



Utrecht University

Universiteit Utrecht
Master Psychology: Applied Cognitive Psychology

THESIS

Quantitative vs. qualitative

A comparison of methods for improved
usability research

Nova A. Eeken (4014693)

Dr. J.S. Benjamins
Dr. A. Keizer

M. Klooster
(Ruigrok NetPanel)

Abstract

One of the most common ways to perform usability research is by direct observation and questioning. However, there are various psychological and social factors that can influence participants' behavior when usability research contains explicit self-reporting. It would therefore be very useful to circumvent the subjectivity of traditional usability research by measuring actual behavior, instead of relying on the verbal report of participants. The purpose of this research was, therefore, to determine whether or not quantitative usability research methods provide different, or better insights into the usability of a product than traditional qualitative usability research methods. To explore this, quantitative research was conducted using eye tracking, mouse metrics, EEG, facial expressions and the System Usability Scale ($n = 48$), as well as qualitative research using observation and questioning ($n = 8$). Results showed that facial expression analysis and EEG were not particularly suitable for usability purposes and did therefore not provide useful insights. However, eye tracking in combination with mouse metrics identified a larger amount of usability problems (which are also more specific and detailed) compared to traditional qualitative research. On the other hand, qualitative analysis revealed the reasons behind usability problems: something which quantitative analysis could only speculate about. By combining both quantitative and qualitative approaches, the results from qualitative usability research could be an excellent starting point for further in-depth usability research with eye tracking and mouse metrics.

Table of contents

Theoretic outline	4
Traditional usability research.....	4
Eye tracking.....	5
Emotion measurement.....	6
The current research	6
Method	8
Participants.....	8
Stimulus	8
Task	8
Procedure.....	10
Data analysis.....	10
Results	13
Eye tracking analysis	13
Clickstream analysis	20
FaceReader™ analysis.....	26
EEG analysis.....	28
Questionnaire analysis	30
Qualitative analysis.....	33
Meta-analysis	36
General discussion	39
References	42
Appendix	46
Informed consent	46
Questionnaire.....	48
Correlation matrix.....	49

Theoretic outline

Traditional usability research

The term usability can be described as a quality feature that assesses how easy it is to use a systems' interface (Nielsen, 2012). This involves facilitating effectiveness, efficiency and satisfaction in a specific context of use, but should not be confused with the concept of functionality. A usability problem, therefore, is a series of negative phenomena that are caused by a combination of design factors and the context of use (Manakhov & Ivanov, 2016). We speak of a usability problem when the interface of a system or webpage causes users to experience discontentment or frustration, perform unneeded or inefficient interactions or even make it impossible for them to achieve their goals. One of the most common ways to perform usability research and get user feedback, is by direct observation and questioning. These procedures are often supported by methods such as thinking-aloud protocols, one of the most widely used usability research techniques since the late eighties (McDonalds, Edwards & Zhao, 2012; Nielsen, 1993; Green 1995; Kuusela, Spence & Kanto, 1998, Haak & Jong, 2003). When users verbalise their thoughts when interacting with a system, the thinking-aloud method makes it possible to analyse the underlying mental process (Kuseela & Paul, 2000). But despite of the previously named benefits, there are various psychological and social factors that can influence participants' behavior when usability research contains explicit self-reporting. Research showed that consciously thinking out loud affects the way information is assimilated, which also influences the cognitive decision-making process (Schooler, Ohlsson & Brooks, 1993; Wilson & Schooler, 1991). Biehal and Chakravarti (1989) state that it causes people to use a more systematic approach, that they tend to make more rational decisions and also have a significantly better understanding of their task (Kuseela & Paul, 2000). It can also increase their motivation to succeed and provoke answers that are more socially desirable.

In addition, the human capability of evaluating their experiences doesn't prove to be very reliable either. An extensive range of research shows that our assessment of events does not take all the details of the actual experience into account (Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Redelmeier & Kahneman, 1996). In fact, our assessment is mainly based on two moments of the actual experience: the moment of peak intensity (for example pain, frustration or joy) and the emotion experienced last. This selective system results in the (often surprising) rejection of details, which makes it evident that how people *think* they have experienced an event can differ greatly from how it actually took place.

Thinking out loud proves to be especially impractical when people can not clearly explain why they made a mistake or approached a problem in a certain way. This phenomenon can be explained by the psychological *system theory* (Parreren, 1971) that assumes that a learning experience always creates psychological "trails". For instance, one of the first techniques you need to master when learning how to cycle, is steering. But in addition, there are other trails such as balancing, breaking and using the pedals. Although these trails start off as individual acts, they now show a strong cohesion and start to form a whole: a system. An experienced cyclist applies all the techniques at the same time without identifying them as separate actions anymore. As a result it can become difficult to tell systems apart. Furthermore, when learning systems show a strong resemblance to each other, system separation becomes more problematic. This is where *interference* can occur: someone actualizes a behavioral trail that belongs to the wrong system. Since we have to deal with a lot of complex

learning systems, it's not surprising that the majority of people cannot completely verbalize their own interference during a thinking-aloud session (Çöltekin et al., 2009).

It would therefore be very useful to circumvent this subjectivity of common usability research. Implicit measurement of visual attention through eye movements and/or implicit measurement of emotion offers promising possibilities (Bojko, 2006; Goldberg et al., 2002).

Eye tracking

Eye tracking is measuring the motion of a subjects' eyes, relative to a visual stimulus presented on a screen (Smith, 2013). This technique is based on the specific reflection pattern that infrared lighting creates when directed at the human eye. This doesn't trouble the subject, since infrared lighting (IR) is invisible to the human eye. When the eye is illuminated with IR, the portion of light that shines on the pupil is not reflected back. This results in a 'dark' pupil with a small glint: the corneal reflection. Although the pupil shifts and moves with the eye, the glare will always stay in the same position relative to the IR light source. This allows the eye tracker to determine precise eye positions at any given moment.

The use of eye tracking for usability purposes leans on the assumption of the *eye-mind principle* (Just & Carpenter, 1976). This theory argues that when someone looks at a visual stimulus, that information is automatically being processed in the brain. Due to the human visual acuity limitation it's not possible to look at the whole visual scene at once, so we have to keep moving our eyes so that we can mentally absorb small sections of the scene one by one (Van der Stigchel, 2015). Consequently, every movement that the eye makes indicates a new phase of visual processing. When our eyes stabilize at a point in a stimulus, this is called *fixating* (Henderson & Hollingworth, 1999). Each fixation varies in duration, depending on the complexity of the visual stimulus and the associated task (Henderson, 2003). To process a new part of the scene, the eyes must move. This results in rapid eye movements, also known as *saccades* (Salvucci & Goldberg, 2000). The pattern of fixations and saccades creates a *scan path*: a map that shows where the eyes have been (Ehmke & Wilson, 2007). Information around the point of fixation is not disregarded completely, but processed in a different manner. This is because the processing of peripheral, visual information is mainly reserved for the selection of future goals for the next saccade, tracking of moving targets and interpreting the essence and layout of the stimuli. Capturing fixations through eye tracking, therefore, is a great measure of what has been consciously analysed in detail (Findlay & Gilchrist, 2003). However, the interpretation of eye tracking data remains highly dependent on the research purpose and the stimuli. For example, the number of fixations is an important search efficiency indicator when a participant is looking for a link. But when the participant is browsing through an online photo album, a higher number of fixations implies increased interest in certain photos. Eye tracking measurements in usability can, amongst other things, be used to:

- Express how easy it is to find links, buttons and targets. This can be determined by looking at the percentage of subjects that fixated on the correct link, button or target in the first place, the number of fixations *before* the first fixation on the target and the time to first fixation on the target.
- Determine whether the underlying action of a link, button or target is understood. This requires determining how often the subject looks at the target (number of fixations on the target) before selecting or clicking it. Also, the time between the first fixation and selection needs to be measured. If the target is meaningful, this number will be close to zero.

- Reveal whether the website contains complex components. To do so, the *dwell time* on certain areas must be determined. Dwell time is the total time spent looking at a certain area, calculated by summing the time that fixations were located within that area. It is important to emphasize that dwell time not only depends on the number of fixations, but also on the average fixation duration of those fixations. A longer fixation time can indicate processing difficulties.
- Pinpoint whether there are parts on the website that are distracting or creating obstruction to the execution of the task. This requires checking the number of fixations on areas that are not relevant to task completion.

Emotion measurement

In order to improve the usability of a product, insight into your users' emotion can be particularly useful. This is because emotions greatly influence attention and the way information is stored in our memory (Talarico, LaBar, & Rubin, 2004; Phelps, Ling, & Carrasco, 2006). Not only does it put users in a bad mood: experiencing negative emotions has a lot of adverse, negative effects on attention. For instance, the ability to store and remember peripheral details is reversed under the influence of frustration or anger. It also reduces the size of the 'spotlight of attention', which causes the attention to automatically shift to flashy features of the visual scene (Fredrickson, 2004; Fredrickson et al., 2003). And although these elements are very noticeable, it does not guarantee that these are the appropriate starting points to accomplish the task. On a webpage that raises negative emotions, users therefore take in information less consciously, which decreases the chance of achieving their goal (Lewiski, 2015). On the contrary, the ability to remember details seems to be significantly higher when one is experiencing positive emotions (Talarico, Berntsen & Rubin, 2009). But positive emotions have another beneficial effect: they broaden an individuals' *thought-action repertoires* (Fredrickson & Branigan, 2005). A *thought-action repertoire* is a series of well known actions that will be executed directly under certain stimulation. Think of fleeing when experiencing fear, or discovering and playing when experiencing joy or excitement. Under the influence of positive emotions, one can thus address a wider range of skills. As a result, this gives users a greater number of tactics to use to achieve their goals, which increases the chance of successfully completing their task. This assumption is supported by studies in which subjects show more flexible and comprehensive thinking patterns under positive stimulation (Isen & Daubman, 1984; Bolte, Goschke, & Kuhl, 2003) and are also more more receptive to information (Estrada, Isen, & Young, 1997). Emotion and usability are therefore highly intertwined.

The current research

The purpose of this research is to determine whether or not quantitative usability research methods on a larger scale create different, or better insights into the usability of a product than the traditional (subjective) qualitative research methods. Additionally, this study hopes to verify relationships between results from different measurement instruments and to discover how they relate to each other. This can be accomplished by testing the usability of a product in different ways, as illustrated in figure 1. The main research question is therefore: “*Are quantitative testing methods (such as eye tracking, EEG, facial expression analysis, mouse tracking and the system usability scale) as effective in identifying usability problems as qualitative research with observation?*”

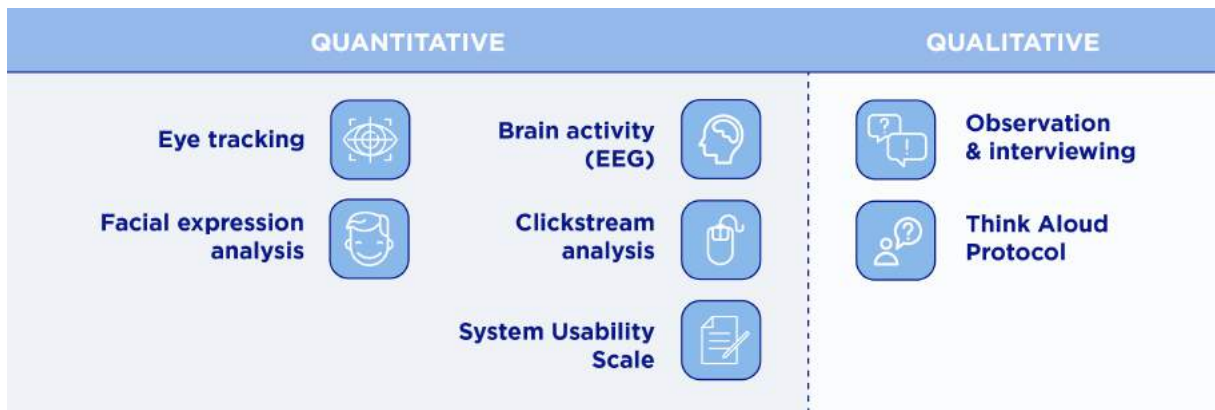


Figure 1: summary of the methods used to identify usability problems

First of all, it is to be expected that mainly eye tracking in combination with mouse metrics will have the ability to identify a higher amount of usability problems than traditional qualitative research. Because of the precise measurement of participant behavior, usability problems are expected to be more detailed and specified. Moreover, since the occurrence of usability problems is known to cause frustration and anger, this will become evident in the form of negative facial expressions. Furthermore, usability problems will likely cause participants to experience stress. Usability problems are expected to be related to higher levels of attention and lower levels of calmness. Lastly, it is likely that the outcome of the System Usability scale and supplementary questions will provide a concise summary of parts of the task in which most usability problems occurred.

The quantitative research methods ($n > 40$) used simultaneously hope to provide insight into:

1. Search patterns, distractions, interface comprehension and complex elements;
2. Concentration, frustration and stress levels;
3. Emotions experienced during the interaction;
4. Bottlenecks, misinterpretations and other non-linear behavior of users;
5. The honest opinion of users in regard to their their experience;

In order to be able to compare the results with traditional usability research, a qualitative session with observation, and in-depth interviews ($n = 8$) will also take place.

Method

Participants

A total of forty-eight subjects participated in this study ($M_{\text{age}} = 34.38$, $SD = 12.73$) including 22 women. All participants had corrected-to-normal eye sight and were asked not to wear dark eye makeup during the experiment. Participants consisted of colleagues, acquaintances and passer-by's who received a "Tony Chocolonely" chocolate bar for their participation.

Stimulus

Website

The website of the Japanese restaurant chain SUMO (<http://restaurantsumo.com/>) was utilized as material for the experiment. Apart from their *all you can eat* concept, the restaurant chain also provides delivery and pick-up services for sushi and other Japanese dishes. To place an order online, the user is taken off the main website and redirected to a separate delivery-environment. This environment is very hard to reach due to the cumbersomeness in the interface and the various bugs in the system make the ordering process troublesome. Moreover, due to the minimum order amount, the 'Place order'-button will not appear until users have selected over €20 worth of dishes. The visual communication of this reason, however, is easily missed. In addition, it is difficult for users to recover from errors, receiving very little feedback from the system and lastly does the website contains many vague content names. Because of these shortcomings, the website only complies with two out of ten usability heuristics by Nielsen and Molich (1990). Therefore, it is particularly suitable to identify usability problems with both (traditional) qualitative research methods and quantitative research methods as reviewed in the introduction. The various quantitative methods used to study the website are discussed below.

Task

Participants were given the task to place an online order using the SUMO website. They could spend around €20 on a sushi set and two separate pieces of sushi.

Tobii X3-120 eye tracker

During the task, eye movements were measured with a Tobii Tobii X3-120 eye tracker that has a sampling frequency of 120Hz. This is a standalone, non-intrusive device that can be placed underneath the screen of any PC or laptop, enabling the subject to have a relatively high degree of freedom to move. The website was projected on a screen under which this eye tracker was placed.

Mouse clicks

During the task, mouse clicks were captured using the Tobii Studio eye tracking software. Based on mouse clicks, each participant's task could be divided into six task segments which will be discussed in more detail in "Data Analysis". The start and end time of participants carrying out parts of the task will vary per individual. Task segments can therefore serve as a guideline to compare data from different instruments. Also considered are click clusters (a high number of mouse clicks within a short period of time), clicks on non-clickable interface elements and non-linear behavior during the order process. These insights prove to be a good trail when looking for usability issues (Kaur & Singh, 2015).

Face reader™

During the task, emotions were measured using FaceReader™: facial expression analysis software that uses precise face modelling (den Uyl & van Kuilenburg, 2005; Drozdova, 2014). The software categorizes facial expressions based on seven universal emotions, identified by Ekman and Rosenberg (1997). These are: a neutral state, happiness, sadness, anger, surprise, fear and disgust. These emotions are logged with a timestamp, allowing them to be linked to specific events in the order process. Recent research shows that FaceReader™ accurately recognizes emotions from 88% of faces and proves to be a reliable indicator of emotional facial expressions (D’Lewis, The Uyl & Butler, 2014; Lewinski, French, & Tan, 2014). The natural bias that subjects may have towards a particular facial expression is corrected by the software using a calibration function, ensuring accuracy of the results. FaceReader™ ran on a separate laptop when participants carried out their task.

Neurosky MindWave (EEG Headset)

During the task, brain activity was measured using the Neurosky MindWave Mobile headset and Myndplayer Pro software. By interpreting brain activity measurements, it is possible to determine cognitive effort. This is done by non-invasively measuring the aggregated signal of action potentials on the skull, also called electroencephalography: EEG (Palaniappan & Mandic, 2007; Lin et al., 2008; Campbell et al., 2010; Peck et al., 2010). The manufacturer Neurosky has developed a wireless EEG headset that distinguishes different types of brain signals based on frequency bandwidth analysis (Hondrou & Caridakis, 2012). The sensor captures frequencies between 3 - 100 Hz within the EEG signal (Neurosky, 2015). Neurosky divides this frequency range into measurements for attention, meditation, and stress. The level of attention and cognitive effort is directly based on the subject's brain activity and produces a value per second on a scale of 0 to 100 for each of the three output types (Crowley et al. 2010).

System Usability Scale

After the task was finished, participants filled out an online questionnaire created with Jambo Software. The questionnaire contained the System Usability Scale as well as additional questions regarding the task. The System Usability Scale was created by Brooke (1996) as a ‘quick and dirty survey scale’ to assess the usability of a system. Comprising ten questions where one half is positively and the other half is negatively correlated, the questionnaire is divided into three segments: effectiveness, efficiency and satisfaction. An example of an item is "I found the system unnecessarily complex." Items are scored on a 5-point Likert scale (0 = strongly disagree, 5 = strongly agree). The total score is determined by adding the sum of all the answers and then multiplying by 2.5. This allows a total score to lie between 0 and 100, with scores less than 60 being considered insufficient, scores between 60-70 sufficient, scores between 70-80 good, scores between 80-90 very good and scores greater than 90 as excellent. Meanwhile it has been one of the most commonly used questionnaires in usability research with more than 4400 citations in scientific articles and publications. The reliability is high, with a Cronbach's α of 0.911 for the total score (Bangor, Kortum, & Miller, 2008). For this study, the questions are translated into Dutch. Because the System Usability Scale cannot identify specific usability problems, additional questions have been added to the questionnaire. These can be reviewed in the appendix.

Equipment

The entire task was performed using a PC with 8GB RAM and a 3.40GHz i7-3770 CPU running Windows 10, working in the Internet browser Mozilla Firefox 42.0.0. During the experiment video recordings were made of the screen, as well as the subject using Tobii Studio 3.4.8 and a Logitech C270 webcam. In addition, there were two monitors connected to the computer with a 55 cm screen diagonal: one displaying the website to the subject and one that visualized eye movements simultaneously for the researcher. The distance between subjects and the screen varied from 60 to 20 cm. Lastly, two keyboards and two mice were present so that both the subject and the researcher could take control of the interface.

Procedure

The experiments were conducted at Ruigrok NetPanel's office in a quiet, secluded observation room where one could not be disturbed. Prior to the test, participants received a brief explanation about the nature of the research and the various measurement instruments. Subsequently, all participants signed an informed consent confirming their agreement to the collection of video recordings and behavioral data. Thereafter the EEG headset was placed on the head and wirelessly connected to a separate laptop via Bluetooth. Then they received the following instruction:

“You fancy some sushi, so you are going to order this online at SUMO. In a moment I’ll show you a visual summary of the task, but I’ll explain it to you first. You are looking for a a sushi set containing only maki (rolls) of your choice and in addition you pick two separate nigiri with prawns. You like to spend around twenty euros, but a few euros above or below doesn’t matter. Furthermore, you would like to have the sushi delivered to you, here in Amsterdam. You may use your own personal details or make them up, but please use the zipcode 1013AL, house number 1. The task is finished after you’ve chosen the desired delivery time. I will then terminate the browser for you. Because we like your behavior to be as natural as possible, just pretend as if I’m not here. This task will take you about five to ten minutes.”

Subsequently, participants were given the opportunity to ask questions before starting the experiment. Right after, the *Myndplay Pro* application was launched to detect the EEG signals. The researcher made sure the subjects face was fully visible within the frame of the webcam using the Tobii Studio software. Thereafter the the X3-120 eye tracker was calibrated on five points on the screen using the built-in Tobii calibration method. The task initiated by displaying the SUMO website on the monitor. After task completion the browser was terminated and all measurement recordings stopped and saved. Thereupon, subjects completed the online System Usability Scale questionnaire and additional questions regarding their experience with SUMO’s order process. Lastly subjects were offered a chocolate bar and thanked for their participation. During the quantitative sessions the observer did not interfere with the course of the task, since this could affect the measurements.

Data analysis

Statistical analysis

Analysis is performed using SPSS, Excel and Tobii Studio. In order to properly compare the different

measurement instruments and datasets, each participants' task will be divided into six task segments based on mouse clicks:

1. Locating the order environment;
2. Selecting the desired dishes;
3. Starting the checkout process. This is possible by clicking the 'Checkout'-button. However, this button will only appear when the participant has over twenty euros worth of dishes selected.
4. Checking the order summary to then click the 'Continue'-button;
5. Entering personal details to then click the 'Continue'-button;
6. Selecting the desired delivery time;

From here on, segments will be referred to as: 'Locating order environment', 'Selecting dishes', 'Starting checkout', 'Reviewing order', 'Entering details' and 'Selecting delivery time'.

Reporting usability problems

All unique usability issues will be assigned a segment number and are expressed in severity, magnitude and the number of participants that experienced this problem. Magnitude will be expressed in either 'global' for problems concerning the interface as a whole, or 'local' for isolated problems that only apply to a part of the interface. Severity will be expressed on a rating scale explained in table 1, ranging from 1 to 4 (Dumas & Redish, 1999).

Severity	Description	Explanation
1	Subtle problem or possible improvement	The problem occurs occasionally and can easily be circumvented. This can also be a cosmetic problem.
2	Has a moderate negative effect on usability	Users are able to use the product, but it requires some effort to get around the problem.
3	Creates significant delay and frustration	Users will try to use the product, but will be severely limited in their ability to do successfully do so.
4	Prevents task completion	Users are unable or unwilling to use the product because of the way it is designed and implemented.

Table 1: an explanation of severity ratings used to classify usability problems

Eye movements and mouse clicks: AOI analysis

For this research's purpose, webpages most relevant to task completion are divided into Area's of Interest (AOI's). An AOI is a specified part of the interface onto which a participant's visual attention can be directed. The data can then be compared relative to these AOIs. Two different AOI groups were specified for this research:

- *Web elements.* Webpages are divided into AOIs based on the interface, such as "header", "title" and "target" that are used for eye movement analysis.
- *Clickable elements.* AOI's will be created for interface elements that are hyperlinks or invoke a change in the interface in any other way, such as buttons, links or adding a dish to the order by clicking on an image. These are used for mouse click analysis.

Furthermore, the number of pages and mouse clicks needed to successfully accomplish the task will be compared to the number of pages and clicks made by participants.

Analysis of EEG measurements

Because the overall strength of raw brainwave signals varies per individual, absolute values cannot be compared across subjects. Therefore, for every type of brainwave signal, the differences per task segment relative to the overall mean value of that participant's signal type will be calculated. Aside from the values of 'attention' and 'meditation' calculated (and already normalised) by the *Myndplay Pro* application, the main focus will lie on the analysis of beta-, theta-, and alpha waves. When individuals experience stress, anxiety or frustration they often show a greater amplitude between the high beta, and low beta waves. The segments in which the average beta-wave difference is large, could therefore indicate usability issues. Alpha brainwaves are related to relaxation and calmness, yet being alert. Theta brainwaves are associated with accessing memory, creative inspiration and excitement (Green & Arduini, 1954). Segments in which alpha or theta brainwaves are low, could therefore also indicate usability issues.

Analysis of facial expressions

The mean value and standard deviations of emotions per task segment will be calculated for every participant and reviewed to determine if expressions differed across segments. FaceReader™ can also compute *valence*: an indication of how positive or negative one is during the course of the task. This is calculated by subtracting the value of the highest negative emotion from the value of happiness.

Qualitative analysis

All data related to relevant observations, comments, answers to open-ended questions and experienced problems will be described in a concise report, while answering the following questions:

- How do users experience SUMO's order process?
- What are the key points to optimize the order environment?

Meta-analysis

For the meta-analysis, the nature and amount of usability problems that are found across qualitative and quantitative methods will be compared. In addition, a correlational comparison of methods using a Spearman's rank-order correlation will be performed to identify possible relationships between data sets from different quantitative methods.

Results

It has to be taken into account that the sample size of this study is relatively small for quantitative analysis. Because the goal of this research is not to generalize results to the population, but to draw conclusions from the current dataset, the p -value will be mentioned, but not used as differentiator for statistical significance. Instead, eta-squared (η^2) will be utilized as an effect size measurement of difference between variable groups.

Eye tracking analysis

Out of 48 participants, 14 were removed from the eye tracking dataset because of bad calibration scores or data validity scores below 66%. Some usability problems such as “I expected another option to choose from” or “I think this font size is too small” that were only verbalised by the participant could not be recorded by eye tracking and are therefore not included in this section. A total of 20 usability problems have been identified using eye tracking analysis. Five of them rated a severity of ‘4’, eight of them rated as ‘3’ and seven problems rated as ‘2’.

	SEGMENT	PROBLEM	SUBJECTS	MAGNITUDE	SEVERITY
1.	1 (Locating Order Environment)	Subjects did not interpret the delivery page link as a target (and therefore could not reach the delivery environment)	33	Global	4
2.	1 (Locating Order Environment)	Subjects were not aware that content existed below the main menu on the home page (and therefore never saw the home page target)	20	Local	2
3.	1 (Locating Order Environment)	Subjects had trouble interpreting the home page link as a target (and therefore took longer to reach the delivery environment)	9	Local	3
4.	1 (Locating Order Environment)	Interpretation of menu items such as ‘menu’, ‘delivery’, ‘restaurants’ and ‘social’ in the main navigation was complex	30	Global	3
5.	1 (Locating Order Environment)	Interpretation of navigation tiles such as ‘menu’ and ‘delivery’ on the home page was troublesome	9	Local	3
6.	2 (Selecting Dishes)	On the delivery home page, subjects look at irrelevant items first since the target element (zip code field) is placed in the right-bottom corner of the page	32	Local	2
7.	2 (Selecting Dishes)	Subjects did not see the main target (zip code field)	2	Local	4
8.	2 (Selecting Dishes)	The meaning of the subtarget (menu item: ‘menu’) is a lot harder to interpret than the call to action ‘Bestellen’ above the main target	16	Local	3
9.	2 (Selecting Dishes)	The implication of the main target (zip code field) seems unclear as subjects keep fixating on other areas afterwards.	13	Local	2

10.	2 (Selecting Dishes)	The interpretation of the left navigational menu item 'Sushi' was troublesome (because participants were searching for nigiri)	26	Global	3
11.	2 (Selecting Dishes)	The interpretation of the left navigational menu item 'Sushi menus' was troublesome (because it had a strong resemblance with 'Sushi sets')	26	Global	3
12.	3 (Starting Checkout)	It was impossible to place an order since there was no visible check-out button	12	Local	4
13.	3 (Starting Checkout)	Subjects experienced great difficulty trying to figure out where to find the check-out button	15	Local	3
14.	3 (Starting Checkout)	Participants did not fixate on the goal information regarding a minimum ordering amount	10	Local	4
15.	3 (Starting Checkout)	Participants did fixate on the goal information regarding a minimum ordering amount, but did not interpret this as the cause of the problem	17	Local	4
16.	3 (Starting Checkout)	Because of the visual hierarchy, subjects did not view the parts of the goal information sequentially	25	Local	3
17.	4 (Review Order)	Participants did not notice the process indicator	29	Global	2
18.	5 (Entering Details)	Participants did not see the progress indicator	22	Global	2
19.	5 (Entering Details)	Subjects interpreted the returning-customers fields as personal detail field	15	Local	2
20.	5 (Entering Details)	The form did not clearly differentiate between new- and returning customers, causing subjects to fixate repeatedly on irrelevant form fields	34	Local	2

Table 2: a listing of usability problems by number, segment, number of participants who experienced it, magnitude and severity rating.

Segment 1: locating the the delivery environment

There are two targets that redirect users to the delivery environment: one on the home page and one on the delivery page. All ($n = 34$) participants fixated on at least one target and some participants ($n = 8$) fixated on both of them. The time to first fixation on a target link was lower for the delivery page target ($M = 5.07$ s, $SD = 7.18$ s) then for the home page target ($M = 28.521$ s, $SD = 25.591$ s). This is due to the fact that the number of gaze visits below the main navigation on the home page is equal to 0 for a large proportion of subjects ($n = 20$). This indicates that participants were not aware that content existed below the fold: the bottom half of the browser window that only becomes visible by scrolling. Moreover, participants looked at- and away from their target repeatedly before actually clicking it ($M_{\text{gazevisits}} = 6.706$, $SD_{\text{gazevisits}} = 3.196$).

The delivery page target was viewed by almost all subjects ($n = 33$) in comparison to the home page target that was only found by a portion of the subjects ($n = 13$). However, on average it took subjects a higher number of gaze visits to interpret the delivery page target ($M_{\text{gazevisits}} = 6.697$, $SD_{\text{gazevisits}} = 4.385$) then to interpret the home-

page-target ($M_{\text{gazevisits}} = 4.385$, $SD_{\text{gazevisits}} = 2.399$). This shows that the delivery page target was found sooner, but also required more gaze visits before subjects clicked it. The amount of fixations between the first fixation and mouse click on target was almost twice as high for the delivery page target ($M = 61.34$, $SD = 51.56$) compared to the home page target ($M = 33.03$, $SD = 35.89$). This is very high, regarding the fact that ideally, this number needs to be as close to zero as possible.

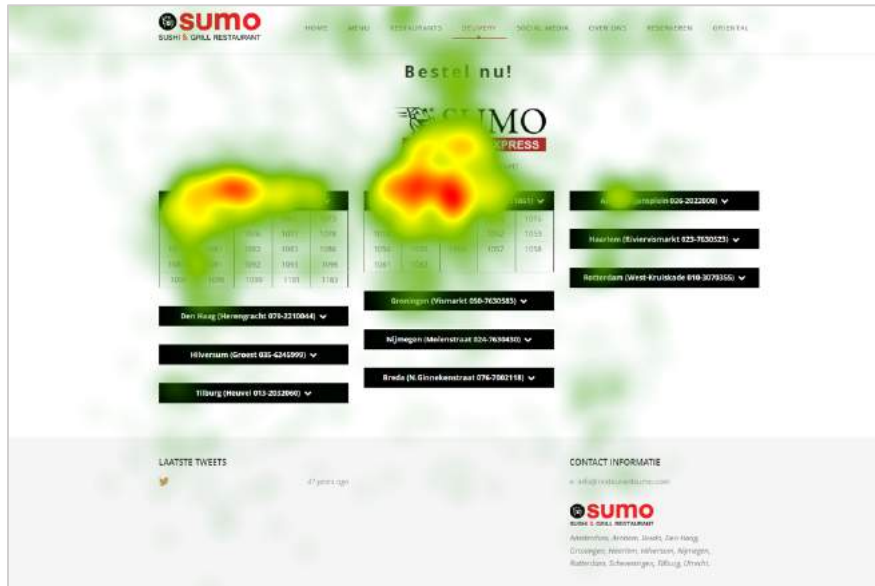


Image 1: heatmap based on relative fixation duration on the Main | Delivery page ($n = 34$). The zip code fields draw the most attention to them, as opposed to the target link: the Sumo Express logo.

The results of the one-way repeated-measures ANOVA showed that there was a large effect of the name of menu-items on the average fixation duration of participants ($F(4.624, 110.967) = 5.728$, $p = .000$, $\eta_p^2 = .193$). Bonferroni post hoc tests showed that participants fixated significantly longer on 'Delivery' ($M = .622$ s; $SD = .242$ s) and 'Restaurants' ($M = .533$ s; $SD = .157$ s) compared to the other menu-items. Moreover, there was a significant effect of the name of menu-items on the number of fixations ($F(4.622, 98.028) = 16.843$, $p = .000$, $\eta_p^2 = .423$). Bonferroni post hoc tests showed that participants fixated significantly more often on 'Delivery' ($M = 7.831$ s; $SD = 3.460$ s) compared to 'Oriental', 'Over ons' and 'Reserveren'. Furthermore, participants fixated on 'Restaurants' ($M = 9.46$, $SD = 4.530$) more often compared to all other menu-items, except 'Delivery'. Lastly, the fixation count was significantly higher on 'Menu' ($M = 6.96$, $SD = 2.851$) and 'Social' ($M = 6.71$, $SD = 2.836$) compared to 'Oriental' and 'Over ons'. This supports the claim that the meaning of 'delivery', 'menu', 'restaurants' and 'social' is less clear to subjects compared to other menu items. Since very few subjects fixated on *all* the home page tiles, it was not possible to run a statistical analysis to differentiate between them. However, the average fixation count of the 'menu tile' ($M = 15.78$, $SD = 9.271$) and the 'delivery tile' ($M = 15.44$, $SD = 10.901$) were almost twice as high compared to the more complex menu-items.

Segment 2: choosing dishes

When first arriving at the order environment home page, almost all participants ($n = 32$) were able to find the main target (zip code field). However, it took them almost 4 seconds ($M = 3.926$ s, $SD = 3.210$ s) and 12 fixations ($M = 12.440$, $SD = 8.332$) on this page to locate it. The sub-target (top navigation item 'Menu') was

seen was seen even later. The participants that located this target ($n = 17$) were able to do so after 7 seconds ($M = 7.194$ s, $SD = 4.018$ s) and 24 fixations ($M = 24.240$, $SD = 13.465$). The time to first fixation on several areas of interest proves that the visual hierarchy of the order environment home page is not optimized for the task, as illustrated in table 3.

Time to first fixation on area of interest	Mean	Std. Deviation
Header	1,794 s	3,446 s
Establishments information (“Vestigingen”)	2,029 s	3,062 s
Discount information (“5% karting”)	2,852 s	2,920 s
Target title (“Bestellen”)	3,478 s	2,973 s
Zipcode field (main target)	3,926 s	3,210 s
Top navigation item ‘Menu’ (subtarget)	7,194 s	4,018 s
Login field	7,413 s	5,225 s

Table 3: The time to first fixation on area of interest on the Order | Home page. The target (zip code field) is does not attract enough attention.



Image 2: The viewing order of participants who visit the Order | Home page for the first time. It shows that the target element (in yellow) which is expected to attract the most attention, is viewed fourth.

Although one would expect the target title to catch the eye, almost all participants looked at the the area explaining discounts and the area containing information about the SUMO establishments first. In addition, the discount information ($M_{\text{visitcount}} = 2.000$, $SD_{\text{visitcount}} = 1.390$), establishment information ($M_{\text{visitcount}} = 3.500$, $SD_{\text{visitcount}} = 1.723$) and login area ($M_{\text{visitcount}} = 2.220$, $SD_{\text{visitcount}} = 1.502$) received multiple unnecessary gaze visits. A ‘gaze visit’ starts with the first fixation within an AOI and ends with the first fixation outside of that AOI. Based on the average fixation duration, the main target (zip code field) seems to be interpreted quicker ($M = .206$ s, $SD = .068$ s) than the subtarget (menu) at the top of the page ($M = .315$ s, $SD = .285$ s), found in the main navigation. This is probably due to the clear statement of the main target title “Bestellen” ($M_{\text{fixationduration}} = .187$ s, $SD_{\text{fixationduration}} = .062$ s). On the other hand, the time between first fixation and mouse click for participants that clicked on the main target zipcode field ($n = 25$) was about 1.91 seconds longer ($M = 4.307$ s, $SD = 5.652$ s) and 1,73 fixations larger ($M = 4.840$, $SD = 2.267$) than the the subtarget ($n = 9$).

The results of the one-way repeated-measures ANOVA showed that there was a large effect of the name of menu-items on the average fixation duration ($F(3.378, 84.459) = 6.58, p = .001, \eta_p^2 = .218$). Bonferroni post hoc tests showed that participants fixated significantly longer on ‘Sushi’ ($M = .630$ s; $SD = .343$ s) and ‘Sushi menus’ ($M = .617$; $SD = .368$) compared to ‘Maki’, ‘Salads’ and ‘Small sushi sets’. However, neither menu items’ fixation duration significantly differed from ‘Sashimi’ ($M = .601$ s; $SD = .432$ s). This supports the claim that the meaning of these menu items is less clear.

Segment 3: starting check-out

In 79% of cases ($n = 27$) the interface did not display a “check-out”-button, due to the fact that the total value of the order remained below the minimum order amount of €20. Eventually, after declaring it was impossible to proceed, 12 participants received a hint from the observer. Only 15 subjects found the cause of the problem by themselves. During the first 20 seconds subjects started their search for the “check-out”-button, 23 participants eventually fixated on the line of text saying “Minimum order amount:” and 18 participants eventually fixated on number “€20”. But although these targets were seen, participants did not interpret this as the cause of the problem. During the first failed search attempts ($n = 12$) it became evident that participants did not fixate on either, or only one of the targets. In the attempts that followed, participants did fixate on both targets, but not sequentially ($n = 25$). Therefore, they did not interpret both parts of the target as one, nor did they link this to the currently selected value of their shopping basket.

The time to first fixation on several areas of interest proves that the visual hierarchy of this page does not lead subjects to their target in the proper order. This is illustrated in image 3.

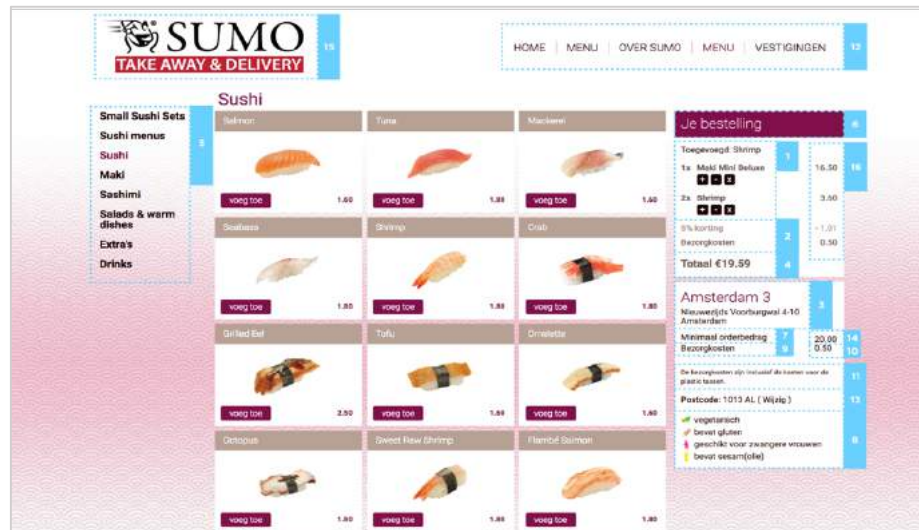


Image 3: the average viewing order of participants when looking for a 'checkout'-button in segment 3. Starting at the shopping basket content, they work their way down, mainly focusing on the left side of the summary elements.

Segment 4: reviewing the order

All participants located the target: the “Proceed”-button. On average, subjects gazed at this target more than once before clicking it ($M_{\text{visitcount}} = 2.324, SD_{\text{visitcount}} = 1.365$). Based on the average fixation duration, participants did not have trouble interpreting the target ($M = .189$ s, $SD = .547$ s). Furthermore, based on the time between first fixation and next mouse click, participants were sure about the meaning behind this goal ($M = 2.374$ s, $SD = 1.355$ s).

All participants fixated on the target, as well as on the main summary of their order. A total of 27 subjects fixated at least once on the top right summary ($M = 2.519$, $SD = 1.365$), 20 subjects fixated at least once on the bottom right information section ($M = 1.850$, $SD = 1.182$), and only 5 subjects fixated on the progress indicator ($M = 1.800$, $SD = 1.304$).

Segment 5: entering personal details

All participants located the target: the “Proceed”-button. On average, subjects gazed at the target button more than once before clicking it ($M = 1.970$, $SD = 1.167$). The time between first fixation and next mouse click was a lot higher compared to segment 4 ($M = 42.054$ s, $SD = 24.344$ s) but this is probably due to the fact that subjects had a tendency to scan the page before they started to fill in their details. Furthermore, the average fixation duration on the target button was also higher than segment 4 ($M = .253$ s, $SD = .170$ s).

Based on the amount of fixations that took place beforehand, the target fields were not the first element that caught participants’ attention ($M = 4.470$, $SD = 5.212$). The time to first fixation ($M = 1.357$ s, $SD = 1.810$ s) confirms this. The main interface elements were viewed top to bottom, as illustrated in table 4.

Time to first fixation on area of interest	Mean	Std. Deviation
Fields for returning customers (“Log in”)	.663 s	.585 s
Fields for new customers (“Gegevens”)	1.357 s	1.900 s
Password fields (“Wachtwoord aanmaken”)	31.877 s	20.556 s
Target button (“Ga verder”)	42.054 s	24.244 s

Table 4: the time to first fixation on area of interest in segment 5 ‘Entering details’. Subjects show a clear vertical viewing direction.

In contrast, the ‘returning-customers’ section was seen almost immediately ($M = .192$ s, $SD = .061$ s) since it is located at the top of the page. Although this field was not applicable to all participants, the time between first fixation and mouse click was much lower ($M = 3.341$ s, $SD = 2.676$ s) compared to the target fields ($M = 6.521$ s, $SD = 7.075$ s). The results of the one-way repeated-measures ANOVA showed that there was a large effect of the form-section on the average fixation duration of participants ($F(1.648, 54.373) = 6.598$, $p = .005$, $\eta_p^2 = .167$). Bonferroni post hoc tests showed that participants fixated a lot longer on the ‘personal details’-section ($M = .232$ s, $SD = .058$ s) compared to the ‘returning-customers’-fields ($M = .196$ s, $SD = .0564$ s) and the ‘password fields’ ($M = .192$ s, $SD = .061$ s). However, the ‘returning-customers’-fields and ‘password fields’ did not significantly differ from each other.

Segment 6: selecting delivery time

No usability problems were found during this segment.

Discussion

Segment 1: ‘Locating order environment’. It became clear that participants found the delivery page target rather quick, as opposed to the homepage target that was first seen very late in the first segment, or not at all. This was due to the fact that a lot of participants did not scroll down the *Main | Home* page. This behaviour could be caused by the screen filling header image, which does not make users aware of the fact that there is content below the main navigation. The amount of gaze visits on the homepage target also indicated that participants were unsure whether or not this was the target they were looking for, even though it contained the

word ‘delivery’. Furthermore, results also showed that although the delivery page target was found more easily than the home page target, it took participants a lot longer to interpret it as the entrance to the delivery environment. Lastly, participants had more trouble interpreting the meaning of the main navigation items ‘Delivery’ and ‘Restaurants’ compared to the other menu items.

Segment 2: ‘Selecting dishes’. The time to first fixation on several interface elements proves that the visual hierarchy of the *Order | Home* page is not optimized for the task. The target (zip code field) is viewed as one of the last items on the page. Moreover, the meaning of the left navigation items ‘Sushi’ and ‘Sushi menus’ seemed vaguer to participants than the other menu items, based on average fixation duration.



Image 4: a screenshot of the shopping basket, where targets are indicated in blue.

Segment 3: ‘Starting checkout’. During the first failed attempts of finding the ‘checkout’-button, it became clear that participants did not fixate on either, or only a portion of the target. In the attempts that followed, participants *did* fixate on both targets, but not sequentially. As a result, they did not interpret both parts of the target as one, nor did they link this to their current order total. Illustrated in image 4, this might be caused by the fact that the text [01] is placed relatively far apart from the amount [02] and the current order total [03].

Segment 4: ‘Reviewing order’ / Segment 5: ‘Entering details’. It became clear that very few participants saw the process bar during segment 4 and 5, indicating they did not have a clear realization of the length of the checkout process. Furthermore, the form fields on the *Checkout | Details* page were viewed top to bottom, making the returning customers field (which was not applicable to any of the participants) the first element participants fixated on.

Clickstream analysis

Eighteen usability problems have been identified using clickstream analysis. Two of them rated a severity of '4', five of them rated as '3', seven problems rated as '2', and and four problems rated as '1'. After subtracting the amount of clicks *needed* for task completion from the amount of clicks *used* for task completion, analysis showed that 51% of 2738 registered clicks were unnecessary. Most of these extra clicks occurred during 'locating order environment' (segment1), 'selecting dishes' (segment 2) and 'starting checkout' (segment 3). From the 865 clicks made in segment 1, 83% exceeded the maximum amount of clicks needed to find the delivery environment. In segment 2 this was 47% out of 633 clicks. During segment 3 this ratio was the highest: 90% out of 501 clicks were unnecessary.

Segment	Clicks necessary for task completion (<i>per subject</i>)	Expected total number of necessary clicks (<i>n = 48</i>)	Number of clicks made (<i>per subject</i>)	
			<i>M</i>	<i>SD</i>
1. Locating order environment	1 - 3	48- 144	18.021	9.786
2. Selecting dishes	7	336	13.188	7.442
3. Starting checkout	1	48	10.438	11.769
4. Reviewing order	1	48	1.063	.247
5. Entering details	7 - 10	336 - 480	9.729	6.115
6. Selecting delivery time	0 - 6	288	4.804	2.986
Total	17 - 28	816 - 1344	57.042	19.742

Table 5: the maximum amount of clicks per segment necessary to successfully complete the task, compared to the average number of clicks made per segment (per person).

The results of the one-way repeated-measures ANOVA on segments showed that there was a very large effect of task segments on the amount of extra clicks needed to complete the task ($F(2.444, 109.987) = 31.860, p = .000, \eta_p^2 = .415$). Bonferroni post hoc tests showed that participants needed significantly more extra clicks in segment 1 ($M = 15.439; SD = 9.786$) and 3 ($M = 9.587; SD = 12.003$) compared to the other segments. Furthermore, this was also the case in segment 2 ($M = 5.978; SD = 7.532$) and 5 ($M = 2.217; SD = 4.487$) compared to segment 4 ($M = .062; SD = .250$) and 6 ($M = .522; SD = 1.362$). As illustrated in figure 2, the amount of extra clicks in segment 1 and 3 did not significantly differ from each other, nor did segment 2 and 3, 4 and 6.

In addition, analysis showed that out of the 1064 page visits made by participants across segments, 64% was unnecessary for task completion. Those page visits were made almost exclusively in segments 1, 2 and 3. Furthermore, the amount of clicks that took place on non-clickable elements of the interface is illustrated in figure 3. The portion of miss-clicks seems to be highest in segment 1 (44%) and segment 3 (38,5%).

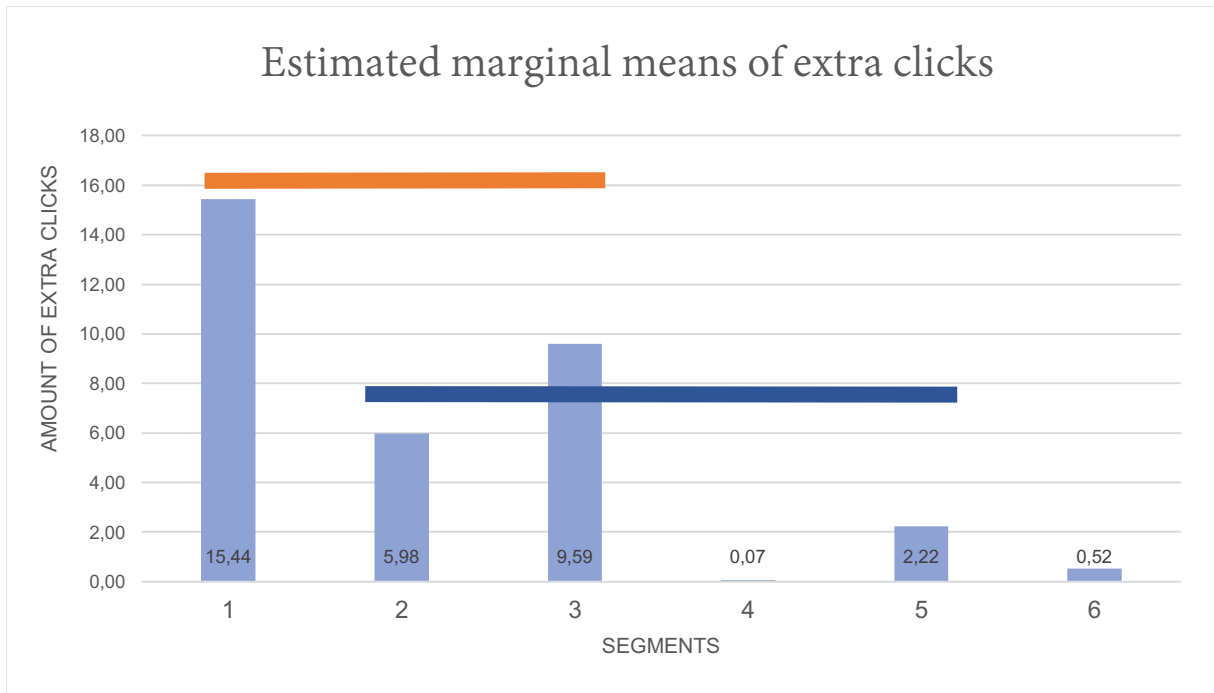


Figure 2: the average amount of clicks (per person) exceeding the amount necessary clicks needed to successfully complete the task. The amount of extra clicks in segment 1 and 3 (below the orange line) are both significantly higher compared to the other segments. The amount of extra clicks made in segment 2 and 5 (below the blue line) significantly differ from each other, and are also both significantly higher compared to the number of unnecessary clicks in segment 4 and 5.

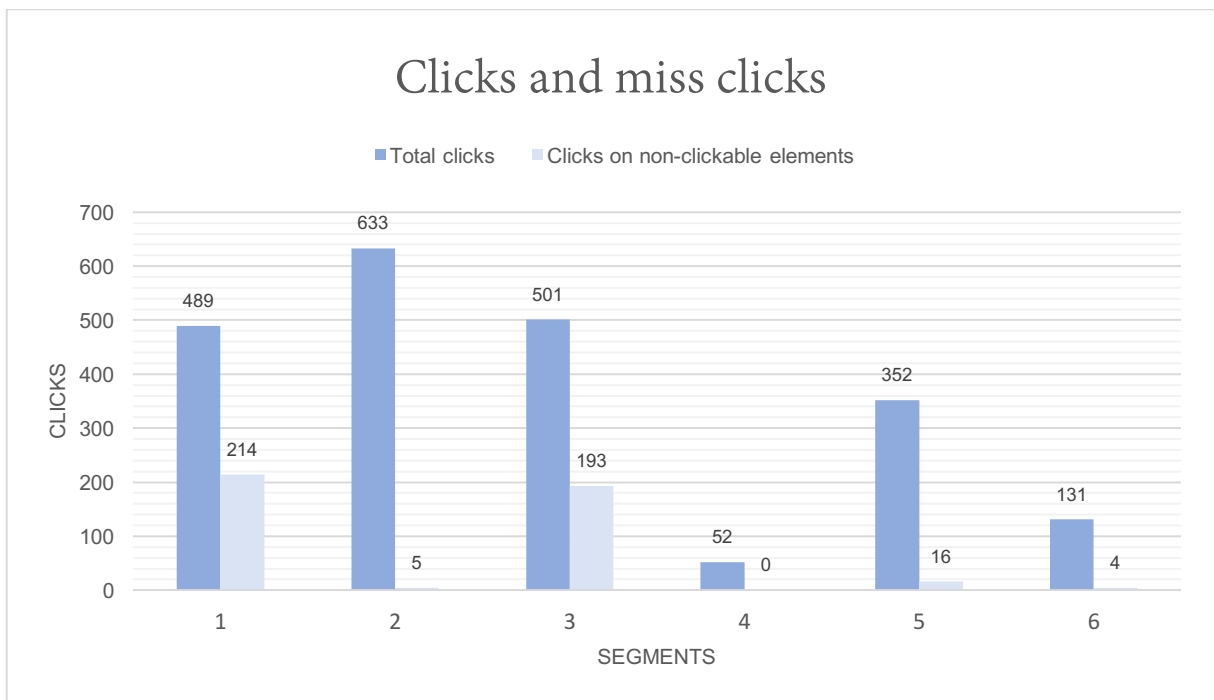


Figure 3: difference between total clicks made per segment, compared to the amount of clicks on non-clickable elements of the interface.

SEGMENT	PROBLEM	SUBJECTS	MAGNITUDE	SEVERITY
1. 1 (Locating Order Environment)	(Main Home) Subjects do not scroll down the home page and therefore miss the target	37	Local	2
2. 1 (Locating Order Environment)	The delivery environment is expected to be found under the navigational item 'Menu', instead of 'Delivery'	30	Global	3
3. 1 (Locating Order Environment)	(Main Menu) Subjects are under the impression they need to inspect the restaurants menu first	28	Local	3
4. 1 (Locating Order Environment)	(Main Menu) Subjects assume the all-you-can-eat PDF menu's are also applicable to their online orders	29	Local	2
5. 1 (Locating Order Environment)	(Main PDF documents) Subjects expect the PDF files (menu's) to contain hyperlinks.	14	Local	2
6. 1 (Locating Order Environment)	(Main Delivery) Subjects expect the zip code listing to be clickable (linking to the order environment)	48	Local	2
6. 1 (Locating Order Environment)	(Main Delivery) subjects do not recognize the SUMO Express logo as a target	48	Global	4
7. 1 (Locating Order Environment)	(Main Reserveren & contact) Subjects expect restaurant establishment address lines and images to be hyperlinks	5	Local	2
8. 2 (Selecting dishes)	(Order Home) subjects repeatedly click on the exemplary zip code in the form field, try to select it or press the 'Delete'-key after doing this.	15	Local	2
9. 2 (Selecting dishes)	(Order Home) although the interface suggest using the zipcode form field as an entrance to the product pages, subject use a different route there.	12	Local	1
10. 2 (Selecting dishes)	If the subjects' zip code is not specified on the Order Home page, the first selected dish will be discarded by the interface, causing them to have to add it again	12	Local	1
11. 2 (Selecting dishes)	Participants visit a remarkable high number of product pages looking for dishes, considering they received a clear task.	10	Global	2
12. 2 (Selecting dishes)	The difference between Order Small sushi sets and Order Sushi menu's was unclear to subjects	18	Local	2
13. 3 (Selecting dishes)	It was impossible to place an order since there was no visible check-out button	27	Local	4
14. 3 (Selecting dishes)	Participants did not understand the absence of the check-out button was a result of their order amount being to low.	12	Local	4

15.	3 (Selecting dishes)	It took participants a very long time to understand that the absence of the check-out button was a result of their order amount being to low.	15	Local	3
16.	3 (Selecting dishes)	The re-use of the SUMO Express logo (which was the entrance to the delivery environment before) makes participants assume this will lead them to check-out	12	Global	3
17.	3 (Selecting dishes)	When re-entering the zip-code (on the homepage or through the order summary) the contents of the order are erased, forcing subjects to start over again	12	Global	3
18.	5 (Entering details)	Participants start filling in details in fields that are meant for returning customers	20	Local	1
19.	5 (Entering details)	The address line is automatically filled in by the system, but not until participants have actually clicked on the field (causing them to start typing)	8	Local	1

Table 6: a listing of usability problems by number, segment, number of participants who experienced it, magnitude and severity rating.

Segment 1: locating the the delivery environment

Subjects show a lot of nonlinear behavior during this segment. Based on mouse clicks it became clear that 19 out of 48 participants (39,6%) came back at least once to look at the *Main | Delivery* page again ($M_{\text{revisits}} = 1.680$, $SD_{\text{revisits}} = .885$). Moreover, a total of 33 participants revisited the *Main | Menu* page at least once again looking for the order environment ($M_{\text{revisits}} = 2.727$, $SD_{\text{revisits}} = 1.547$). Out of those 33 subjects, 16 of them even revisited this page more than three times. Also, when participants first visit the home page, all clicks are made *on*, or above the main navigation bar. Even at the end of this segment, only a few participants ($n = 11$) clicked on elements that were situated on the bottom half of the page. It becomes clear that although participants visited the goal pages multiple times, targets were not clicked until a considerable amount of time ($M = 00:02:08s$, $SD = 00:01:21s$).

Always starting the task from the *Main | Home* page, participants used the main navigation to navigate to ‘Menu’ ($n = 30$) or ‘Delivery’ ($n = 17$). One participant chose ‘Reserveren & Contact’. This means, that despite the fact that there is a menu item called ‘Delivery’, 62,5% of participants are under the impression they need to look at the menu first. The *Main | Menu* page contains four hyperlinks, linking to PDF-files containing the dinner and lunch menu for guests eating at one of the establishments. When clicked, the PDF-file opens in a new tab within the browser. Most participants ($n = 29$) clicked on one of the links at least once ($M = 1.650$, $SD = .950$). Almost half of those participants ($n = 14$) tried clicking on dishes that were pictured in these PDF-files. When participants visit the *Main | Delivery* page, a repeated amount of clicks ($n = 48$) is registered on the names of establishments in Amsterdam ($M = 3.350$, $SD = 2.239$) and the (non-clickable) zip codes that fold out from underneath ($M = 2.100$, $SD = 2.562$). Based on this behavior it is evident that participants assume the zip codes listings to be clickable and that they will lead them to the delivery environment. In addition, almost every element on this page is expected to be a hyperlink, except for the *actual* hyperlink: the SUMO Express logo.

Segment 2: choosing dishes

When subjects reach the order environment, they can proceed to picking dishes once they have clicked the zip code form field containing an exemplary zip code, and start typing in their own. However, it stood out that a portion of the subjects ($n = 15$) clicked on the zip code form field on the *Order | Home* page more than twice and tried to select the exemplary zip code or press the 'Delete' key on their first page visit. If participants decided not to use this form, they could proceed anyway by utilizing the top navigation or clicking on the header ($n = 12$). Once they tried to add their first dish, a form field popped up asking users to provide zip code details. It struck that after having done this, all 12 subjects clicked and added the same dish *again*.

On average, participants needed almost 7 product pages to select their desired dishes ($M = 6.730$, $SD = 4.703$). 10 out of 48 participants visit more than 10 pages (21%). The most revisits occur on the pages *Order | Small sushi sets* and *Order | Sushi menu's*. Also, subjects seem to go back and forth between these two pages a lot.

Segment 3: starting check-out

In 79% of cases ($n = 27$) the interface did not display a "check-out"-button, due to the fact that the total value of the order remained below the minimum order amount of €20. During this segment a lot of random clicks on non-clickable elements occur. Most clicks are centred around the title text "Your Order", the allergy information and the establishment information. A few participants even tried hitting the 'Enter'-key ($n = 5$). Moreover, subjects tried to check-out by clicking the Sumo Express logo at top of the page ($n = 24$) or clicking the zipcode link ($n = 2$). Based on clicks and page visits, however, order contents were erased when zip codes were entered a second time. This happened to 25% of the subjects ($n = 12$), of which 5 of them had to go through this loop twice.

Segment 4: reviewing order

No usability issues were found during this segment.

Segment 5: entering personal details

Only a few usability problems occurred in the fifth segment, because the form contained a validation-function (reminding users when they made mistakes or failed to provide the required details). However, 41,7% of participants ($n = 20$) clicked on the field at the top, where returning customers are required provide their email address. Halfway through, participants started noticing this and moved on to the new customer's section. Only a small portion of subjects ($n = 4$) also continued to the password field. Lastly, 8 participants started typing in a street address into the 'street' field, only to notice half way through that this field is automatically filled out by the system after it has been clicked.

Segment 6: choosing delivery time

No usability issues were found during this segment.

Discussion

Based on the analysis, the amount of unnecessary clicks appears to be a good indicator for usability problems. As expected, most problems were found in segment 1, followed by segment 3, 2, and 5. Starting with the first segment, only 35% of participants clicked 'Delivery' from the main navigation when they first arrived on the website. Considering the clarity of the task, the name of this menu item does not clearly convey its meaning. Moreover, it becomes clear that although participants did visit the goal pages, the targets were not clicked. This

indicates that the targets were either not noticed, or not interpreted as a target. As for the *Main | Home* page, there is a large probability that the target is not seen at all during the first page visits. This is supported by the fact that no clicks were registered below the main menu, probably due to the large header size. On the *Main | Delivery* page it seemed to be the other way around. Based on mouse clicks, participants expect almost every element on this page to be a hyperlink, except the SUMO Express logo in the middle of the page. This could be caused by its' strong resemblance to the restaurant logo, withholding subjects to link this logo to the delivery environment. The mixed meaning of this logo proves to be a problem in segment 3 as well, when participants are looking for a checkout-button. Although the SUMO Express logo is placed in the top left corner (indicating it will bring users back to the homepage) participants still click it. The experience they had with this logo in the first segment might have influenced their decision during segment 3.



Image 5: an illustration of the similarity between the restaurant logo (left) and the 'express' logo (right), meant for delivery options.

Furthermore, it struck that when participants were asked to provide a zip code after adding a dish to their order, afterwards they added the same dish *again*. This indicates that when zip code details are not provided beforehand and a dish is added regardless, this action is discarded. Moreover, the most revisits occur on pages *Order | Small sushi sets* and *Order | Sushi menu*'s. Also, subjects seem to go back and forth between these two pages a lot. This either means that subjects cannot find what they are looking for, or are unsure about the content it contains. Lastly it becomes clear that instead of adding more dishes to the order, participants started randomly clicking interface elements during segment 3. This indicated that participants did not understand the absence of the check-out button was a result of their order amount being too low. The way the minimum ordering amount is currently communicated, therefore, is not clear enough.

FaceReader™ analysis

Four participants were removed from the sample because of too much missing data. In these cases, FaceReader™ was unable to generate an accurate model of the face. This was due to participants leaning into the monitor, causing parts of their face to fall out of frame. The change in intensity of the six facial expressions across task segments is illustrated in figure 4. During all segments, except segment 3, sadness seemed the most expressed emotion by participants.

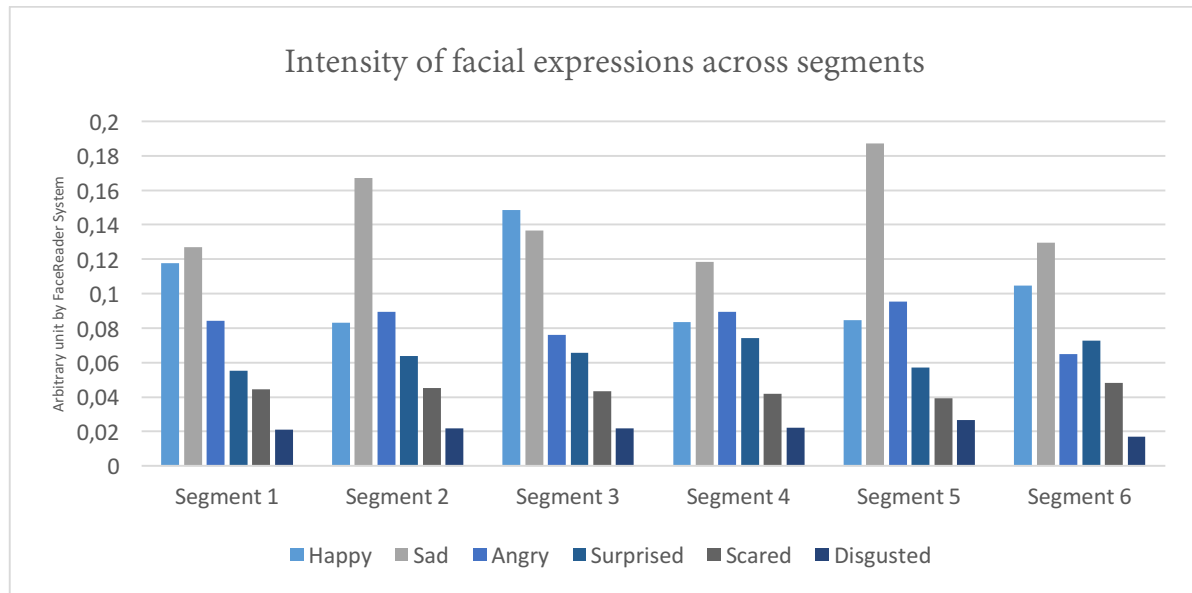


Figure 4: a visual representation of expressed emotions in facial expressions of participants ($n = 44$). 'Sad' seems to be the most dominant emotion across segments, followed by 'happy' and 'angry'.

The results of the one-way repeated-measures ANOVA showed that there was a large effect of task segment on the level of sadness in the facial expressions of participants ($F(3.317, 136.000) = 9.291, p = .000, \eta_p^2 = .185$). Bonferroni post hoc tests showed that participants expressed a significantly higher level of sad facial expressions during segment 2 ($M = 0.167, SD = .139$) and 5 ($M = 0.189, SD = .143$) compared to the other segments. Surprisingly, the overall intensity of happy facial expressions was relatively high. However, the results of a one-way repeated-measures ANOVA showed that there was a very small effect of task segment on the amount of happy facial expressions that participants disclosed ($F(3.479, 142.654) = 4.013, p = .006, \eta_p^2 = .089$). Bonferroni post hoc tests showed that participants expressed a higher level of happy facial expressions during segment 3 ($M = 0.154, SD = .146$) compared to segment 2 ($M = 0.085, SD = .098$), 4 ($M = 0.086, SD = .107$) and 5 ($M = 0.082, SD = .116$). Segment 1 and 6 did not significantly differ from each other, nor did the rest of the segments. Furthermore, although the amount of disgusted facial expressions was lowest of all emotions, the results of a one-way repeated-measures ANOVA showed that there was a large effect of task segment on the amount of disgust that participants expressed ($F(3.337, 136.834) = 8.306, p = .000, \eta_p^2 = .168$). Bonferroni post hoc tests showed that participants looked at the screen with more disgust during segment 5 ($M = 0.027, SD = .025$) compared to the other segments. Lastly, there did not seem to be a significant difference between levels of expressed emotions across segments for angry ($\eta_p^2 = .045$), surprised ($\eta_p^2 = .036$) or scared ($\eta_p^2 = .011$).

Based on negative and positive emotions, FaceReader™ computes overall valence. Based on a one-way repeated-measures ANOVA, there was a medium effect of task segment on the the overall negative valence in participants' expressions ($F(3.796, 155.616) = 5.860, p = .000, \eta_p^2 = .125$). Bonferroni post hoc tests showed that valence was significantly *less* negative in segments 1 ($M = -0.053, SD = .177$) and 3 ($M = -0.022, SD = .219$) compared to segment 2 ($M = -0.124, SD = .184$) and 5 ($M = -0.152, SD = .201$).

Discussion

The analysis showed that when using SUMO's website to make an order, the most experienced emotions were 'sad', 'happy' and 'angry'. However, not all measurements were significantly different across segment. Based on the outcomes of a repeated-measures ANOVA on task segment, the following conclusions can be drawn:

	Effect	Effect size
Segment 1	provokes less negative emotions compared to segment 2 and 5	$(\eta_p^2 = .125)$
Segment 2	provokes more sad facial expressions compared to all other segments	$(\eta_p^2 = .185)$
Segment 3	provokes more happy facial expressions compared to segment 2, 4 and 5 and provokes less negative emotions compared to segment 2 and 5	$(\eta_p^2 = .089)$ $(\eta_p^2 = .125)$
Segment 4	has no influence on facial expressions	-
Segment 5	provokes more sad facial expressions compared to all other segments and provokes more disgusted facial expressions compared to other segments	$(\eta_p^2 = .185)$ $(\eta_p^2 = .168)$
Segment 6	has no influence on facial expressions	-

Figure 5: summary of the significant effects of segments on various facial expressions during the task.

Based on the expectations described earlier in the method section, segment 1 and 3 are expected to provoke the highest levels of negative emotion. However, the measured facial expressions during the task suggest the exact opposite: valence levels are higher during the first and third segment, and the level of happy facial expressions is significantly higher during segment 3 compared to the other segments. A study by Hoque and Picard (2011) clarifies this. They found that there is a significant difference between acted vs. natural frustration in facial expressions. Besides the fact that acted frustration is much easier to detect by a computer, they also discovered that almost all individuals smile during natural frustration. It would, therefore, be possible to find increased levels of happy expressions during segment 1 and 3. It makes it impossible, however, to differentiate between actual happy facial expressions and smiles out of frustration. Moreover, participants exhibited significantly higher levels of sad facial expressions during segment 2 (selecting dishes) and segment 5 (entering personal details). This suggests that although participants might have been most frustrated during segment 1 and 3, they disliked the process of selecting dishes and entering personal details the most. Lastly, although levels of disgust were moderate, participants experienced significantly more disgust during segment 5 (entering details) compared to the other segments. This implies that participants particularly dislike the process of entering their personal details. So while the negative emotions in segment 1 and 3 are mostly related to usability problems, the negative emotions in segment 2 and 5 are mostly related to the overall experience.

EEG analysis

Results

Four participants were removed from the sample because of software malfunction. After the data was analysed, the average computed differences in attention, meditation and zone levels per task segment turned out to be very minimal. As illustrated in figure 6, participants seemed most concentrated in segment 2 ($M = 47.593$, $SD = 9.545$) and segment 4 ($M = 47.820$, $SD = 14.610$). Participants were least calm during in segment 3 ($M = 52.596$, $SD = 9.611$) and segment 6 ($M = 52,342$, $SD = 13.632$). The results of the one-way repeated-measures ANOVA showed, however, that there was no significant effect of task segment on the average levels of attention ($\eta_p^2 = .036$), meditation ($\eta_p^2 = .009$) nor zone ($\eta_p^2 = .037$).

Estimated marginal means of computed EEG signals

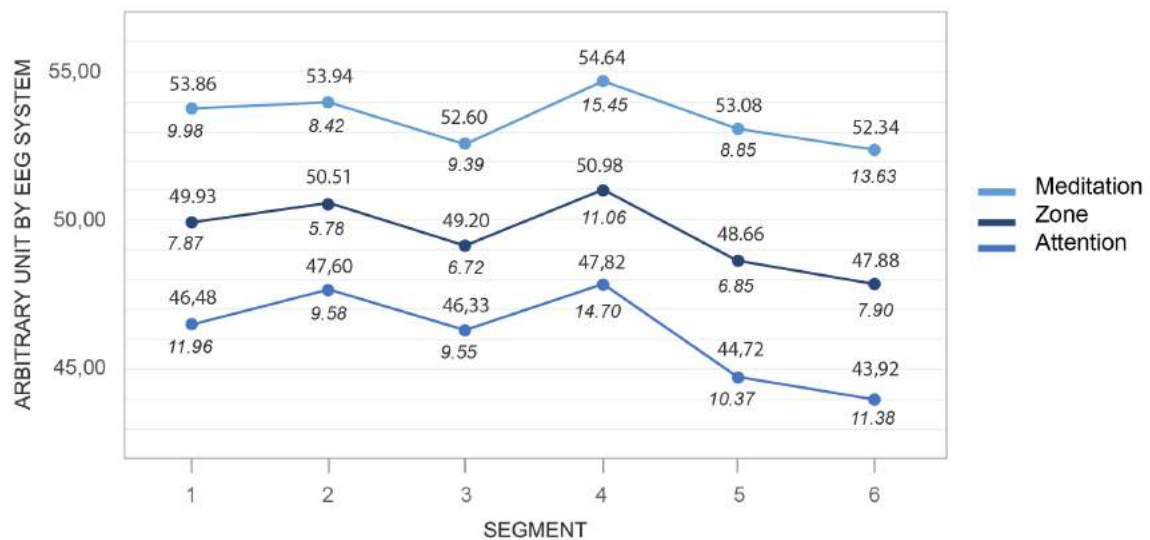


Figure 6: Visual representation of the computed difference between Meditation, Zone and Attention levels. The mean of each segment is situated above the data point and the standard deviation is situated underneath.

The average difference in amplitude of low- and high beta waves per task segment were very minimal as well. As illustrated in figure 7, the biggest amplitude differences can be found in segment 3 ($M = -3.812$, $SD = 4.472$) and 5 ($M = -2.309$, $SD = 3.005$).

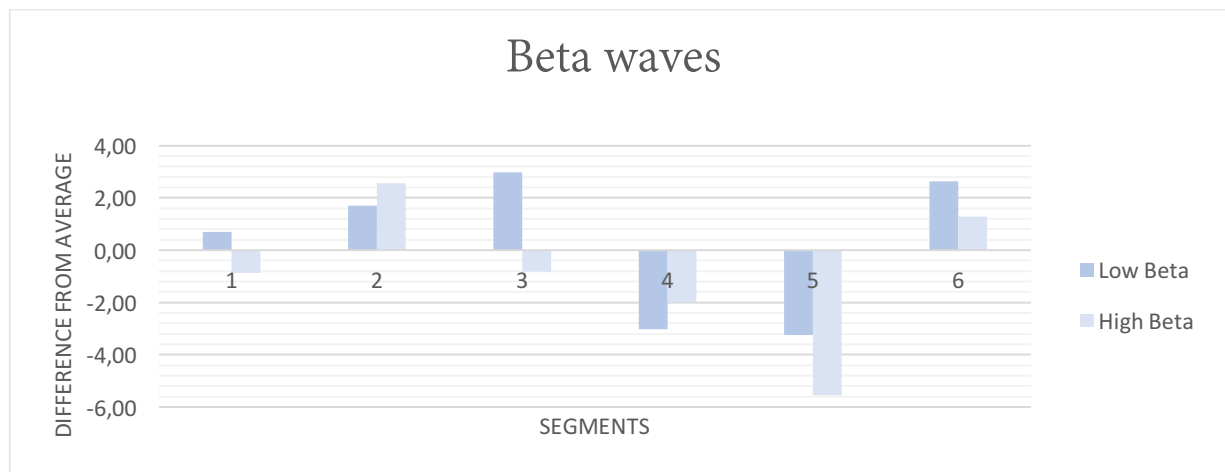


Figure 7: Visual representation of the difference from average for all participants ($n = 44$) in low- and high Beta waves per segment.

Nevertheless, the results of the one-way repeated-measures ANOVA showed that there was no effect of task segment on the average amplitude differences between low and high beta waves ($\eta_p^2 = .017$).

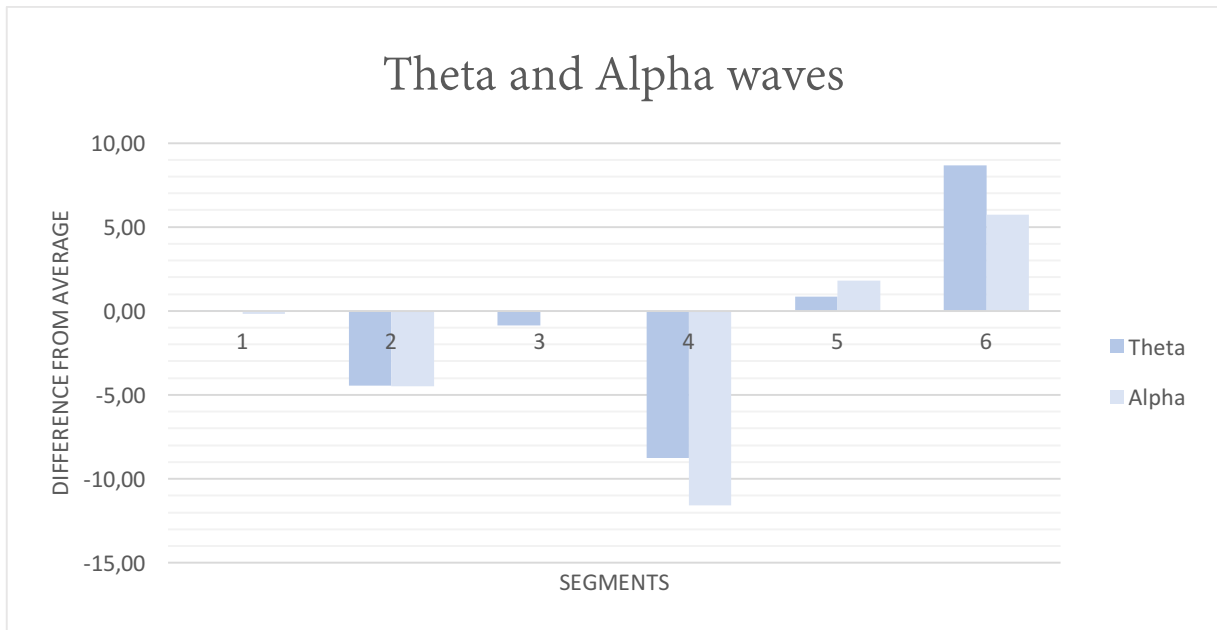


Figure 8: Visual representation of the difference from average for all participants ($n = 44$) in Theta and Alpha waves per segment.

The average amplitudes of alpha waves were lowest in segment 2 ($M = -4,482$, $SD = 13,424$) and 4 ($M = -11,599$, $SD = 54,434$) and highest in segment 6 ($M = 5,735$, $SD = 48,407$). The amplitude of theta waves were also lowest during segment 2 ($M = -4,465$, $SD = 12,544$) and 4 ($M = -8,773$, $SD = 50,338$) as well, and highest during segment 6 ($M = 8,687$, $SD = 56,413$). The results of the one-way repeated-measures ANOVA showed that there was no effect of task segment on the average levels of alpha ($\eta_p^2 = .030$) of theta waves ($\eta_p^2 = .026$).

Discussion

Due to minimal differences across task segments and the lack of a significant effect of segments on concentration and stress-levels, it is not possible to draw conclusions from this data set in regard to usability problems. One of the causes could be the reliability of the Neurosky MindWave Mobile headset, which is doubtful. While laboratory tests use EEG systems with about 20 to 200 electrodes to capture and amplify the signal, the mobile headset only has *one* electrode. This means the data will contain a lot of noise, making it harder to differentiate between segments. Also, because nature of the tasks is very similar, EEG does not appear to be a very suitable method to measure usability or identify usability problems.

Questionnaire analysis

Based on the experience of 48 subjects within this data sample, SUMO's order process was graded with an average SUS-score of 34 ($M = 33.56$, $SD = 15.04$). The overall degree of unpleasantness across task segments appeared to be greater than the degree of experienced difficulty. This is made visible in figure 9. However, a repeated-measures ANOVA on difficulty and unpleasantness proved that there was no significant difference between difficulty and unpleasantness within task segments.

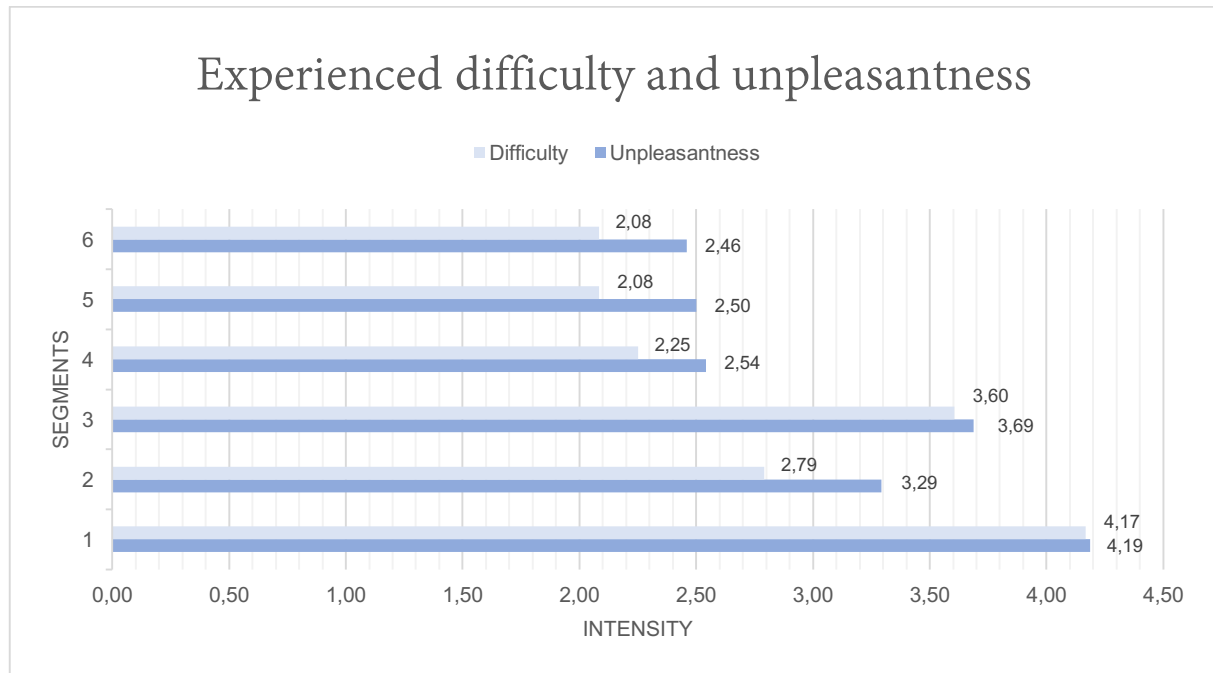


Figure 9: Visual representation of the estimated marginal means of experienced difficulty and unpleasantness across task segments ($n = 48$).

Difficulty

A repeated-measures ANOVA on task segments proved that segments had a very large influence on the degree of experienced difficulty ($F(3.221, 151.406) = 55,434$, $p = .000$, $\eta_p^2 = .541$). Bonferroni post hoc tests showed that participants found segment 1 ($M = 4.166$, $SD = 0.781$) and segment 3 ($M = 3.604$, $SD = 1.250$) significantly more difficult than the other segments, but 1 and 3 did not significantly differ from each other. Furthermore, participants found segment 2 ($M = 2,792$, $SD = 1,071$) significantly more difficult than segment 5 ($M = 2,083$, $SD = 0,821$) and 6 ($M = 2,083$, $SD = 0,739$). Segment 4, 5 and 6 did not significantly differ from each other.

Unpleasantness

A repeated-measures ANOVA on task segments proved that segments also had a very large influence on the degree of experienced unpleasantness ($F(3.488, 163.939) = 42,478$, $p = .000$, $\eta_p^2 = .475$). Bonferroni post hoc tests showed that participants found segment 1 ($M = 4.188$, $SD = 0.816$), segment 2 ($M = 3.292$, $SD = 1.010$) and segment 3 ($M = 3.688$, $SD = 1.114$) significantly more unpleasant than the other segments. Segment 1, 2 and 3, however, did not significantly differ from each other, nor did segment 4, 5 and 6.

Influence of difficulty and unpleasantness on SUS scores

A Shapiro-Wilk's test ($p > 0.05$) (Shapiro & Wilk, 1965; Razali & Wah, 2011) and a visual inspection of the histogram, normal Q-Q plot and box plot showed that the SUS-scores were approximately normally

distributed for almost all participants with a skewness of .253 ($SE = .354$) and a kurtosis of -.212 ($SE = .695$). Three outliers (P07, P44, P45) were removed from the sample. A multiple regression was run to investigate the influence of the experienced difficulty and unpleasantness of the task on SUS-score. The regression model with the SUS-score as dependent variable and difficulty and pleasantness per task segment as independent variable predicts the SUS-score significantly, $F(12,32) = 3.781$, $p = 0.001$, $R^2 = 0.586$. This means the model explains 59% of the variation in the SUS-score. Experienced difficulty in segment 1, 2 and 4 added statistically significantly to the prediction, $p < .05$.

Model		B	SE B	β	t-value
Difficulty	Segment 1: 'Locating order environment'	-9.730	2.964	-.582	-3.283*
	Segment 2: 'Selecting dishes'	-4.214	1.654	-.374	-2.548*
	Segment 3: 'Starting checkout'	-3.693	2.210	-.385	-1.672
	Segment 4: 'Reviewing order'	7.378	3.448	.522	2.140*
	Segment 5: 'Entering details'	-6.507	4.530	-.454	-1.436
	Segment 6: 'Choosing delivery time'	7.776	4.100	.487	1.897
Unpleasantness	Segment 1: 'Locating order environment'	4.258	3.072	.246	1.386
	Segment 2: 'Selecting dishes'	-.160	1.766	-.013	-.091
	Segment 3: 'Starting checkout'	-1.175	2.637	-.108	-.445
	Segment 4: 'Reviewing order'	-6.305	4.030	-.365	-1.565
	Segment 5: 'Entering details'	-3.450	3.457	-.266	-.998
	Segment 6: 'Choosing delivery time'	1.968	2.840	.148	.693

Table 7: Results of the multiple regression analysis of difficulty and unpleasantness on SUS-scores ($p < 0.05$).

Comments

One of the most discussed problems by participants in regard to the last question of the questionnaire ('What could have made this experience better for you?') was the trouble they experienced trying to find the delivery environment in the first segment. For instance, one subjects mentioned: "The link to the delivery environment is tucked away so far, it's ridiculous. It almost makes you think that SUMO does not want you to order from them." A couple of other subjects also made it clear that if this would have been a real situation, SUMO would have lost them as a customer: "Even just *finding* the delivery environment was hard! I would have left this website half way through and order sushi from a competitor." There was one participant who clarified *why* she experienced trouble locating the delivery environment. She said: "I was intuitively searching for something like 'Bestel hier' instead of 'Delivery'." Furthermore, subjects also expressed their displeasure about the checkout-button not being visible: "I didn't see that the minimum ordering amount was €20 anywhere." Lastly, there was one subject also expressing his discontent with segment 5: 'entering details': "This process is too complicated and too long. Nobody want to go through so many steps just to place an order. Especially having to fill in so many personal details makes me never want to return to this website."

Based on the answers to this question, it was possible to identify 9 general usability problems described in table 8.

	SEGMENT	PROBLEM	SUBJECTS	MAGNITUDE	SEVERITY
1.	1 (Locating Order Environment)	The menu items in the top navigation are unclear to participants.	3	Global	2
2.	1 (Locating Order Environment)	The meaning and purpose of the PDF-menu's is unclear to subjects	1	Local	2
3.	1 (Locating Order Environment)	Participants have a hard time finding the delivery environment because the website does not 'lead' them to it.	15	Global	3
4.	1 (Locating Order Environment)	The functionality of the shopping basket seemed unclear.	1	Local	1
5.	2 (Selecting Dishes)	It's unclear what type of dishes are to be found on the various product pages.	1	Global	2
6.	2 (Selecting Dishes)	Subjects feel like they lack guidance, helping them navigate through the order process.	2	Local	2
7.	3 (Starting Checkout)	Participants do not see the minimum order amount, because it is not clearly visible or communicated with the user	7	Local	3
8.	3 (Starting Checkout)	The absence of the checkout-button causes a lot of frustration and confusion among subjects.	5	Local	4
9.	5 (Entering details)	The website contains too many form fields, needlessly lengthening the order process.	1	Local	2

Table 8: a listing of usability problems found through the questionnaire, categorized by number, segment, number of participants who experienced it, magnitude and severity rating.

Discussion

Based on (Bangor, Kortum, & Miller, 2008) SUS scores below 70 need to be considered insufficient. Since SUMO's average SUS score is 34, it can be expected that the overall usability of the order process is poor. Based on the multiple regression analysis this score was mostly effected (59%) by the high values of experienced difficulty in segment 1 and 2 and lower values of experienced difficulty in segment 4. This is confirmed by the answers to the open question regarding the overall experience, where most comments address problems in the first three segments. However, although these comments provide enough information to pinpoint a couple usability problems, it becomes evident that this method is best used in combination with other usability research methods.

Qualitative analysis

First impressions

For the traditional qualitative research, 8 subjects were observed and questioned. The first impressions on SUMO's website appeared to be positive. The images that rotated on the page header were appealing and attracted attention. However, one subject pointed out that the changing imagery made the website feel cramped. According to participants, the images that were used across the website also portray various different atmospheres. As a result, they leave participants wondering about what kind of restaurant SUMO is supposed to be. Some images make you think it is a pub or café, while others give the impression that is a chic restaurant. Moreover, a number of participants are able to navigate to the order environment directly through the alternating images on the homepage. In this case they saw the text "Order via sumossushiexpress.com. Now 5% discount & no delivery fee!" and realized they could place their order by clicking it. Other participants expressed that the images were changing too quickly, and were therefore unable to properly process the information.

The main navigation

It became clear that visitors reached the order environment through different routes, explained in the following two situations:

- **Situation A:** participants reach the order environment by chance because they clicked the alternating header image. This route causes less frustration.
- **Situation B:** participants search the main navigation. Before they finally reach the order environment, they have already clicked several menu items and visited multiple pages (menu, restaurants, delivery or reservations). This leaves them looking for the order environment for a relatively long period of time, causing frustration among some of the participants.

Within situation B, subjects have a tendency to click on 'menu', instead of 'delivery'. This is a result of expecting to immediately having to view and select dishes. The first encountered problem is that not everyone understands what the meaning of 'Dinner inside' and 'Dinner outside' is supposed to be. When participants click on any of these links, most of them recognize the document as a PDF-file. However, some participants do not notice this and actually try to click on dishes they would like to order.



When subjects do click on ‘delivery’, the addresses and zip codes of the establishments are the first thing that catch the eye [01]. Thereafter, subjects try to click on the zip codes, thinking this will be the hyperlink to the delivery environment. What struck, was that the SUMO Express logo was not associated with a button or hyperlink at all. The texts ‘Sumo Sushi Express’ and ‘Order now’ do not attract much attention either [02]. Only with help from the observer do participants eventually click the logo, leading them to the order environment. Moreover, participants expect they can click on a specific restaurant on the restaurant page to place their order. When they subsequently only find opening hours, prices and an address, subjects do not look around further and leave the page.

The order environment

When participants have reached the order environment eventually, they expect to start ordering right away. It seems unclear to them why they are faced with another page *before* the actual order environment. Furthermore, to some participants it remains unclear in which categories they can find certain dishes [01]. For example, ‘maki rolls’ are found fairly easy, but finding the ‘nigiri shrimp’ appears to be more difficult. Participants pointed out that a combination of texts and imagery might help to circumvent this. When the desired dishes were found and added to subjects’ shopping baskets, they received a pop-up, asking for a zip code to check whether or not SUMO delivered in their area [02]. Participants thought this appeared too late and preferred to receive these kinds of pop-ups a lot earlier in the order process.



Image 7: a screenshot of the Main | Delivery page

Lastly, most subjects did not notice their order total or the minimum order amount listed in the summary on the right [03]. For some, this led to confusion and frustration. Only by chance or with help from the observer did participants notice their orders should be over 20 euros. Participants would like to be informed about this sooner. In addition, a number of participants pointed out that for customers who are ordering solely for themselves, a minimum order amount of 20 euros is too high. Some did not think of this as very customer-friendly and pointed out that if this would have been a real situation, they would not have placed the order. All observations mentioned above resulted in the following 15 usability problems:

SEGMENT	PROBLEM	SUBJECTS	MAGNITUDE	SEVERITY	
1.	1	Participants find the main navigation ambiguous.	6	Global	3

	(Locating Order Environment)				
2.	1 (Locating Order Environment)	Participants experience the meanings of menu names (inside/outside) as unclear.	5	Local	2
3.	1 (Locating Order Environment)	The menu being a PDF document causes confusion	3	Local	1
4.	1 (Locating Order Environment)	Products on the PDF menu are not clickable, causing participants to get stuck in the order process.	5	Local	3
5.	1 (Locating Order Environment)	Addresses and zip codes draw away the attention from the order-button.	6	Local	3
6.	1 (Locating Order Environment)	Participants wonder that the use of the zip code listings are.	6	Local	2
7.	1 (Locating Order Environment)	Participants do not recognize the Sumo Express logo as an order-button or link to the delivery environment.	4	Global	3
8.	1 (Locating Order Environment)	Participants expect to choose a specific establishment on the <i>Main Restaurants</i> page and place an order there.	2	Local	2
9.	2 (Selecting Dishes)	Participants feel like they receive very little cues or assistance during the order process. As a result, they would place an order with the competitor or try to reach SUMO by phone.	7	Global	4
10.	2 (Selecting Dishes)	Participants feel the need to receive more feedback from the system, because it is unclear to them when actions have been successful or unsuccessful.	5	Global	1
11.	2 (Selecting Dishes)	Participants do not expect to find another page <i>before</i> the actual delivery environment.	1	Local	1
12.	2 (Selecting Dishes)	It remains unclear to participants in which categories they can find certain dishes. Finding 'nigiri' appears to be hard.	7	Local	3
13.	2 (Selecting Dishes)	The pop-up, asking for a zip code to check whether or not SUMO delivers in a certain area, appears too late, according to participants.	2	Local	1
14.	3 (Starting Checkout)	The order total in the summary is not noticed.	5	Local	2
15.	3 (Starting Checkout)	'Minimum order amount' is only noticed by chance or with help from the observer.	5	Local	3

Table 9: a listing of usability problems found through the questionnaire, categorized by number, segment, number of participants who experienced it, magnitude and severity rating.

Meta-analysis

Descriptive comparison of usability problems

Together, the quantitative and qualitative research methods pinpointed 63 usability problems, of which 36 were unique. It became clear that most usability issues occurred during the first three segments. First of all, out of 19 usability issues that eye tracking was able to point out, 8 of them were unique for this research method. Second, out of the 20 problems found by clickstream analysis, 7 of them were not found using any of the other methods. Moreover, out of 15 problems found by qualitative research, 4 problems could not be pinpointed by other research methods that were used. Lastly, one out of 9 problems based on the questionnaire comments was unique for this method. Unfortunately, FaceReader™ and EEG did not provide sufficient data to identify separate usability problems and are therefore not included in this summary.

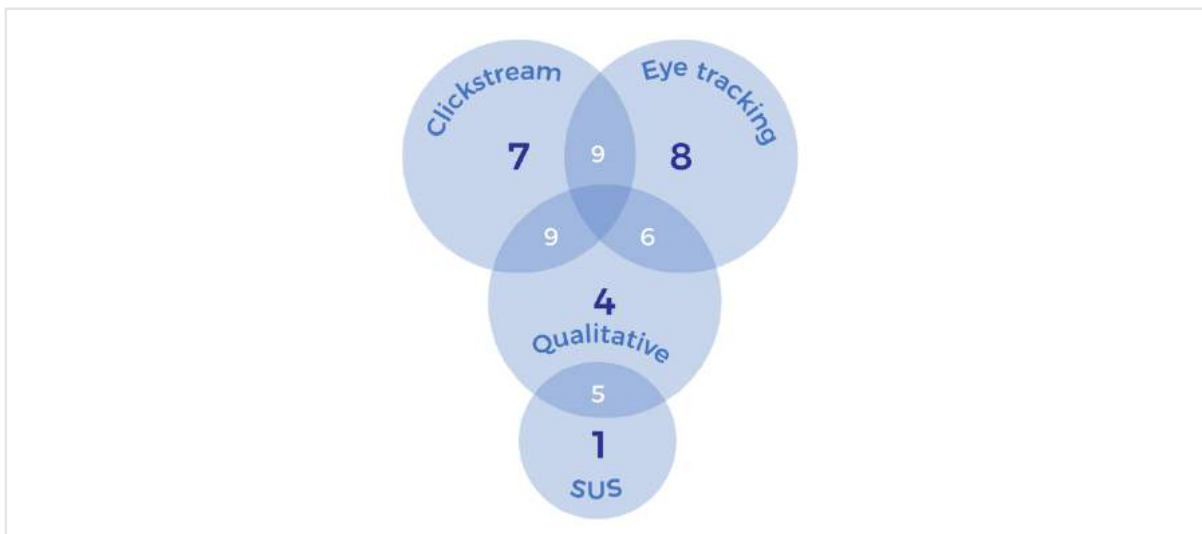


Figure 10: number of usability problems found per research method. Dark blue numbers indicating problems found solely by that method alone, white numbers indicating overlapping problems.

Correlational comparison of methods

A Spearman's rank-order correlation with a significance level of 5% was run to determine the relationship between all variables, separated by segment, across methods. No correlations based on a cut-off point lower than $\alpha = 0.05$ were examined in this analysis. Also, solely correlations above $r_s = .450$ or below $r_s = -.450$ will be discussed. A complete overview of the correlation matrixes per segment can be found in the appendix.

SEGMENT 1			
Variable A	Variable B	Correlation	p-value
Attention	Task Time	$r_s(30) = -.499$	$p = .005$
unnecessary clicks	Unpleasantness	$r_s(30) = .584$	$p = .001$
unnecessary clicks	Task time	$r_s(30) = .910$	$p = .000$
unnecessary pages	Difficulty	$r_s(30) = .523$	$p = .003$
unnecessary pages	Unpleasantness	$r_s(30) = .763$	$p = .000$
unnecessary pages	Task time	$r_s(30) = .920$	$p = .000$
Arousal	Unpleasantness	$r_s(30) = .504$	$p = .005$
Difficulty	Task time	$r_s(30) = .504$	$p = .005$
Unpleasantness	Task time	$r_s(30) = .784$	$p = .000$

Table 10: a summary of significant correlations above $r_s = .450$ or below $r_s = -.450$ in segment 1: 'locating delivery environment'.

During the first segment, ‘locating delivery environment’, there was a moderate, negative correlation between attention (EEG) and task time. This means that when participants were highly concentrated, they also spent less time trying to find the delivery environment. Furthermore, analysis showed there was a moderate to strong positive correlation between unnecessary clicks, unnecessary pages, difficulty and unpleasantness in relation to task time. On their turn, unnecessary clicks were positively correlated with unpleasantness and unnecessary pages were correlated with difficulty. This means that when participants needed more clicks to find the delivery environment, they also found this experience less pleasant. Moreover, participants who visited a higher level of pages during the first segment also found it significantly more difficult. During this segment, arousal positively correlated with unpleasantness as well, meaning that when participants showed more arousal in their facial expressions, this part of the task was experienced as less pleasant.

<i>SEGMENT 2</i>			
Variable A	Variable B	Correlation	p-value
Alpha waves	Disgusted	$r_s(30) = -.546$	$p = .002$
Extra clicks	Fixation duration ‘Sushi’	$r_s(30) = .496$	$p = .009$
Extra clicks	Scared	$r_s(30) = .503$	$p = .005$
Extra clicks	Task time	$r_s(30) = .717$	$p = .000$
Extra pages	Fixation duration ‘Sushi’	$r_s(30) = .566$	$p = .001$
Extra pages	Scared	$r_s(30) = .497$	$p = .005$
Extra pages	Task time	$r_s(30) = .774$	$p = .000$

Table 11: a summary of significant correlations above $r_s = .450$ or below $r_s = -.450$ in segment 2: ‘selecting dishes’.

In the second segment, ‘selecting dishes’, the amount of unnecessary clicks and pages were positively correlated with task time again. More striking was the negative correlation between the amplitude of alpha waves (EEG) and disgusted facial expressions (FaceReader™). What this means, is that when participants’ brainwaves showed higher alpha levels, they also expressed less disgust when they were selecting dishes. Furthermore, the amount of unnecessary clicks and pages were positively correlated with the average fixation duration on the menu item ‘sushi’. This indicates that when participants had more trouble interpreting this target, they also exhibited more clicks and page visits while selecting dishes. This also seemed to be the case with the amount of unnecessary clicks and pages in relation to scared facial expressions.

<i>SEGMENT 3</i>			
Variable A	Variable B	Correlation	p-value
Clicks on non clickable	Task time	$r_s(30) = .515$	$p = .004$
Extra clicks	Task time	$r_s(30) = .821$	$p = .000$
Extra pages	Task time	$r_s(30) = .686$	$p = .000$
Fixation count	Task time	$r_s(30) = .575$	$p = .001$
Surprised	Difficulty	$r_s(30) = -.499$	$p = .005$
Surprised	Unpleasantness	$r_s(30) = -.493$	$p = .006$
Difficulty	Task time	$r_s(30) = .544$	$p = .002$
Unpleasantness	Task time	$r_s(30) = .484$	$p = .007$

Table 12: a summary of significant correlations above $r_s = .450$ or below $r_s = -.450$ in segment 3: ‘starting checkout’.

During the third segment, ‘starting checkout’, the variables unnecessary clicks, unnecessary page visits,

amount of clicks on non-clickable elements, fixation count on ‘minimal order amount’ and experienced difficulty and unpleasantness were all positively correlated with task time. Also, the expression surprised (FaceReader™) was negatively correlated with difficulty and unpleasantness during this segment. This indicates that participants who showed more surprise in their facial expressions, experienced the search for the checkout-button as less difficult and unpleasant.

During the fourth segment, ‘reviewing order’, only one correlation was found. Logically, the time between first fixation and mouse click on the ‘Ga verder’-button was strongly positively correlated with task time ($r_s(30) = .641, p = .000$), meaning that when participants took longer to click this button, task time was also higher. In segment five, ‘entering details’, it appeared that meditation (EEG) was negatively correlated with difficulty ($r_s(30) = -.468, p = .009$). Therefore, participants who were more calm during this segment also found it less difficult. Moreover, alpha waves seemed negatively correlated with the amount of clicks made on non-clickable interface elements ($r_s(30) = -.467, p = .009$). This would indicate that participants with higher amplitudes of alpha waves also made less wrongly placed clicks.

In the final segment, ‘selecting delivery time’, there was a moderately negative correlation between attention (EEG) and (the by FaceReader™ computed value) valence ($r_s(28) = -.490, p = .008$). It suggests that participants who experienced higher levels of attention had lower levels of valence, indicating that higher concentration had a negative effect on facial expressions.

<i>OVERALL</i>			
Variable A	Variable B	Correlation	p-value
Clicks on non clickable	SUS	$r_s(30) = -.781$	$p = .000$
Clicks on non clickable	Task time	$r_s(30) = .687$	$p = .000$
Extra clicks	Scared	$r_s(30) = .523$	$p = .008$
Extra clicks	Task time	$r_s(30) = .696$	$p = .000$
Extra pages	SUS	$r_s(30) = -.465$	$p = .010$
Extra pages	Task time	$r_s(30) = .741$	$p = .000$
SUS	Difficulty	$r_s(30) = -.500$	$p = .005$

Table 13: summary of significant correlations above $r_s = .450$ or below $r_s = -.450$ across the entire task

When one looks at the task as a whole, there seems to be a strong positive correlation between the percentage of clicks on non-clickable elements, unnecessary clicks and unnecessary pages in relation to task time. More interestingly, three variables seem tightly coupled: the percentage of clicks on non-clickable elements shows a strong negative correlation with SUS score. The amount of unnecessary visited pages is negatively correlated with the SUS-score as well. And on his turn, SUS score is negatively correlated with experienced overall difficulty of the task.

Discussion

Between all correlations that were found, the only variable that showed consistent correlations across segments was task time in relation to unnecessary clicks, unnecessary pages, difficulty and unpleasantness and SUS-score. This is rather logical, since extra clicks and page visits indicate inefficiency. This confirms that one of the most well known usability metrics, *time on task*, is a good predictor for usability issues.

General discussion

One of the most common ways to perform usability research is by direct observation and questioning. However, there are various psychological and social factors that can influence participants' behavior when usability research contains explicit self-reporting. It would therefore be very useful to circumvent the subjectivity of traditional usability research by measuring actual behavior, instead of relying on the verbal report of participants. The purpose of this research was to determine whether or not quantitative research methods give different, or better insights into the usability of a product than the traditional methods. The main research question was therefore: “*Are quantitative testing methods (such as eye tracking, EEG, facial expression analysis, mouse tracking and the system usability scale) as effective in identifying usability problems as qualitative research with observation?*” It was expected that mainly eye tracking in combination with mouse metrics could identify a larger amount of (more detailed) usability problems compared to traditional qualitative research. Moreover, since usability problems are known to cause frustration and anger, participants' facial expressions were expected to show this. Furthermore, usability problems were expected to be related to higher levels of stress and concentration. Lastly, it was expected that the outcome of the System Usability scale and supplementary questions would provide a concise summary of task segments in which most usability problems occurred.

The results of the meta-analysis showed that out of 19 usability issues that were discovered by eye tracking analysis, 8 of them were unique for this research method. Moreover, clickstream analysis identified 20 usability issues of which 7 were unique for this research method. Out of the 15 usability issues detected by qualitative research, 4 issues were unique for this research method. In summary, when clickstream analysis and eye tracking analysis are bundled together, these quantitative methods exposed 15 usability problems that the other research methods did not, compared to 4 usability problems by qualitative analysis. Furthermore, FaceReader™ and EEG measurements did not provide sufficient data to identify separate usability problems or significantly differentiate between task segments. Based on the correlational comparison of methods, there seems to be a strong positive correlation between the percentage of clicks on non-clickable elements, unnecessary clicks and unnecessary pages in relation to task time. Also, three variables appear related: the percentage of clicks on non-clickable elements (clickstream) shows a strong negative correlation with SUS score. The amount of unnecessary visited pages is negatively correlated with the SUS-score as well. Subsequently, SUS score is negatively correlated with experienced difficulty of the task.

When we look at how one identical usability problem is explained by quantitative and qualitative methods, it becomes clear how different the insights are. Table 13 provides an example of three insights based on the same usability problem. Based on qualitative analysis, participants *did not notice* the target, whereas eye tracking analysis showed that the target was *seen*, but not interpreted as such. Subsequently, clickstream analysis showed that participants *did not understand* the target was even clickable.

Eye tracking analysis	Clickstream analysis	Qualitative analysis
Subjects fixated on the Sumo Express logo several times, but did not interpret this as a button.	Subjects did not understand the SUMO Express logo was a button, because they clicked everywhere on the page, except on the logo.	Participants had a hard time finding the delivery environment because they did not notice the Sumo Express logo.

Table 13: a comparison of insights across three research methods, based on the same usability problem.

Another interesting situation was where the qualitative analysis described an observation in which half of the participants (50%) utilized the homepage header to enter the delivery environment. This was striking, because in the quantitative data sample only 4 out of 48 participants (8%) followed this route. This is a clear example of the bias that can be created by qualitative research with smaller samples. On the other hand, qualitative analysis can reveal the *reason* behind a usability problem: something about which quantitative methods can only speculate. Table 14 shows that clickstream analysis could only determine that participants had trouble finding the desired dishes. Eye tracking analysis suspected the problem was caused by solely one menu item: ‘sushi’. Qualitative analysis actually clarified that most menu items were vague and ambiguous, making it especially hard for participants to find a dish called ‘nigiri’.

Eye tracking analysis	Clickstream analysis	Qualitative analysis
The average fixation duration on the menu item ‘Sushi’ was higher compared to other items, indicating trouble interpreting it.	Participants visited a remarkable high number of product pages looking for dishes, while they only needed 2 types of sushi.	It remained unclear to participants in which categories they can find certain dishes. Especially ‘nigiri’ was hard to find because of the ambiguous navigational item names.

Table 14: a comparison of insights across three research methods, based on the same usability problem.

In addition, qualitative research provides other insights that cannot be collected from quantitative data, such as: “Participants pointed out that a minimum order amount of 20 euros is too high for customers who are ordering solely for themselves. They did not find this very customer-friendly and pointed out that if this was a real scenario, they would not have placed an order.”

Ultimately, the expectation that eye tracking in combination with mouse metrics will have the ability to identify a higher amount of usability problems than traditional qualitative research, proved to be true. Because of the precise measurement of participant behavior, usability problems are more detailed and specified. Moreover, it became clear that a simple usability task such as ordering sushi online, did not provoke enough different emotions and stress levels to be used for data analysis. Furthermore, insight into stress and frustration levels seemed very useful to usability research in *theory*, but mostly just confirmed the usability problems that were already identified by other methods. Also, no obvious cohesion between variables across quantitative methods was found (aside from task time, SUS-scores and unnecessary clicks and page visits). Lastly, the outcome of the System Usability scale and supplementary questions have indeed provided a concise summary of parts of the task in which most usability problems occurred.

In conclusion, it becomes clear that the use of eye tracking in combination with mouse metrics has the ability to identify a higher amount of specific and detailed usability problems than traditional qualitative research. Nevertheless, using a quantitative approach is considerably more labor intensive in terms of data processing. For this reason, this approach is not recommended when research questions are still explorative or broadly orientated, such as: 'Does this website contain usability issues?' Results from qualitative research can, however, provide an excellent starting point for further in-depth research using eye tracking and clickstream analysis.

References

- Ashby, F. G., & Isen, A. M. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological review*, 106(3), 529-550.
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human-Computer Interaction*, 24(6), 574-594.
- Biehal, G., & Chakravarti, D. (1989). The Effects of Concurrent Verbalization on Choice Processing. *Journal of Marketing Research*, 26(1), 84-96.
- Bojko, A. (2006). Using eye tracking to compare web page designs: A case study. *Journal of Usability Studies*, 1(3), 112-120.
- Bolte, A., Goschke, T., & Kuhl, J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological Science*, 14(5), 416-421.
- Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.
- Campbell, A., Choudhury, T., Hu, S., Lu, H., Mukerjee, M. K., Rabbi, M., & Raizada, R. D. (2010, August). NeuroPhone: brain-mobile phone interface using a wireless EEG headset. In *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds* (pp. 3-8).
- Çöltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the effectiveness of interactive map interface designs: a case study integrating usability metrics with eye-movement analysis. *Cartography and Geographic Information Science*, 36(1), 5-17.
- Crowley, K., Sliney, A., Pitt, I., & Murphy, D. (2010, July). Evaluating a brain-computer interface to categorise human emotional response. In *Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on* (pp. 276-278).
- D'Arcey, J. T. (2013). *Assessing the validity of FaceReader using facial EMG* (Doctoral dissertation, California State University, Chico).
- Den Uyl, M. J., & Van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. In *Proceedings of measuring behavior*, 30(1), 589-590.
- Drozдова, N. (2014). *Measuring Emotions in Marketing and Consumer Behavior: Is Face Reader an applicable tool?* (Master's thesis).
- Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Intellect books.
- Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it - Volume 1* (pp. 119-128). British Computer Society.
- Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.
- Estrada, C. A., Isen, A. M., & Young, M. J. (1997). Positive affect facilitates integration of information and decreases anchoring in reasoning among physicians. *Organizational behavior and human decision processes*, 72(1), 117-135.
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing* (No. 37). Oxford University Press.

- Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical transactions-royal society of London series B Biological Sciences*, 1367-1378.
- Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & emotion*, 19(3), 313-332.
- Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology*, 65(1), 45-55.
- Fredrickson, B. L., Tugade, M. M., Waugh, C. E., & Larkin, G. R. (2003). What good are positive emotions in crisis? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of personality and social psychology*, 84(2), 365.
- Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002, March). Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (pp. 51-58). ACM.
- Green, A. (1995). Verbal protocol analysis. *The psychologist*, 126-129.
- Green, J. D., & Arduini, A. A. (1954). Hippocampal electrical activity in arousal. *J. Neurophysiol*, 17(6), 533-557.
- Haak, M. J. van den, & Jong, M. D. T. de. (2003). Exploring two methods of usability testing: concurrent versus retrospective think-aloud protocols. *Professional Communication Conference. Proceedings. IEEE International, IPCC 2003*.
- Hasselmo, M. E., & Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural networks*, 18(9), 1172-1190.
- Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11), 498-504.
- Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, 50(1), 243-271.
- Hondrou, C., & Caridakis, G. (2012, May). Affective, natural interaction using EEG: sensors, application and future directions. In *Hellenic Conference on Artificial Intelligence* (pp. 331-338). Springer Berlin Heidelberg.
- Hoque, M., & Picard, R. W. (2011, March). Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Proc. of IEEE Int. Conf. on Automatic Face & Gesture Recognition* (pp. 354-359).
- Isen, A. M., & Daubman, K. A. (1984). The influence of a ect on categorization. *Journal of personality and social psychology*, 47(6), 1206-1217.
- Just, M. A., & Carpenter, P. A. (1976). The role of eye-fixation research in cognitive psychology. *Behavior Research Methods*, 8(2), 139-143.
- Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological science*, 4(6), 401-405.
- Kaur, K., & Singh, H. (2015). Analysis of Website using Click Analytics. *International Journal of Science, Engineering and Computer Technology*, 5(6), 185.
- Kuusela, H., & Paul, P. (2000). A Comparison of Concurrent and Retrospective Verbal Protocol Analysis. *The American Journal of Psychology*, 113(3), 387-404.

- Kuusela, H., Spence, M. T., & Kanto, A. J. (1998). Expertise effects on prechoice decision processes and final outcomes: A protocol analysis. *European Journal of Marketing*, 32(5/6), 559-576.
- Lewinski, P., Franssen, M. L., & Tan, E. S. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, 7(1), 1-14.
- Lewiski, P. (2015). The role of facial expression in resisting enjoyable advertisements.
- Lin, C.-T., Chen, Y.-C., Huang, T.-Y., Chiu, T.-T., Ko, L.-W., Liang, S.-F., Hsieh, H.-Y., Hsu, S.-H., and Duann, J.-R., 2008. Development of wireless brain computer interface with embedded multitask scheduling and its application on real-time driver's drowsiness detection and warning. *IEEE Transactions on Biomedical Engineering*, 55(5), 1582-1591.
- Manakhov, P., & Ivanov, V. D. (2016). Defining Usability Problems. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 3144-3151). ACM.
- McDonalds, S., Edwards, H. M., & Zhao, T. (2012). Exploring Think-Alouds in Usability Testing: An International Survey. *IEEE Transactions on Professional Communication*, 55(1), 2-19.
- Neurosky. (2015, 01 maart). Technical Specs. Geraadpleegd op 29 maart, 2017, van <https://store.neurosky.com/pages/mindwave>
- Nielsen, J. (1993). *Usability Engineering*. Mountain View, Californië, Verenigde Staten: Academic Press.
- Nielsen, J. (2012, 4 januari). Usability 101: Introduction to Usability. Geraadpleegd van <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- Nielsen, J. (2013, 04 juni). How Many Test Users in a Usability Study? Geraadpleegd van <https://www.nngroup.com/articles/how-many-test-users/>
- Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256). ACM.
- Palaniappan, R. and Mandic, D.P., 2007. Biometrics from brain electrical activity: a machine learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4), 738-742.
- Parreren, C. F. (1971). *Psychologie van het leren*. Van Loghum Slaterus.
- Peck, E., Chauncey, K., Girouard, A., Gulotta, R., Lalooses, F., Treacy Solovey, E., Weaver, D., and Jacob, R., 2010. From brain to bytes. *XRDS*, 16(4), 42-47.
- Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science*, 17(4), 292-299.
- Pretorius, M. C., Calitz, A. P., & van Greunen, D. (2005, July). The added value of eye tracking in the usability evaluation of a network management tool. In *Proceedings of the 2005 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries* (pp. 1-10). South African Institute for Computer Scientists and Information Technologists.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), 372.
- Rebolledo-Mendez, G., Dunwell, I., Martínez-Mirón, E. A., Vargas-Cerdán, M. D., De Freitas, S., Liarokapis, F., & García-Gaona, A. R. (2009, July). Assessing neurosky's usability to detect attention levels in an assessment exercise. In *International Conference on Human-Computer Interaction* (pp. 149-158).

- Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain*, 66(1), 3-8.
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye Tracking Research and Applications symposium* (pp. 71-78). New York: ACM Press.
- Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts Beyond Words: When Language Overshadows Insight. *Journal of Experimental Psychology: General*, 122(2), 166-183.
- Smith, T. J. (2013). Watching You Watch Movies: Using Eye Tracking to Inform Cognitive Film Theory. In A. P Shimamura (Red.), *Psychocinematics: Exploring Cognition at the Movies* (pp. 165-191). New York, Verenigde Staten: Oxford University Press.
- Talarico, J. M., Berntsen, D., & Rubin, D. C. (2009). Positive emotions enhance recall of peripheral details. *Cognition and Emotion*, 23(2), 380-398.
- Talarico, J. M., LaBar, K. S., & Rubin, D. C. (2004). Emotional intensity predicts autobiographical memory experience. *Memory & cognition*, 32(7), 1118-1132.
- Tobii Technology. (2015). *Accuracy and precision Test report (X3-120 fw 1.7.1)*. Geraadpleegd van <https://www.tobii.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-pro-x3-120-accuracy-and-precision-test-report.pdf>
- Wilson, T. D., & Schooler, J. W. (1991). Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions. *Journal of Personality and Social Psychology*, 60(2), 181-192.

Appendix

Informed consent

Introductie

U bent hierbij uitgenodigd deel te nemen aan onderzoek naar onderzoeksmethoden gericht op het vaststellen van usability. Traditioneel usability onderzoek is sterk afhankelijk van de zelfrapportage en subjectieve evaluatie van participanten. Het is daarom erg nuttig om te weten te komen of een kwantitatieve benadering, waarbij voornamelijk naar gedrag wordt gekeken, betere inzichten biedt op het gebied van gebruiksvriendelijkheid.

Aanpak

Wanneer u besluit deel te nemen aan het onderzoek, wordt u gevraagd een aantal vooraf vastgestelde taken te volbrengen op de website van SUMO restaurants. U zult hierbij geobserveerd- en gefilmd worden en na het uitvoeren van de opdracht een aantal vragen beantwoorden over de moeilijkheidsgraad en de algemene kwaliteit van het systeem. Tijdens het uitvoeren van de opdracht worden oogbewegingen, muisklikken, hersengolven en gezichtsuitdrukkingen geregistreerd. De opdracht en vragenlijst zullen in totaal zo'n 15 minuten in beslag nemen.

De waarnemer is gemachtigd het onderzoek ten alle tijden vroegtijdig te beëindigen of besluiten de resultaten niet in het onderzoek op te nemen in het belang van- en zonder vooraf toestemming te vragen van de proefpersoon. Andersom is de proefpersoon ook altijd gemachtigd onderzoek de test ten alle tijden beëindigen, om wat voor reden dan ook.

Risico's

Er zijn geen risico's verbonden aan de deelname in dit onderzoek.

Privacy

De data die verzameld wordt is vertrouwelijk en zal alleen anoniem in het onderzoek verwerkt worden. De data wordt gepubliceerd in de vorm van een master thesis, geïnitieerd door Universiteit Utrecht. De gegeneraliseerde data van het onderzoek zal gebruikt worden door de originele opdrachtgever Ruigrok NetPanel en de onderwijsinstelling Universiteit Utrecht. Alvorens de publicatie van het onderzoek zal de data ten alle tijden digitaal beveiligd zijn met een wachtwoord.

Uw rechten als proefpersoon

Deelnemen aan dit onderzoek is op vrijwillige basis. U heeft het recht te besluiten om niet meer te willen deelnemen aan het onderzoek, of het onderzoek vroegtijdig te beëindigen. Mocht u besluiten zich terug te trekken zullen hier geen consequenties aan verbonden zijn en u zal niet verplicht worden tot het verschaffen van nadere informatie.

Contactgegevens voor vragen of problemen

Voor algemene vragen kunt u contact opnemen met de onderzoeker. Nova Eeken is te bereiken op het telefoonnummer +31 638278172, of stuur een email naar n.a.eeken@students.uu.nl.

Neem contact op met Jeroen Benjamins, externe begeleider van Universiteit Utrecht via 030 253 1244 of J.S.Benjamins@uu.nl wanneer u vragen of bezwaren heeft met betrekking tot uw rechten als proefpersoon of wanneer zich ongewone activiteiten voordoen.

Toestemmingsverklaring

Bij wijze van een handtekening verklaart de proefpersoon akkoord te gaan met bovenstaande informatie.

Zowel de proefpersoon als de waarnemer krijgen een kopie van het document in bezit.

Naam proefpersoon

Handtekening

Datum

Questionnaire

Naam:

Leeftijd:

Geslacht:

In hoeverre ben jij het eens of oneens met onderstaande stellingen over de website die je zojuist hebt gebruikt?

(1 = helemaal mee oneens, 5 = helemaal mee eens)

System Usability Scale	1	2	3	4	5
Ik denk dat ik deze website vaker zou willen gebruiken.					
Ik vond de website onnodig complex.					
Ik vond de website makkelijk in gebruik.					
Ik denk dat ik hulp van een meer technisch onderlegd persoon nodig heb om deze website te kunnen gebruiken.					
Ik vond de functionaliteiten goed geïntegreerd in de website.					
Ik vond de website erg inconsistent.					
Ik kan me voorstellen dat veel mensen deze website snel leren gebruiken.					
Ik vond de website erg lastig te gebruiken.					
Ik voelde me erg zelfverzekerd tijdens het gebruiken van de website.					
Ik moest een hoop leren voor ik aan de slag kon met deze website.					

Hoe heb je de moeilijkheidsgraad van de handelingen ervaren die je hebt uitgevoerd op de website?

(1 = heel makkelijk, 5 = heel moeilijk)

	1	2	3	4	5
Het bereiken van de bestelomgeving					
Het vinden en selecteren van de gewenste gerechten					
Het vinden van de knop 'Afrekenen' om het check-out proces te starten					
Het overzicht van de bestelling controleren					
Het invullen van mijn persoonsgegevens					
Het kiezen van de bezorgtijd					

Wat was jouw ervaring tijdens het uitvoeren van de handelingen op de website?

(1 = heel prettig, 5 = heel vervelend)

	1	2	3	4	5
Het bereiken van de bestelomgeving					
Het vinden en selecteren van de gewenste gerechten					
Het vinden van de knop 'Afrekenen' om het check-out proces te starten					
Het overzicht van de bestelling controleren					
Het invullen van mijn persoonsgegevens					
Het kiezen van de bezorgtijd					

Is er iets dat deze ervaring had kunnen verbeteren?

Correlation matrix

Segment 1

* Correlation is significant at the 0.05 level (2-tailed).

Segment 1	Aluminum	Aluminum	Kochium	Zinc	Boa Diff	Thoa	Alma	Club not	Club extra	Days extra	TTT	Thudium	Treant	Idexy	Sid	Ampr	Shurton	Harad	Expend	Vance	Alvost	Edmich	Yurham	Treant			
Aluminum	Correlation Coefficient Sig. (2-tailed) N																										
Kochium	-.145 .000 30	Correlation Coefficient Sig. (2-tailed) N																									
Zinc	.435 .000 30	.435 .000 30	Correlation Coefficient Sig. (2-tailed) N																								
Boa Diff	.111 .000 30	.111 .000 30	.435 .000 30	Correlation Coefficient Sig. (2-tailed) N																							
Thoa	.183 .000 30	.183 .000 30	.435 .000 30	.111 .000 30	Correlation Coefficient Sig. (2-tailed) N																						
Alma	.192 .000 30	.192 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	Correlation Coefficient Sig. (2-tailed) N																					
Club not	.132 .000 30	.132 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	Correlation Coefficient Sig. (2-tailed) N																				
Club extra	.272 .000 30	.272 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.132 .000 30	Correlation Coefficient Sig. (2-tailed) N																			
Days extra	.444 .000 30	.444 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	Correlation Coefficient Sig. (2-tailed) N																		
TTT	.117 .000 30	.117 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N																	
Thudium	.052 .000 30	.052 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N																
Treant	.042 .000 30	.042 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N															
Idexy	.132 .000 30	.132 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N														
Sid	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N													
Ampr	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N												
Shurton	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N											
Harad	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N										
Expend	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N									
Vance	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N								
Alvost	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N							
Edmich	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N						
Yurham	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N					
Treant	.084 .000 30	.084 .000 30	.435 .000 30	.111 .000 30	.183 .000 30	.192 .000 30	.272 .000 30	.272 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	.444 .000 30	Correlation Coefficient Sig. (2-tailed) N					

Segment 5

Program	Component	Coversation																						
		Amurto	Medison	Zane	Res. DT	Thia	Alma	Chas. MP	Chas. sum	Page sum	TTTRIC	Relacion	Happy	Est	Amy	Barrio	Scors	Dagard	Vance	Alvord	Edwards	Chasart	Talano	
Amurto	Component	348	-243	506	247	-151	-28	240	-142	11	65	-122	125	131	-262	125	131	129	176	206	202	222	222	222
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Medison	Component	442	-410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Zane	Component	298	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Res. DT	Component	287	-112	-1018	297	-109	297	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	-109	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Thia	Component	124	931	931	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Alma	Component	424	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	918	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Chas. MP	Component	606	629	249	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Chas. sum	Component	442	410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Page sum	Component	234	-147	154	-224	-189	-117	440	477	477	477	477	477	477	477	477	477	477	477	477	477	477	477	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
TTTRIC	Component	-298	299	190	478	-169	-108	-283	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	-272	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Relacion	Component	645	233	342	465	472	472	472	472	472	472	472	472	472	472	472	472	472	472	472	472	472	472	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Happy	Component	445	440	129	434	-109	-104	148	148	148	148	148	148	148	148	148	148	148	148	148	148	148	148	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Est	Component	324	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	344	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Amy	Component	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	408	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Barrio	Component	442	410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Scors	Component	298	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Dagard	Component	442	410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Vance	Component	606	629	249	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Alvord	Component	442	410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Edwards	Component	298	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	299	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Chasart	Component	442	410	401	432	-414	406	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	422	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	
Talano	Component	606	629	249	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	678	
	Sig. (2-tail)	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	

* Correlation is significant at the 0.05 level (2-tail).

** Correlation is significant at the 0.01 level (2-tail).

Segment 6

Program	Activity	Component	Account	Metric	Zone	Item	Alpha	Click out	Click rate	Click cost	Page cost	CTR	Estimate	Imps	Sd	Appr	Spend	Budget	Variance	Actual	Efficiency	Lift/loss	Tadpole									
																								Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)	Seg. (2-adj)
Program 1	Activity 1	Component 1	Account 1	Metric 1	Zone 1	Item 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00			
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
							0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

* Conversion is calculated as the ratio of Clicks to Impressions

