.6	0 - 14

Appendix C

Consent form

Dear participant,

Thank you for participating in this usability study.
This study is part of a bachelor thesis done in cooperation with Kennispunt Twente, the University of Twente, and the faculty of behavioural, management and social sciences (BMS).

In this study, you will be asked to navigate through the website of Kennispunt Twente and solve tasks given to you by the researcher. While doing so you are asked to equip an eye-tracking device.

The eye-tracking device will record your gaze. Afterwards, you will rewatch the recording of your gaze and are being asked to verbalize your thought process. This session will be audio recorded. All recordings will be stored in a cloud which is secured with multifactor authentication. In the process of transcribing the interview all information that can identify like name or residency, will be anonymized. All records will be deleted after the 21 of July 2023.

The anonymized transcribing will be archived by the University of Twente so they can be used for future research and learning.

Please be aware that you can withdraw from this study at any point without giving any reason. You will receive no repercussions for withdrawing from the study. Your participation is completely voluntary.

If you have any further questions or inquiries, you are welcome to contact the researcher Aldo Matraku via a.matraku@student.utwente.nl or the supervisor Joyce Karreman via j.karreman@utwente.nl.
This study has been approved by the ethics committee of the University of Twente. You can contact the ethics committee at ethicscommittee-bms@utwente.nl.

Appendix D

Instruction used in the study

Instruction:

Kennispunt Twente is a non-profit government-related research agency. It has been accommodated as a guest at Regio Twente, as a recognizable independent unit with its own identity, management and budget. As a member of the national Association for Statistics and Research (VSO), we are committed to the national code of conduct for research agencies. This code of conduct has been laid down by the Dutch Data Protection Authority.
Kennispunt Twente offers policy research and information and data provision for the entire region of Twente. With this, Kennispunt Twente wants to strengthen the knowledge position of the Twente municipalities and the Twente region. We have a lot of data for Twente municipalities. For example, in the areas of population, social domain, housing, employment and safety.

Instruction
In the first part of this experiment, you will navigate on the Kennispunt website. The website is in Dutch. Please remain silent while solving the task.
After each completion you will be asked to fill out a small survey. Then the next task will be given to you.

After completing all tasks, you will be asked to watch a recording of your navigation. In that recording your eye gaze will be shown. Hence, you know where you look at. Please share your thoughts out loud. You are welcome to pause the recording.

Remember: The website is being tested, not you!

Please take the role of policy advisor. You are being asked to gather information which is aimed at creating, reflecting and adjusting policies in the region of Twente. For this you should use the Kennispunt website.

Appendix E

Heatmaps of the Kennispunt Twente – Twente monitor social domain website

Figure 7.1

Heatmap aanbieders description

Figure 7.2

Heatmap achtergronden description

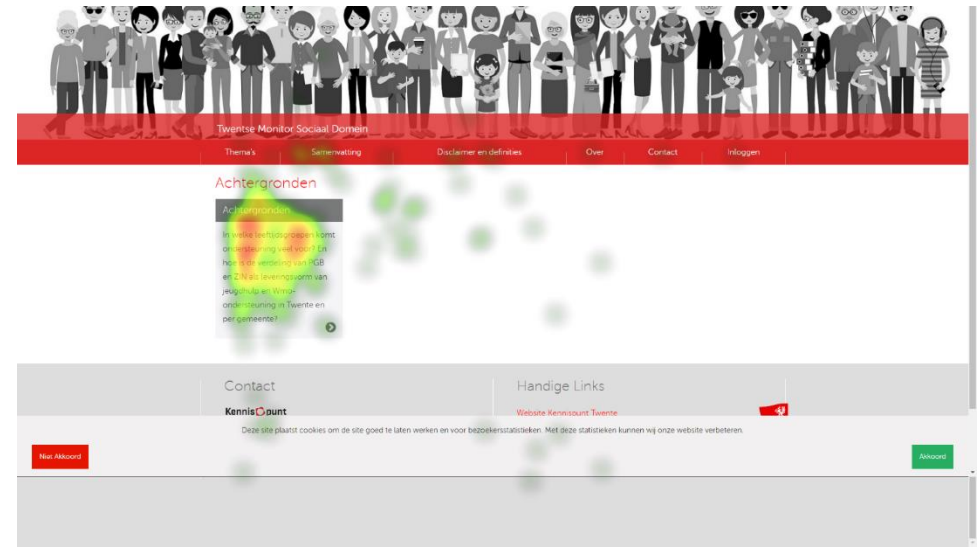
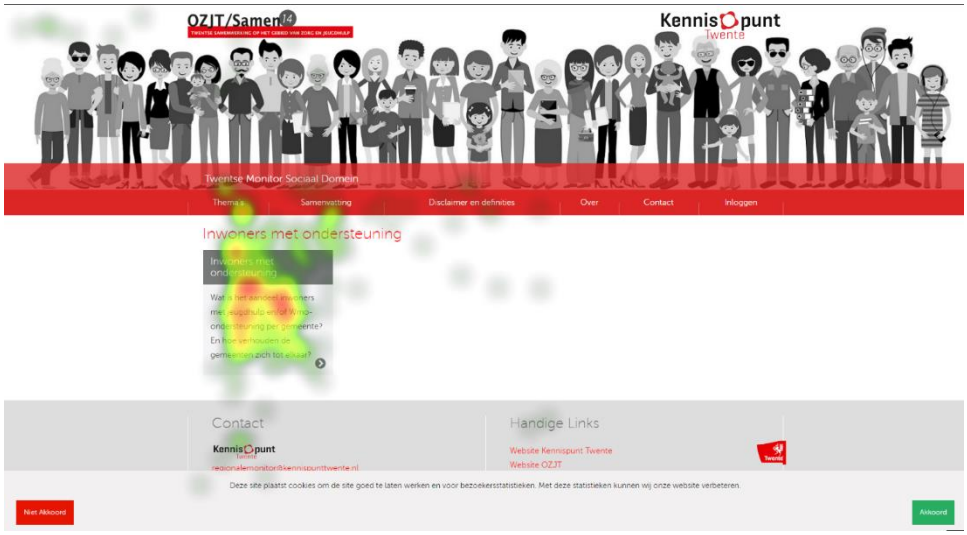


Figure 7.3

Heatmap of inwooners met ondersteuning



Appendix F

Search Log

Figure 7.4

Search log

Source	Search string (databases) or search method (other sources)	Total hits	Remarks
Database name	*Search string (databases) or search method (other sources)*	*Total hits*	*relevant hits*
Web of Science	https://www.webofscience.com/wos/woscc/summary/98ac1199-aa56-48b2-83e4-df3678d70e9d-8101ef08/relevance/1	62	6
Web of Science	"Think aloud method" OR "think aloud proto*" AND Usa* https://www.webofscience.com/wos/woscc/summary/796e18b4-c7f8-49a7-b2f3-86fb20fe99d5-82ced6ef/relevance/1	76	7
Scopus	e-gov* AND eye-tracking	6	2
Scops	retrospective think aloud AND Eye tracking	5	2
Scopus	e-gov* AND usability AND think aloud	7	1

Scopus	eye tracking AND RTA	25	6
Scopus	system usability scale	455 6	5
Scopus	dashboard AND ("usability" OR "Eye Tracking" OR RTA)	350	15