MASTER THESIS

INFORMATION STUDIES

AALBORG UNIVERSITY

COPENHAGEN

CHATBOT VS. ONLINE CUSTOMER EXPERIENCE

Author

Anna Maria Bak

Supervisor: Birger Larsen

Number of characters: 106622

Semester: 10th

Date of delivery: 12th September 2019

TABLE OF CONTENTS

01. INTRODUCTION

The official definition of Chabot (also referred to as talk bot, chatterbot, Bot, IM bot, interactive agent, Conversational interface or Artificial Conversational Entity) describes it as a "computer program that process natural language input from a user and generates smart and relative responses that are then sent back to the user" (Khan & Das, 2018)
The first-ever chatbot was developed by Joseph Weizenbaum in 1966 at MIT Artificial Intelligence Laboratory. Named Eliza, it was designed to mimic a psychotherapist (Khan & Das, 2018). The term chatbot, however, was first introduced in 1994 by Michael Mauldin, creator of verbal robot Julia.

Starting in 2014, massaging systems started implementing support for bots - conversational agents that can interact with users directly through the messaging apps themselves (Klopfenstein at.al, 2017). Today's rising interest in chatbot technology is owned mainly to Facebook and Slack. In April 2016 during the F8 developer conference, Facebook opened its platform for developers, providing them with an enormous user database that can be used to provide instant assistance via chat (Khan & Das 2018). According to Pandorabots (www.pandorabots) self-proclaimed world's leading chatbot platform they already host 285 thousand chatbots in 2016 alone and since its announcement in April the same year, Facebook reported 11 thousand built (Dale, 2016).

There are several papers regarding the development of chatbot for different purposes as well as evaluation or comparison of specific bots. On the other hand, there is a very limited research material investigating the relationship between chatterbots and user's satisfaction of their overall online customer experience (OCX). Does it change the user expectation? How much of an impact does the chatbot performance have on website's OCX?

Customer experience is a subjective response of an individual to a set of interactions between them and the company or its products (Gentile, Spiller, and Noci, 2007; Meyer and Schwager, 2007). Following this definition, I assume, that OCX can be seen as customer experience applied in the online environment. Studies suggest that providing a compelling OCX has many positive outcomes for website owners and thus for the business (Hoffman and Novak 1996; Gentile et al. 2007). It is, therefore, a critical component that cannot be overlooked.

In this paper, I aim to investigate the impact and its potential magnitude of chatbot implementation on OCX. I strive to illustrate potential changes in visitor's expectation and perception related to adding

a chatbot as one of the touchpoints in the online customer journey and ending with listing possible concerns that should be addressed while planning the introduction of the tool.

As different companies offer many different products whilst communicate through various channels, depending on their audience, it is impossible to formulate a general answer to those questions. For the purposes of this work, I am using VELUX as my case example. They introduced a chatbot in April 2019 to their customers in Great Britain. To address these phenomena, I formulated the following problem statement:

1.1. PROBLEM STATEMENT

Does the implementation of chatbot impacts VELUX online customers experience (OCX)?

RQ1: *What are the most important factors that influence customer perception of their online experience?*

In order for an experience to be good or bad, there are sets of components that need to be assessed. Although some of the factors are more person and situation-dependent, there are some that are universal. In this question, I investigate the most significant factors that impact people's perception of OCX in order to identify those that could potentially change the visitor's opinion of VELUX OCX.

RQ2: *How do the client expectations regarding OCX change after the implementation of a chatbot?*

Including a new interactive element to a website can change visitors expectations regarding its performance. By answering this question, I try to define additional determinants of OCX that are tied to the introduction of new technology to the VELUX website. Together with factors identified in the previous question, it will create a base for discovering whether or not the chatbot impacts the OCX of VEUX website.

RQ3: *What do customers think of VELUX OCX after chatbot implementation?*

Through the final question, I aim to learn what the subjective opinion of VELUX website visitors is. Connecting the results with the list of factors created beforehand, I will be able to draw conclusions and answer the problem statement.

1.2. CASE DESCRIPTION

1.2.1. VELUX

VELUX is a Danish manufacturing company founded in 1941, aiming to "create better-living environments using daylight and fresh air through the roof – for life, work and play" (VELUX A/S, n.d). Their products include roof windows, skylights as well as corresponding accessories such as roller shutters, blinds, remote control units, and installation solutions.
The company is currently a part of VKR Holding A/S with seventeen production facilities in nine countries and individual sales companies in forty countries worldwide (VELUX A/S, n.d).

When it comes to online presence, VELUX owns more than sixty individual websites in total. They differentiate between types of a webpage depending on their purpose and product range.

The corporate site (velux.com) - Its an official global website focusing on the company itself, its history, purpose, vision and values. Visitors can also find information regarding the company's' health research, social initiatives as well as get an overview of the product range.

Marketing site - This type of website has a local version in all of the markets the company is currently present at. Those sites are divided into two sections depending on the type of the customer: professionals (architects, builders, installers) and homeowners. Its purpose is to present all the information regarding the available products and services. It also provides visual project inspiration including replacement, loft conversion and room gallery as well as all the technical specification and installation instructions. Visitors can use a variety of online tools such as *ContactPro* where people can find a certificated installer near them or *Roof Window Price Calculato*r which allows them to get a total cost of products required for their projects.

VELUX shop - Online stores are selling branded VELUX window accessories. Besides blinds and shutters, visitors can also order electric solutions as well as maintenance kits and locks. Webshops are not available in all the countries, and the availability of individual models differ from market to market.

Itzala - Those are online stores marketing both VELUX and Itzala window accessories. Itzala is a brand run by VELUX owned subsidiary Altarterra Ltd. Itzala branded products have a more narrow range but can fit more than one brand of roof windows. For example, they can also fit FAKRO and RoofLITE goods. Similarly to VELUX shops, they are not available on all the markets.

Solstro - Solstro is a part of VELUX and VKR Group. Those sites are online stores offering a wide range of roof windows and roof window accessories at more affordable prices. Additionally, they are also selling VELUX branded roof window products and solutions. Solstro webshops are only present on some of the European markets.

Other - VELUX also has some stand-alone domains. Those sites vary from market to market and can relate to various of the company's' global and/or local initiatives

1.2.2. CHATBOT IMPLEMENTATION

VELUX is constantly looking for ways to improve their experience on the website.
In April this year, the company introduced chatbot function on marketing sites in Netherlands and United Kingdom, with hope to reduce complexity regarding product choice as well as the entire process of planning, buying and installing their products.

Previous similar activities include live chat on some of the markets (e.g. Poland, United Kingdom) and chatbot experiment in the USA. There is, however no comprehensive overview of those type of existing solutions across markets, and there is no data available regarding their performance at the moment.

This particular chatbot launch is targeted primarily at the End users in the upper part of the marketing funnel (consideration phase). Those visitors are interested in information about a replacement, better space or more space project for their homes, seeking inspiration or more technical material. The chatbot is meant to open a new communication channel and provide a shortcut through the customer journey by guiding visitors to relevant content, answering key questions.

Implementation in the United Kingdom was initially divided into three phases with around two-week intervals between each phase; however in the end only two of them have been completed. During the initial run, the chatbot is presented only on the following subpages:

- windows overview section;
- In-page search result pages;
- help and advice - where to buy;
- Help and advice - VELUX guarantee

- VELUX Extensions budgeting
- Room Gallery
- VELUX Extension bundle (product page)
- VELUX Loft bundle (product page)

In the second phase, more product pages including electrical manoeuvred roof windows and black-out blind product page (list of all URLs can be found in Appendix 2)

Upon writing this section, the timeline for the implementation in the Netherlands is not yet fully confirmed.

1.2 SCOPE

This paper will primarily focus on one of the markets. Due to time constraints and language barrier, this work concentrates on the VELUX marketing site in the United Kingdom. Additionally, because of time constraints, all the potential testing take place will not concern the final form of the solution as any potential optimisation and/or expansion are scheduled to take place in the last quarter of the year.

02. LITERATURE

2.1 LITERATURE SEARCH

According to Randolph (2009), a literature review can help the author not only to gather valuable insights but also to demonstrate their knowledge regarding a topic including relevant vocabulary, theories, as well as important variables or influential researchers in the field. In this part of my work, In his section, describe the methodology behind the literature search process that I used in my literature review. It also contains a description of the sources and the evaluation process. The majority of the literature was obtained based on the through systematic search and the semester's curriculum.

2.1.1 TAXONOMY

For this work, I based my literature search and review process on Cooper's Taxonomy of Literature Reviews (1988) in which he categorises the literature review process into six components.

The first component is **the focus**. This component helps to define the foci of the review (Cooper, 1988). Cooper identifies four main foci categories: research outcomes, research methods, theories and practices/applications. It is common for the dissertation to have more than one focus, as they are not mutually exclusive (Cooper, 1988). In my paper, I decided to centre my review on research outcomes methods. It will assist me in identifying already existing variables, measures, and methods used in the field of online customer experience as well as the results produced. **The goal**, which is the second component, outlines the objective of the review (Cooper, 1988). My goal is to synthesise and critically examine the previous work within the field. **Perspective**, the third component concerns the influence of the reviewer's own subjective opinions on literature discussion (Cooper, 1988). According to Randolph (2009), the choice of perspective depends on whenever the review is quantitative or qualitative. As my work is primarily qualitative research, I choose what is called an espousal of position. This perspective means that as a researcher, I have a more editorial impact on the review, revealing biases and reflecting upon findings in order to make a certain point. When it comes to **coverage,** Cooper (1988) mentions four types of this component, based on an extent in which the author chooses to find research and include them in their review. I decided on an exhaustive review with selective citation. I aim to obtain every work written about a specific topic. However, this research must fall within a predefined scope. My scope excludes all the materials unavailable online. Additionally, I will not look into blogs and website articles. Two remaining components include **organisation** and **audience** (Cooper, 1988). I organised this review based on concepts that are relevant as opposed to arranging them in chronological order or grouping them according to methods and theory. I find this format to be easiest to follow, keeping in mind that this work will also be presented to the company used as the case.

2.1.2 SEARCH QUERIES

Following Cooper's (1988) taxonomy, I identified different themes and sub-themes based on my problem statement and individual research questions:

Online customer experience

Being the main focus of my paper, I need a better understanding of the OCX concept. Particularly it's various components that have the most influence on how the user perceives OCX. Identifying those factors will help me to create unified measurements strategy for this paper.

Chabot

Here, I look into Chabot's definition as well as its applications in peoples daily use. Knowing user preferences will help me establish their potential expectations towards chatbot technology.

Chabot assessment

While introducing a new interactive touchpoint in the online customer journey, one must evaluate not only if it's working but also its performance level. Investigating chatbot assessment methodologies will help me identify additional, relevant factors that might have an impact on customer's OCX perception.

Keywords

The use of keywords is the most popular way of searching databases. It is important to carefully choose the said keywords in order to produce the most relevant results (Cronin, Ryan & Coughlan, 2008). In my process, I created initial keywords for each theme, gradually expanding them with corresponding synonyms.

As my knowledge in the area was very limited, I relied heavily on "*pearl growing*" search strategy. This strategy is using found literature as a base for retrieving other work (Rowley, Slack 2004). I used articles from the first search to expand my keywords list further using topic-specific terminology. Additionally, I used their already existing reference as an inspiration to bring more material into my literature review.

2.1.3 SOURCES

In order to cover a broad range of articles, I used several databases available online. Keeping in mind that databases do not overlap perfectly, my choice of using different sources provides more material but also adds validity to the search process. The material gathered for the literature review comes from the following sources: Aalborg University library catalogue, Google Scholar and ResearchGate.

2.1.4 SELECTION

To analyse gathered papers, I followed the PQRS method introduced by Cohen (1990) as described in Cronin et al (2008). It is designed to ease the process of identifying and selecting papers that are most relevant for the research. The method consists of four steps: preview, question, read and summarise.

In the **preview**, I read through the abstract of each paper found. Additionally, I scan the introduction and conclusion to preliminary exclude works that I found irrelevant. In **the question** phase, I matched initially selected articles, based on my research questions. Those became the base for further elimination. The papers that did not make it through this phase were stored regardless in a separate folder. During **the reading** stage, I thoroughly went through selected work while making notes. I focused primarily on methodology and findings keeping in mind previously identified themes as well (see section 2.1.2). When it comes to **summarising**, I collected all my notes, transforming them into a longer, more coherent summary.

Lastly, I went back to articles discarded in question stage - reading those more carefully to make sure I did not make a mistake.

2.2 LITERATURE REVIEW

This literature review aims to present gathered knowledge regarding online customer experience as well as chatbots to identify relevant components that create the base for my research. I start with online customer experience definition, including its various determinants. Next, I present an introduction and history of a chatbot, focusing on its usage and motivation behind it. Finally, I believe that the overall performance of the newly introduced tool is also important. Thus I look into chatbot usability, presenting an overview of various metrics involved, taking into consideration different perspectives.

2.2.1 Online Customer Experience

In today's media-heavy society, the experience became the focal point of businesses differentiation strategies. Customers, having access to numerous contact points with companies, shifted attention from transaction-based customer relation to the concept of Customer Experience (CX) (Gentile, Spiller & Noci, 2007). There are several works naming CX as an important component in creating value for customers as well as the business (Gentile et al. 2007; Verhoef, 2009; Rose, Clark, Samouel & Hair, 2012). Klaus and Maklan (2013) tested number hypotheses, investigating relations between a positive CX and outcomes such as customer satisfaction, potential loyalty intentions and word of mouth behaviour. All of the hypotheses were validated.

Further, in their paper, Gentile et al. (2007) define customer experience as a subjective reaction of a customer to their interaction with company's product, part of an organisation or the firm on its own. According to them, CX consist of sensorial component (senses stimulation: vision, hearing, touch, etc.), emotional component (moods, feelings) and a cognitive component (mental processes). Pragmatic, lifestyle and rational components are also mentioned. Verhoef (2009) expanded this definition with a more holistic approach, separating customer responses into cognitive, affective, emotional, social and physical. At the same time, he highlighted the fact that the experience includes all stages of the purchasing process (search, purchase, consumption and after-sales. In today's world, it also involves multiple channels including, but not limited to, an online website.

Hoffman and Novak (1996) suggested that in order for a website to be successful in compiling its visitors, it must facilitating a state of flow. They defined as " *the state occurring during network navigation, which is (1) characterized by a seamless sequence of responses facilitated by machine interactivity, (2) intrinsically enjoyable, (3) accompanied by a loss of self-consciousness, and (4) self-reinforcing*". Visitors would achieve flow when they achieve a balance between their skill level and challenges posed by the interaction. Flow increases user's exploratory and participatory behaviour as well as positive subjective experiences. On the other hand, as the flow state can be a reward on is own, there is a risk of visitors being distracted. They might want to concentrate on exploring the website as opposed to aiming to find the information they need. Authors used the concept of flow as a basis for their process model of network navigation in the hypermedia Computer-Mediated Environments. They identify four Flow determinants, all of which were connected with user's cognitive responses: perceived congruence of skills and challenges; focused attention; interactivity; and telepresence. First two components are necessary for the flow state to occur, whares the other will enhance it.

In their paper from 2002 Hoffman, Novak and Yung created a structural model incorporating factors that make up for a good OCX. Using results from multiple surveys dispatched to a large web-based consumer sample, they tested thirteen individual constructs from the model introduced in their previous paper. The table below presents relationships between those constructs that have been investigated.

Direct influence on flow	Indirect influances on flow	Consequences of flow
Higher skill at using the Web & higher perceived control during interaction increases flow.	Greater importance corresponds to greater focused attention	Greater flow corresponds to greater exploratory behavior
Greater challenge & arousal increse flow	Higher speed of interaction increases focused attention, telepresence and time distortion & flow	Greater flow corresponds to greater positive affect.
Greater telepresence & time distortion increase flow		
Greater focused attention increases flow		

Table 1:Overview of relationships between variables in Hoffman et al. (2002). Source : Own creation

Additionally, the authors researched relationships between constructs and consumer behaviour as well as web usage. According to them, the feeling of fun and exploration declines, the longer visitors use the web. They also claimed that the flow would be greater for visitors using web for experimental purposes rather than goal-oriented tasks. This particular statement was disproved in their later work, where it was found that experiencing flow was more prevalent among goal-oriented users (Novak,Hoffman & Duhachek 2003).

Most of the constructs were confirmed through extensive testing and validation processes, however, there are some that did not prove true. Increased focused attention did not boost the flow. Nonetheless, it did correspond with greater telepresence and time distortion making it an indirect factor. There were also some new correlations that emerge in the process. The interactive speed proven to have a direct positive effect on the flow, challenge was positively connected with focused attention and importance was beneficial to the skill level of the user. Lastly positive affect was completely excluded as it became untestable due to elimination of all corresponding variables.

Rose et al. (2012) explore CX in online shopping context. The study, on top of cognitive components, incorporates also affective ones.

Cognitive state variables	Affective state variables		OCX Outcomes
Interactive speed	Easy-of-use		Satisfaction
Telepresence	Customization	Preceived Control	Trust
Chellenge	Connectedness		Repurchase Intentions
Skill	Aesthetics		
	Preceived Benefits		

Table 2 Overview of variables from Rose et al. Source: Own creation

In the table above, three affective variables: ease-of-use (navigation, functionality); customization (tailoring website appearance and functionality); and connectedness (ability to connect and share with others virtually) are connected through perceived control. Authors used this attitudinal variable to explain consumer behaviour.

Testing the relationships between those variables, the authors formulated eighteen hypotheses. Initial tests proved "partial empirical validation of the theoretically assumed relationship between latent variables" (Henseler, Ringle, and Sinkovics 2009 as cited in Rose et al. 2012) for all of them. Based on that, authors defined online customer experience as "*psychological state manifested as a subjective response to the website"(*Rose et al. 2012).

Other studies suggest additional variables such as playfulness and personal innovativeness (Agarwal & Karahanna, 2000); content/interface (Choi, Kim & Kim, 2007); novelty (Huang, 2003) or perceived usefulness (Agarwal & Karahanna, 2000; Hsu & Lu, 2003; Sanches-Franko, 2006).

Lastly, Hoffman & Novak (2009) reviewed their model by creating overview of other work in online customer experience field. They highlighted a variety of measurements proposed in different papers. They found unidimensional measures to be simpler, easier to gather the data for. On the other hand, multidimensional measures allow a more holistic definition of flow able to be tested for statistical fit in a structural model. Multiple measures of flow were recomened to use when possible.

2.2.2 CHATBOT

Current Chatbots can be defined as independent computer programs running on "rules-driven engines or artificial intelligent (AI) engines that interact with users via a text-based interface primarily" (Khan & Das, 2018). They can be plugged into multiple messaging systems as long as those have an opened platform API. Initially, chatbots have been developed for entertaining purposes, with imitating human speech as a primary goal Shawar& Atwell, 2007). Created in 1966 Eliza was impersonating a psychoanalyst. SmarterChild, developed in 1995 by ActiveBuddy was the first bot that, next to entertain value, provided users with useful information including , for example, sports scores and weather. Today's voice-activated personal assistants available in Apple (Siri) and Samsung (S Voice) phones are descendants of SmarterChild (Khan & Das, 2018). Although the terms *chatbot* and *conversational agent are* sometimes used interchangeably (Khan & Das, 2018), Wilks (2010) made a clear distinction between the two. While the former is designed to perform specific tasks, for example, ticket ordering, the later only mimics a conversation having a limited set of responses.

The application possibilities of chatbot technology are wide. In their paper, Shawar and Atwell (2007) refer to the work of Fryer and Carpenter (2006) in which they present a case of chatbot assisting in language studies for non-beginner users. Although the bot cannot detect any spelling or grammar mistakes, students who took part in the research reported that they felt more relaxed talking to chatbots as opposed to a life partner. Additionally, students could repeat the material and use the text-based response to practice reading and listening. Another opportunity is online content exploration. H&M clothing brand launched a chatbot which provided users with purchase suggestions based on their personal wardrobe (Brandtzaeg & Følstad, 2018). Chatterbots can also serve as a more efficient alternative to customer service agents. "Do Not Pay" helps users with filing a complaint form in case they received a parking ticket and chatterbot created for Babylon Health gives its online visitors medical advice (Bandtzaeg & Følstad, 2017).

Despite the initial enthusiasm, some mention that the adoption of chatbot technology is smaller than originally predicted (Simonite, 2017 as cited in Brandtzaeg & Følstad, 2017). One of the main reason

for this situation is a common chatbot's failure to fulfil user needs caused by unclear purpose, unclear responses and fault usability Coniam, 2014 as cited in Brandtzaeg & Følstad, 2017).

Authors Brandtzaeg and Følstad (2017) investigate individuals' main motivations for using chatbots. They surveyed bot users age 16 to 55, based in the United States. Among their 146 responders, 52 of them were male, and the remaining 94 were women. The majority (64%) was familiar with chatbots and had been using it for at least 2 years.

Productivity was reported as the main motive for using chatterbots (68%). Participants pointed towards easy of use, convenience and access to information as central benefits of the tool. 20% chose entertainment purposes as the main motivation, describing the tool as "fun" and by that having an additional value. The third most reported reason was the potential social benefits that Chatbots can provide (12%). Here, participants (10% of the whole sample) mentioned using chatbot as a mean to avoid loneliness by interacting with the tool rather than socializing with other people. The novelty of the chatbot was the fourth most reported motivation for using the tool (10%). Those people were looking into the bots abilities and their limits. Other responses could not be clearly categorised and usually occurs as a single response. Those motives included: Chatbots being a default method of customer support or providing an automated answer when no alternative is available.

2.2.3 CHATBOT ASSESSMENT

At the moment, there is a lack of a unified, widely applicable metric framework for chatbot evaluation (Io & Lee, 2017 as cited in Peras, 2018).

One of the most widely used chatbot evaluation frameworks is PARAdigm for DIalogue System Evaluation (PARADISE) (Cahn, 2017). The model separates what bot needs to accomplish in terms of task requirements from how the task is fulfilled from a dialogue perspective (Walker et al., 1997). Using questionaries to collect users ratings, authors asses subjective factors such as ease of usage, naturalness, friendliness, or willingness to use the system again. PARADISE also attempts to objectively evaluate the effectiveness of the bot through maximizing task success whilst minimizing dialogue costs. Dialogue costs are defined as efficiency costs (total elapsed time, number of systems turns,

total number of system turns per task, and total elapsed time per turn) and qualitative costs (number of re-prompts, number of user barge-ins, number of inappropriate system responses, concept accuracy, turn correction ratio) (Walker et. al, 1997). Hung et al. (2009) used PARADISE to measure the effectiveness and naturalness - namely how well it can sustain the natural flow of the conversation - of the chatbot. For this purpose, the authors defined efficiency costs as resource consumption used to accomplish an individual task and quality costs as the content of the conversation.

Kuligowska (2015) proposed a number of metrics for the evaluation of chatbots used in the business sector. The evaluation was applied to 29 polish- speaking bots using a 5 point rating scale. Author asses the following bot attributes: visual appearance, implementation form on the website, speech abilities, knowledgebase, knowledge presentation and its usability(click-through links, ability to scroll through past dialogues etc), conversational abilities, personality,personalization options, responses in unexpected situations and possibility for users to rate the chatbot and website. The framework, although providing a commercial point of view, is subjective. Venkatesh et al. (2018) attempts to reduce this subjectivity by introducing measurements that correlate with human judgement. At the same time, the framework enables a detailed analysis, not possible with user ratings. The unified method can be used to compare chatbots against each other, including non-goal oriented dialogue systems. It combines user engagement, coverage of topics, consistency, variety of content and depth of the conversation as measurements. The validity of those metrics was confirmed by finding an undisputed connection between selected factors and hundreds of thousands of ratings provided by users during live chat sessions.

Chakrabarti & Luger (2012) use Grice's conversational maxims as the base for bot appraisal. Chatbot would meet the quality maxim if the provided responses were factually correct; the quantity maxim if the information included was adequate; the relation maxim if the responses matched the context of the conversation; and the manner maxim if responses were clear. Jwalapuram (2017) further expand on the idea, adding a user perspective. He investigates the correlation between human judgment and Grice's maxims using Likert scale.

Shawar and Atwel (2007) claim that the evaluation methodology should be adopted based on the chatbot's purpose and user needs. They applied that mindset when evaluating three different chatbot prototypes developed base on ALICE chatbot system. They evaluate chatbot's learning techniques success using dialogue efficiency and quality as well as user satisfaction; its ability to be used as an information retrieval system in comparison with a search engine using a number of matched

responses and user preference question; its ability and usefulness as an information source access tool relying on quality assessment of prototype potential users.

Cahn (2017) mentions four different perspectives of chatbot evaluation. Information retrieval perspective, measuring bot's effectiveness through metrics such as accuracy, precision, recall and correctness of the response. User perspective focuses on user satisfaction mostly measured through questionaries. Linguistic perspective requires experts evaluation of aspects such as producing full, meaningful sentences that are grammatically correct. The last, artificial intelligence perspective asses how human a chatbot can come across as, typically using Turning test.

According to Peras (2018), most research concentrates on one perspective whilst ignoring other aspects. The author, therefore, presents a framework incorporating all of the identified contexts in which chatbots can be evaluated, adding a business angle as the fifth perspective. Each of these perspectives was further divided into categories, attributes, metrics and approaches (quantitative vs qualitative) respectively. Below table shows an overview of the perspectives with their categories. The full framework can be found in Appendix 3.

Perspective	User	Information Retrival	Language	Technology	Business
Category	Usability	Accuracy	Quality	Humanity	Business Value
	Performance	Accessablity	Quantity		
	Affect	Eficiency	Relation		
	Satisfaction		Manner		
			Grammatical accuracy		

Table 3: Overview of Chatbot's evaluation metrics based on Peras (2018). Source: Own creation

Each perspective has its own advantages and disadvantages (Peras, 2018). Peras (2018) highlights that not all chatbots need to be evaluated using all five perspectives. The framework should rather serve as a basis for choosing the right approach. Similarly to Shawar and Atwel (2007) athor suggest that the approach should be determinate by chatbot application area and user needs.

03 THEORY

In this segment, I present an overview of the theories selected for this project that will help me with the problem statement. The chapter includes not only the description of the theories but also how each theory connects with my problem statement and thus helps me answer my research questions.

3.1 FLOW THEORY

Customer experience is the main focus of this work. Throughout the literature review on OCX, Flow theory popped up constantly as the main base for any mentioned model. In order to get a better understanding of OCX, I decided to look into the original theory. With this I aim to see if there are more factors that need to be taken into consideration during the designing of the research.

3.1.1 OPTIMAL EXPERIENCE

The concept of Flow was officially introduced by Mihály Csíkszentmihályi in 1975. It was described as an optimal experience. In his work from 1982 Csíkszentmihályi points out that researchers in psychology field tend to focus on behaviour rather than experience although even when interacting with another person, people look into what the other party would do because it will have a direct impact on their experience. He justifies the negligence, describing behaviour as more reliable and more objective measure.

The author defines optimal experience as a subjective state of mind that can occur in the ordered state of consciousness. This means that the level of information received by the person cannot be too little or too much. There is a risk of distraction in attention which leads to failure in processing experience. Optimal experience can be described by what there is to do (challenges) and what the person is capable of doing (skills) . To reach the optimal experience, one must attain equal ratio of both of those variables simultaneously. Higher levels of capabilities over challenges cause a state of boredom, and higher levels of challenges over boredom can lead to anxiety.

3.1.2 FLOW

Flow which with time became synonymous with optimal experience is "*a subjective state that people report when they are completely involved in something to the point of forgetting time, fatigue, and everything else but the activity itself.*" (Csíkszentmihályi 2014 p.230). Theory has its origins in intrinsically motivated or autotelic activity - activity that can be a reward in itself.

Throughout his works, Csíkszentmihályi (1975/2000) investigated the nature of enjoyment, interviewing people from all kinds of professions that pointed towards enjoyment as the main motivation behind their work. He formulated following conditions of flow: balance between challenges and perceived skills;lear goals and immediate feedback on the performance which allows for timely adjustments. Those conditions allow person to reach the state of flow, characterised by: (1) **Intense concentration**, focused solely on the task at hand; (2) **merging the action and awareness**, filtering out irrelevant thoughts and feelings; (3) **loss of reflective self-consciousness**; (4) **sense of**

control, the feeling that one is capable of handling the situation; (5) **time distortion**, feeling that it has passed quicker that it actually did; (6) **experiencing activity as a reward**, where potential goal of it is just an excuse to engage and continue with the process.

When it comes to measuring the phenomena, researchers can use methods such as interview, questionnaire and experience sampling method (ESM). In ESM, participants are given a pegged device that sends a signal within pre programmed time interval, reminding them to fill out a questionnaire regarding their state at the time of the signal. Questions cover cognitive, emotional and motivational aspects of the situation. This level of detail distinguishes ESM from diary method when the description is fully subjective and skewd either positive or negative way.

Next to positive affect and lasting satisfaction, the author hypothesized that flow can be a factor that facilitates personal grow. As one experiences the flow, one tends to replicate the feeling by repeating the action that with time becomes less and less challenging as one require higher level of skills. He also highlights that depending on personality, some people can experience flow more often than the others. Flow comes easier for people with what is called an autotelic personality. Those individuals are curious, persistent and enjoy a high-challenge activities.

Although flow is connected with enjoyment and pleasure., it has its downsides. Csíkszentmihályi (2014) writes that it can have a negative effect in the form of addiction. While they can potentially improve quality of life, they can also trap one's consciousness in a certain order state, without which, one is unable to function in everyday life.

3.2 TECHNOLOGY ACCEPTANCE THEORY

Chatbot is a relatively new technology, especially in the context of business tool. Zarouali, Van den Broeck, Walrave and Poels (20018) used the technology acceptance theory, specifically CAT model to predict consumers response to company's chatbots on the Facebook platform assessing its commercial effectiveness. Results shown that five out of six elements from the model were significant in respect of usage prediction. Keeping in mind that VELUX chatbot is the first one presented to its visitors I decided to look into this particular model as well as others related to this theory. Findings will be used to extend the pool of OCX factors to create a broader yet more accurate overview of important determinants appropriate for VELUX case.

3.2.1 TAM & TAM2 MODEL

Technology acceptance model (TAM) describes factors that determine the user's acceptance of information technology. Davis (1986) claimed that usage of technology can be predicted based on a person's motivation. Whereas motivation is influenced by external factors, namely the system's features. He created the TAM model suggesting that **perceived ease of use (PEOU), perceived usefulness (PU)** and **users attitude** towards using a system, are the main determinants of whether a person will or will not use a particular technology (see figure below).

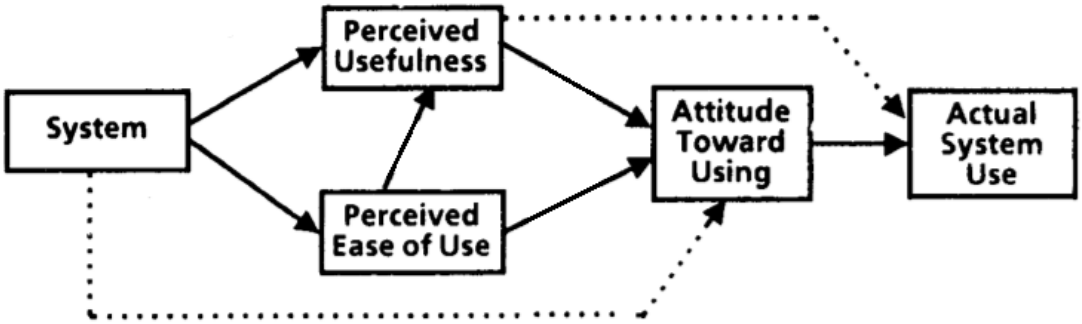

Figure1: TAM Model. Source: Davis (1986)

Davis (1986) hypothesized that the attitude of a user toward the system was the most important, determinant heavily influenced by PEOU and PU which in turn were affected by systems design and its other characteristics. His testing revealed additional relations that were not initially taken into consideration (represented on the figure by the dashed arrow). PU proved that next to its indirect effect on use, it also has a significant direct effect on it. Another example is the system and its characteristics. Although it does affect PU and PEOU it also proved to have a strong direct effect on attitude component.

In his later paper, Davis (1989) removed Attitude as a mediating factor due to lack of support of this hypothesis. In his work written with Bagozz and Warshaw (1989) **behavioural intentions** were introduced. Factor directly influenced by PU. The revised model can be seen in the figure below.

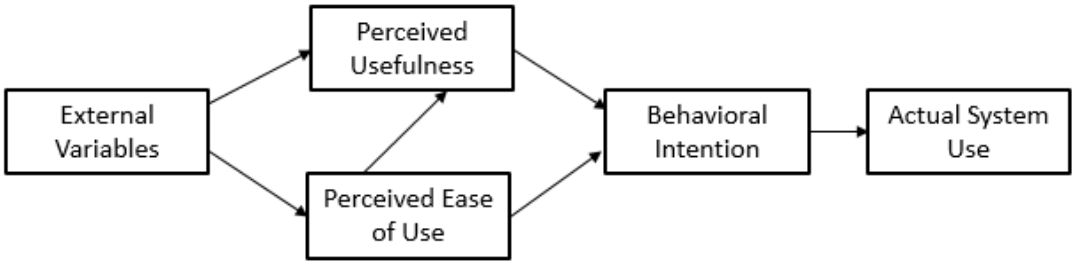

Figure2 Revised TAM Model. Source: Bagozz and Warshaw 1989

Authors suggested that a person might develop behavioural intentions, based on PU without forming an attitude. The change also explained the strong, direct influence of PU on actual usage. Potential external variables were identified as well, being s**ystem characteristics and nature of its implementation process; user training as well as user participation design**.

Venkatesh and Davis (2000) sought to further identify variables that influence PU as this particular factor was one of the strongest determinants of actual use. Their research led to the creation of **TAM2** - extended version of TAM (see figure below).

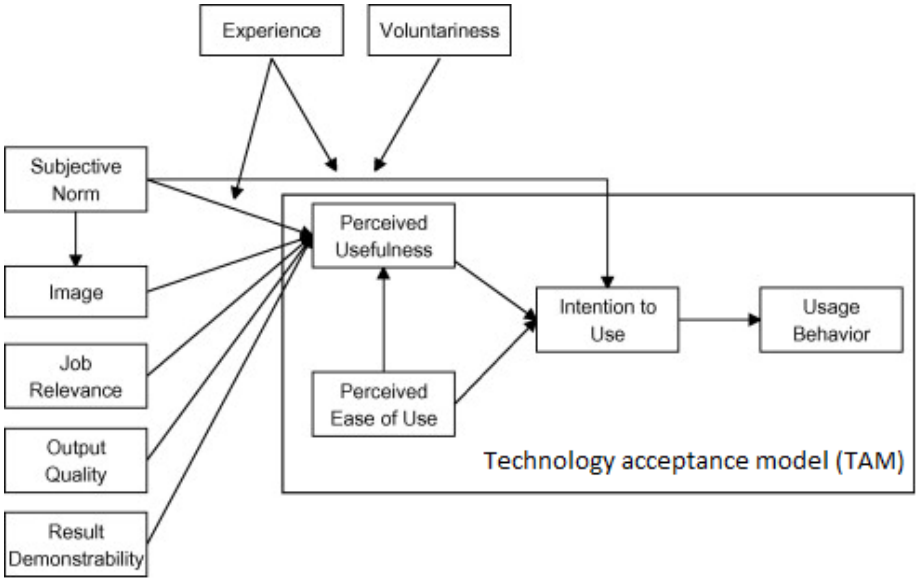

Figure 3 TAM 2. Source: Venkatesh and Davis 2000

There are five distinguish variables influencing PU:

subjective norm: other peoples influence on user's decision whether to use or not to use the system;

Image: (also influenced by subjective norm) user's desire to maintain a beneficial status among others;

job relevance: the degree to which the technology can be used for a particular task;

output quality: whether or not the system performed the task and how good was it executed.

result demonstrability: ability to produce tangible results.

Previous experience and level of voluntariness were also introduced as moderators for the subjective norm.

The model was tested in the form of longitudinal research in both voluntary and involuntary environments. All added variables and their relationships were confirmed. What is more, subjective norm, PU and PEOU proved to be direct determinants of intention to use (behavioural intention in TAM, figure 2)

3.2.3 Unified Theory of Acceptance and Use of Technology (UTAUT & UTAUT2 model)

In their paper, Venkatesh, Morris, Davis and Davis (2003) empirically compare eight separate models describing technology acceptance, each with different sets of determinants. Those models were then unified to create the unified theory of acceptance and use of technology (**UTAUT**).

The model draws from similarities across models of (1) theory of reasoned action,(2) the technology acceptance model, (3) the motivational model, (4) the theory of planned behavior, (5) model combining the technology acceptance model and the theory of planned behavior, (6) the model of PC utilization,(7) the innovation diffusion theory, and (8) the social cognitive theory.

Each model was tested using longitudinal field studies across four organizations splitting participants into voluntary and mandatory partakers. The results showed that the identified seventeen core concepts explain 17 to 53 percent of discrepancy in the user's intention to use information technologies.

Authors selected four constructs out of seven with a significant direct influence on intention or usage. **Performance expectancy** refers to users expectation of the system to perform its job. **Effort**

expectancy is defined as the degree of easiness expected by the user whilst using the technology. **Social Influence** describes how important user think, other people, believe they should use the technology. Lastly, **facilitating conditions** is the belief user have about having an appropriate organizational and technical infrastructure to support the system. Variables such as gender, age, experience and voluntariness of use are used as moderators between constructs and behavioural intention as well as usage (see figure below)

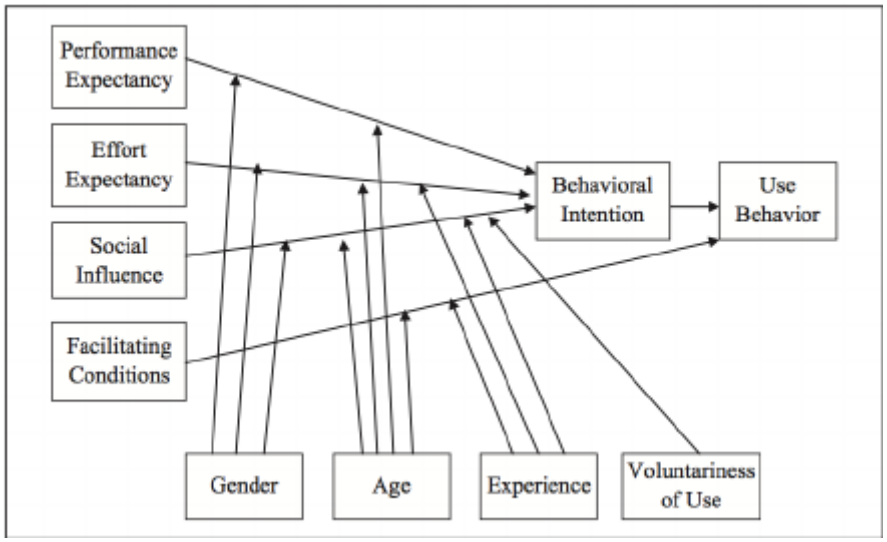

Figure 4 UTAUT Model. Source: Venkatesh et al. 2003

Venkatesh et al. (2003) tested their model and concluded it outperforms previously mentioned individual models, explaining up to 70 percent of variations in the user's intention to use information technologies.

Venkatesh, Thong and Xu (2012) expanded the UTAUT model further by identifying key additional constructs to be integrated. **UTAUT 2.** Extra concepts include **hedonic motivation**, defined as a feeling of fun and pleasure user experience as a consequence of using technology; **price value**, referring to monetary costs private user bear in connection to using technology; and **experience/habits** of the user. The model including relationships between variables can be seen in the figure below.

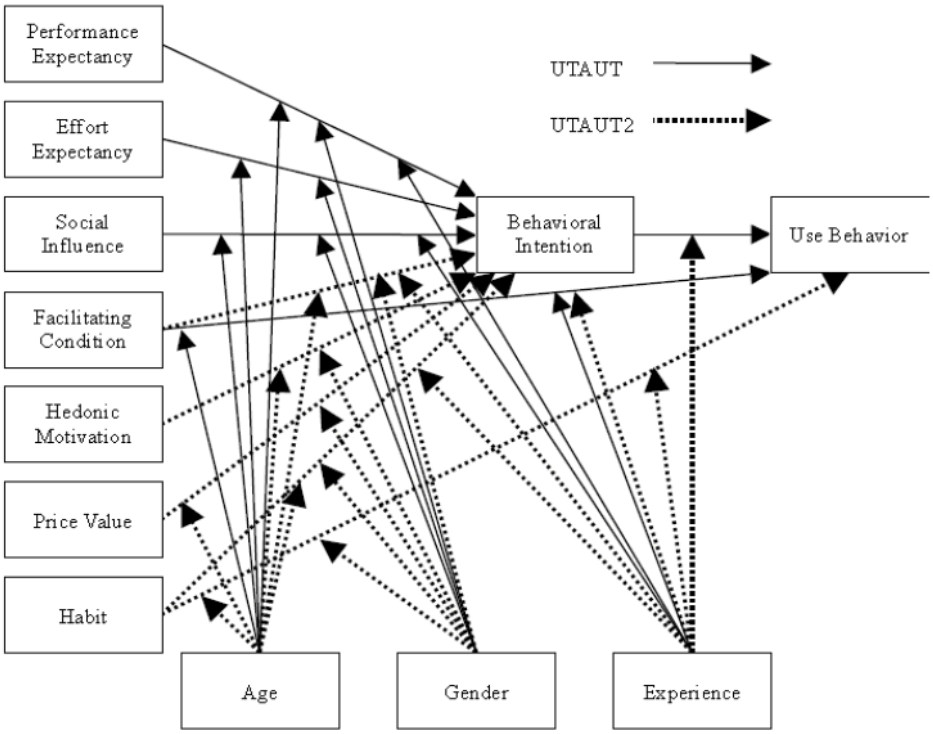

Figure 5 UTAUT 2 Model. Source: Venkatesh, Thong and Xu (2012)

The new model was tested on mobile internet technology users in Hong Kong where the use of mobile internet was a voluntary choice. Results supported newly created relations within new constructs. A substantial rise in the variance explanation in behavioural intention and technology use was noted, making the new version more comprehensive.

3.2.3 CAT MODEL

This model, unlike many other models which focused on cognition, incorporates affect perspective as well. Kulviwat, Bruner II, Kumar, Nasco, & Clark, T. (2007) combined PAD paradigm with the TAM model, creating a new theoretical framework the Consumer Acceptance of Technology **(CAT)** model.

PAD theory was developed by Mehrabian and Russell (1974 as cited in Kulviwat et al., 2007). It stated that all the user's emotional responses to their environment- both physical and social - can be captured with the following dimensions: pleasure, arousal and dominance.

Pleasure refers to a level of enjoyable reaction (such as happiness, enjoyment, satisfaction); **arousal** describes mental awareness and physical activity an individual might experience while exposed to

some stimuli; **dominance** is defined as the extent to which user feels in control of stimuli. Graphic representation of the CAT model can be seen n the figure below.

Figure 6 CAT Model. Source: Kulviwat, Bruner II, Kumar, Nasco, & Clark, T. 2007

Results supported the majority of relations presented in the model. Overall CAT explained over 50 percent of the variance in adoption intention. Three out of four variables (pleasure and arousal) from PAD paradigm were significant determinants of attitude towards adoption. Dominance did not have a significant influence on attitude, but the author suggests that it might have to do with testing settings, rather than the online experience itself. Settings might have not invoked types of emotions, usually related to this particular variable. **Pleasure** and **arousal** respectively come right after **PU** which had the most impact on attitude concept.

3.2.4 CRITICISM

TAM model has been criticised for oversimplifying human motivation behind technology adoption. Bagozzi (2007), claims that one model could not possibly explain behaviours and decisions in every adoption scenario. He points out that although there are many papers expanding on the model by adding additional determinants, there are not a lot of them that explain each construct (e.g PEOU) in more detail. Venkatesh (2003) did introduce additional variables (age, gender,experience) serving as

moderators of however they lack the theoretical background of "why?" were they chosen (Bagozzi, 2007). Benbasat andBarki (2007) highlighted the fact that with the amount of papers adding to already established TAM, the model lost its usefulness. As such, researcher can simply used "preferred version" , matching their particular paper.

Another weak point of TAM model is its lack of social and cultural aspect of the decision making process (Bagozzi, 2007). Culture have always helped with shaping human behaviour and therefore its a very important has an undenying impact on any kind of decision we make. Thus it also impacts conceptualisation of all the variables in TAM. The UTAU includes social influence as a variable however it refers mostly to interpersonal influence rather than society as a whole.

Lastly Bagozzi (2007) indicated that TAM considers usage behaviour as an ultimate end goal. Creator of the model fail to recognise that the usage in itself can, and often is, a means to achieve other, more fundamental goal therefore, usage as a goal achieving gap is completely neglected in the model. Intention of use is also often made prior taking an action often with significant time gap between the two, meaning that there are a lot more reasons that can formulate behind the actual usage.

04. METHODOLOGY

In this chapter, I present methods used to gather and analyse my data. Starting with a list of methods that were taken into consideration while planning my work, I include their description, benefits and disadvantages. Further I show chosen approach to finally introduce employed procedure in details.

4.1. CONSIDERATIONS

4.1.1. SURVEY

The purpose of a survey is to gather quantitative data, describing specific aspects of the studied population (Fowler 2009). Main method of data collection includes asking questions. A researcher can either ask questions personally or deploy a self-completion questionnaire which is completed by respondents themselves (Bryman, 2012). Questions can be open - where participants are free to reply however they choose - or closed, where subjects need to choose their answer form the set of fixed options.

Self-completion questionnaires are often quicker and more convenient for potential participants. They can decide when to fill it out and do it from the comfort of their homes if they want to (Bryman, 2012).

On the other hand, survey as a method has some disadvantages. Unlike during an interview, there is no possibility of further probing for a more detailed explanation of the given answer. It also cannot be too long as a responder might get tired and abandon the survey before finishing. Participant might read the questionnaire as a whole before attempting to answer meaning that their answer might be biased from the beginning (Bryman, 2012).

Other difficulties might arise when the answer of the participant is not truthful. Geisen and Bergstrom (2017) refer to it as a measurement error. This error has multiple variations depends on what causes this situation. In an interview where it is the researcher reading out the question, lie may occur due to the way question was presented (vocabulary, tone of voice, manner of speaking). This scenario author called the **interviewer error**. When the respondent is not truthful due to survey wording, design, it is the **instrument error**. Another type of measurement error is connected with participants ability to recall certain situations (e.g remembering how many times did they visit a doctor in the last six months). Interpretation also influences the answer (e.g some people would, for example, include seeing a nurse as a doctor visit.). This error is regarded as a **respondent error**. Lastly, **mode effects error** depends on the device choice. The layout of the device can prevent them from seeing all the answers at once and therefore he or she is not aware of other options.

When conducting the survey, researchers mostly question a sample of the whole population. It is important that the sample is representative, otherwise, the results will not reflect characteristics of the group as such (Fowler 2009). This problem can be minimized by choosing a bigger sample.

To evade some misunderstandings, Bryman (2012) advise the reader to avoid ambiguous terms and double-barred (questions asking for two things simultaneously) or long questions

4.1.2. INTERVIEW

An interview is a form of conversation with a purpose and some degree of structure (Kvale, 2008). It is is one of the most commonly used data gathering methods in qualitative research (Bryman, 2012). Unlike structured interview in quantitative research, it allows greater flexibility as it is primarily focused on the participant's perspective. It is also less restricted when it comes to time and going off-script. The interviewer would encourage, hoping to reveal what is really important to the subject, even if it was not initially considered by the researcher.

Bryman (2012) identifies two types of qualitative interview based on the interviewer's approach to the process: **unstructured interview** and **semi-structured interview**. First one is the closest to conversation form. The interviewer uses a set of prompts as a guide to a range of topics needed to be covered in the interview. Interviewee responds freely and the researcher follows up on parts that he finds worthy. In the letter one, the researcher tends to ha a list of questions regarding specific topics pre-prepared (Interview guide). Subject still has the freedom to answer whatever he or she deems appropriate. The order of questions from the guide can be changed at any time and the possibility to

use questions outside of the guide, as the follow up is available. The choice of a particular type of qualitative interview depends on a variety of factors. If for example there is more than one person conducting the interview, a semi-structured interview would be more appropriate. Interview guide would ensure cohesion amongst the responses of different interviewees.

Kvale (2008) distinguish forms of interview based on its purpose. **Factual** and **conceptual** interviews focus on facts and concepts gathering o top of their personal perspective. **Narrative** and **discursive** interviews are characterised by creating knowledge through discussion and narrative. The interviewer is actively engaging in the discourse as opposed to acting as a voiced questionnaire. Keeping in mind the number of different interview types, the author highlights that there is no set procedure to conduct interview research.

Interviews as a qualitative method can be analysed using **Thematic** or **Narrative** analysis (Bryman, 2010). Thematic analysis not easily identifiable due to lack of distinctive techniques. There is however a general strategy one can employed called Framework. Framework can be described as "*matrix based method for ordering and synthesising data*" (Ritchie et al. 2003 p.219 as cited in Bryman, 2010). Matrix consists of central themes and subthemes that are defined through rereading of interview's transcripts. Data is then organized in themes and presented in form of subthemes for each category. There is no set technique for theme identification, however, it is assumed that it reflects the researcher's "*awareness of the recurring ideas and topics in the data*"(Bryman, 2010 p. 580).

Narrative analysis is suitable for data that is sequence sensitive (stories) It focuses on how participants process what happen rather than factual description of the event. One of the danger of this approach is that researchers tend to neglect using critical thinking and treat the story as its been told.

Interview as a method has been criticised for the potential lack of reliability if conduct poorly (Bryman, 2012; Opdenakker 2006). Not being careful with formulation, the researcher can affect respondent's answers by asking leading questions.

4.1.3. USABILITY TESTING

Rubin, and Chisnell, (2008) pointed out that the term usability testing is often overused to describe any kind of method which goal is to evaluate a product or system. Goodman, Kuniavsky, and Moed, (2012) present a more accurate definition, they refer to usability testing as "*structured interviews focused on specific features in an interface prototype*"(Goodman et al, 2012 p.275).

It produces insights into how people use working system or a prototype. The test is constructed based on specific tasks, that are designed to evaluate prototypes features and therefore are not an ideal method to study experience as a whole. Bevan (2009) points out that although user experience and usability measures have no fundamental difference between them. One concentrates on user perception of the product, the second one focuses on whether or not the product is usable.

Goodman et. al (2012) further explains that a typical usability testing procedure consists of a series of tasks performed by existing or potential user of the product/service. Tasks need to be : reasonable in length and otherwise, described in terms of end goal, specific, doable, and in realistic sequence.

The session is recorded and researcher analyses the results in terms of success, mistakes and opinions. The measurements could be both qualitative and quantitative in nature. Quantitative measurements can include : time to complete the task, number of errors, number of successful task completion. Qualitative data can be extracted through asking questions during the test.

When it comes to how many participants should evaluate the product/service, there is no clear answer. Rubin, and Chisnell (2008) argue that larger groups still deliver valuable results , especially for more complicated tests. However, they also say that after around eight to nine users the reported problems become highly repetitive. The strategy recommended was to follow Jacob NIelsen study and recruiter at least 5 participants.

Authors suggests that a usability test can be performed even in the earliest stage of product/service development. Nano- usability test is an example of such an early test. It requires a minimal number of resources as it does not require specialised equipment or specific settings. Participant is asked to perform a task, preferable something they care about and the researcher limits themselves to berly observing the interaction. In this case participants are drawn from among family or friend group, thats why its a variation of Family Usability Testing. The results will not be detailed bu will give researcher an initial overview of potential issues and user general experience.

A Micro-usability test is closer to a proper evaluation. It still takes a short amount of time, however it does require to recruit participants that fall into the potential target group of the product/service. This way the test becomes more objective. A list of tasks should also be created based on main functionalities of the product/service.

Maurer and Ghanam (2008) also introduced a simplified usability testing that, similarly to previous ones, does not require specialised equipment. Discount usability testing relies on techniques such as think aloud, heuristic evaluation, card sorting, scenarios or walkthroughs making it a little more advanced. Think aloud prompts the user to voice their thoughts as they go through tasks. This allows researchers to get insights into why participants used the system in a particular way. On the other hand the participantøs might decrease as they need to divide their attention. In heuristic evaluation more than one participant is involved. During the process they analyse the interface separately based on predefined principles and combine the results, identifying usability issues. Card sorting can help with uncovering potential user's mental information architecture . It can be used as basis for menu navigation evaluation.

Overall discount usability testing is easy and quick to conduct. It does not require experts or even actual end users. The recruitment process is very flexible.

On the other hand, oversimplification of the tests can produce partial results. Moreover, involving participants outside of user group might lead to implementation of features that are normally not desired.

4.1.4. EYE-TRACKING

Eye tracking is a technique used in usability testing that follows and measures different attributes of human eye including where the person looked and in which order (Poole & Ball, 2006). What one is looking at is a subconscious decision, indicating what is important to he or she. The method can therefore provide insights into one's cognitive processes such as problem solving, reasoning and search strategies.

At the same the, eye tracking uses quantitative measures to analyse the results, making the method objective. Authors highlight four groups of measurement: fixation, saccades, scan paths and blink rate together with pupil size.

Fixation indicate instances when the eye is relatively stable, focusing on a particular element. Measurements in this category can tell the researcher which parts of the website for example are more important for the user (fixation per area of interest) or if they have potential difficulties extracting information (fixation duration). Fixation spatial density can demonstrate searching process quality (fixations in small area -> efficient search; evenly spread fixations -> inefficient search).

Saccades, are jerky movements of both eyes simultaneously, always followed by fixation (Goldberg and Wichansky, 2003). This measurement can indicate a change in observer goals or the fact that the server does not match hers or his expectation. It occurs if the difference between two neighbour saccades is more than 90 degrees (Poole & Ball, 2006).

Scan paths describes the whole sequence of movements needed for task completion. Its duration, length, density and direction, demonstrate searching efficiency as well as strategy.

Finally, **blink rate and pupil size** can help interpreting cognitive workload of the observer however those particular reactions can be caused by external factors.

The measurements are taken using a specialty equipment that can either be mounted on top of an observer's head or placed independently from them. The first set up is preferable when tasks require substantial amount of head motion whereas the second one is typically used for usability evaluations (Goldberg and Wichansky, 2003).

No matter the choice of set up, eye tracking technology poses some challenges. Poole and Bali (2006).Point towards eyewear and contact lenses as something that prevents an accurate measures of eye movement. The equipment is also constantly calibrated for new users to optimise the tracking (Goldberg and Wichansky, 2003).

Poole and Bali (2006) also turn reader's attention to the importance of test design in ensuring reliability of the study. If a researcher is interested in analysing fixations, he or she needs to calibrate the equipment accordingly but also define the minimum time of fixation itself. Further, defined area

of interest should be large enough to register all relevant eye movement Lastly, tsks for participants should be well-defined for eye motion to be accurately matched to actual cognitive processing.

4.1.5. CONTENT ANALYSIS

According to Bryman (2012), content analysis is an approach to analysing various texts. It's objective is to systematically quantify content using a predetermined manner.

Content analysis allows researchers to counted different components of the text, namely (1) significant actors -main figures of a particular interview; (2) words; (3) subjects and themes - phenomena categorization; (4) dispositions - sentiments of the text as a whole, particular phenomena or an individual actor.

The methods is very transparent with clear coding scheme. Its unobtrusive, meaning that there is no need for a researcher to interfere with the study, which otherwise could lead to participants altering their behaviour. Finally it is highly flexible, making it suitable to apply to a wide range of various types of unstructured and structured texts. Text data can be obtained via interview/focus group responses, open survey questions, observation or through printd/digital media (Kondracki & Wellman, 2002 as cited in Hsieh and Shannon 2005).

Next to its advantages, content analysis has been criticized for a number of its shortcomings. The quality of analysis depends heavily on the quality of sources used for the process which means it will be only as good as selected texts. Developing a coding scheme its never fully free of subjective interpretation of the coder. Lastly, as a quantitative method, it focuses strongly on measurement, which could lead to highlighting what is measurable rather what is truly important.

The procedure consists of seven steps (Kaid,1989 as cited in Hsieh & Shannon, 2005). Formulation of the research question determines the goal of the research, types of media and text used. One needs to then select an appropriate sample for analysis and define potential categories. Designing coding process is the next step, followed by coding implementation, determining process trustworthiness and finally analysing coding results.

Hsieh and Shannon (2005) propose three approaches to content analysis: conventional, directed, or summative.

Conventional content analysis is often used in research aiming to to describe a phenomena, which have a limited knowledge available in terms of theory or academic literature. In this case, researchers formulate coding categories based on the data. If data was gathered through interview/open question

survey, they read data repeatedly to gain an overall understanding of source as a whole. Using notes, they record their first impressions, thoughts and initial analysis transforming into initial coding scheme as the procedure continuous. Codes are sorted into categories and appropriate definitions for each category, subcategory as well as the individual codes are developed (Morse & Field, 1995 as cited in Hsieh and Shannon, 2005).

Direct content analysis has a more structured procedure. It relies on already existing theory and researchers to develop initial coding categories (Potter & Levine-Donnerstein, 1999). Any part of the text that does not fit into any of the initial categories is given a new code. Hsieh and Shannon, (2005) suggest that in order to create more trustworthiness in owns study, one must be sure to cover every possible occurrence (e.g. of the phenomenon emotional reaction)

Summative content analysis is an approach where the researcher starts quantifying specific words aiming to understand its contextual use. It also expands to interpretation of the word rather thatn relying solely on quantitative measures. This analysis is typically applied to manuscripts, journals and specific content in textbooks.

The main difference between all approaches development of initial coding. While choosing the appropriate approach, one must remember the purpose of the study and proceed with the most appropriate version of content analysis.

4.2. RESEARCH DESIGN

Putting my problem statement in the centre of research design, I followed the **pragmatic paradigm**. According to Mackenzie & Knipe (2006), paradigm creates the basis for choices regarding methods and research design as a whole. Pragmatism grants researchers freedom to choose methods and procedures that are best suitable to gather appropriate data to solve the problem (Creswell, 2014). It opens the possibility of a mixed-method approach that is used I use in this paper.

Mixed method choice is also supported by picking a singular **case study** of VELUX company as my framework for research design. Case studies are often associated with qualitative research but

Bryman (2012) poses that there are many instances when both approaches were employed.

Mixing qualitative and quantitative methods gives a more detailed, reliable insight into the problem, to some degree eliminating each other's weaknesses (Creswell, 2014). The author describes three individual models of mix-method approach: one where researcher merges qualitative and quantitative methods simultaneously (**convergent parallel mixed methods**); one where he or she starts with quantitative methods to then follow up the results with a qualitative approach (**explanatory sequential mixed methods**); and one where approaches are reversed (**exploratory sequential mixed**

methods). I decided to hold both my methods at roughly the same time hoping to get as the most comprehensive understanding of the problem.

The main source for answering the **first two RQ** was **desk research**. Combining findings from literature review and theory section I extracted number of factors that served the basis for designing further methods.

In order to gather insights needed for **RQ3** I deployed a combination of methods. Starting with **usability testing**, I used it as a tool for acquainting participants with the VELUX website. Performing a series of tasks they had a chance to explore the webpage and use the chatbot tool. Although eye tracking would have provided a valuable, more exhaustive view on how the website its used, it was omitted due to time and equipment limitations. Based on the usability session they fill out a series of **surveys** indicating their experience with the website. Finally I conducted a short **interview** with each of the user to investigate topics related to the problem statement but also to follow up on the usability testing. The interview provided me with the qualitative overview for the same questions.

I use **thematic analysis** to examine the outcomes of the interview and descriptive statistics to summarise the survey results. Thematic and content analysis are very similar approaches as they have the same aim. However, thematic analysis allows researchers to explore the meaning of participants responses within a particular context (Vaismoradi, Turunen & Bondas, 2013). This is not the case with content analysis. Here researchers focus of occurrence frequency, often disregarding context, and therefore risking the change in meaning of particular response.

The recordings of user testing will be evaluated by both thematic analysis and **descriptive statistics**. I use thematic analysis to categorise user voice comments and descriptive statistic will give a numerical indication of web performance. The visualisation of the whole research design can be found in Figure 7 below

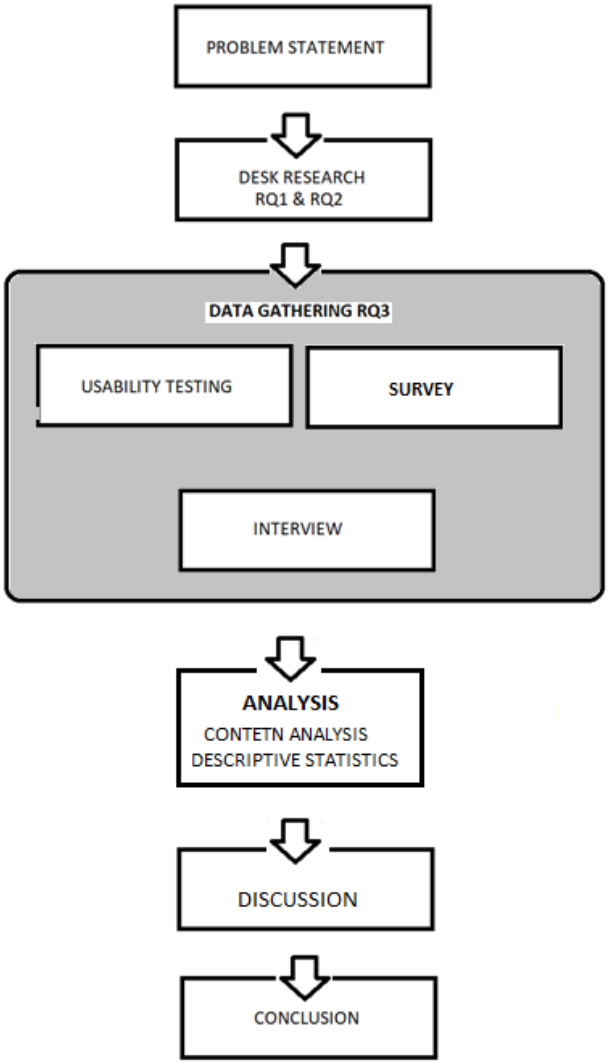

Figure 7 Visualisation of the research design applied in this paper

4.3 PARTICIPANTS

In order for research to represent the potential respondent's point of view, participants were recruited based on a predefined target audience of the VELUX website. They are defined as United Kingdom homeowners between the ages of 35 and 65. I further limit the pool of subject by involving only the main decision-maker per household as he or she would most likely browse the internet in search for renovation inspirations. In total, I collected data from five individual users all found through personal connections. A demographic overview of each subject is presented below including their gender, age and the city where their home is located.

	Gernder	Age	City
User 1	*Female*	*38*	*Hamilton*
User2	*Male*	*70*	*Eastkilbride*
User 3	*Male*	*53*	*Eastkilbride*
User 4	*Female*	*52*	*Blantyre*
User 5	*Female*	*41*	*Edinburgh*

Table 4, Participants demographic information: gender, age, city location of their owed houses. *Source: own creation*.

One of the participants is outside of scope when it comes to the age, however he is still a house owner and do use computer/internet on daily basis. With the shortage of volunteers, I decided to involve him in my study.

4.4 PROCEDURE

Data gathering for this project occur through a singular meeting with each participant, during which, all the methods were employed. As each of them intertwine with each other, I decided to present the procedure for the process as a whole, rather than describing it individually for each method.

The individual sessions took place in subjects home using equipment provided by me. For maximum privacy, trying to eliminate unnecessary interruptions, we opted for a separate room in case there were other people present. Each participants were introduced to the process before the start of the test.

Opening with the pre- test survey, they proceed with trying to find the information they were asked about. Users were asked if they prefer to read the questions themselves or they want me to do it for them. This differed from user to user. They were also offered a possibility of using the mouse if they were uncomfortable using a touchpad. After completing each scenario, I deployed scenario related surveys for them to fill out. Surveys were presented digitally on an ipad so that users did not have to exit VELUX website to answer them. Next step involved a longer more detailed survey after which user could take a 10 min break. Not everyone used this opportunity. While on the break I had time to save the video recording and change the equipment for conducting interviews. Finally I end with short interviews for which I used my interview guide.

Overall, the whole process was rather smooth and without major issues, technical or otherwise. There was no time constraint on study duration however I tried to fit within one hour or one hour fifteen minutes.

4.5 . QUANTITATIVE METHODS

4.5.1 USABILITY TESTING

Development

When designing the usability testing tasks, I kept in mind that this particular method was supposed to serve as a tool for participants to get familiar with the website and use the chatbot tool. I created individual exercises based on three web usage scenarios: buying a roof window, buying a roof blind and converting the window into an electrical one as an upgrade to the already purchased manual one.

As velux.co.uk is a marketing site, it does not allow visitors to purchase the actual product, instead redirecting the to an appropriate velux online store. Thus, assignments, centered around finding relevant information, needed before an actual purchase of each item.

To ensure subjects use the chatbot tool, answers to some of the tasks (delivery time and installation costs, installation duration) could only be obtained via VELUX Virtual Assistant. Other information can be found within the website.

Another way I made sure participants have a chance to use the chatbot was informing them that if they were stuck they could ask me for a hint. When that happen I pointed them towards the tool. Not to lead users on, I only did that when they specifically ask for it.

The process of designing specific question was performed in collaboration with a person who falls into target audience category. I consult the subject in terms of what kind of information would be of interest if they were to accomplish the goals of all three scenarios.

Tasks from window conversion category were designed to be the hardest ones to accomplish. The information available on the website were limited and chatbot could not provide concrete replies. With this, I wanted to expose the participant to some degree of a negative experience and see if there are any changes in their attitude towards the website/chatbot. The topic was chosen based on chatbot performance data available in my organization. Although the specific numbers are confidential, I looked for a topic that had the biggest difference between the number of times it appears in conversation and fulfillment rate, given by the user at the end of interaction.

Initially the conversion scenario included blinds and shutters instead of widow however during the pilot test - conducted with a user within the target group - it appears to be less complicated than I originally anticipated, thus the change was made. All scenarios can be seen in Appendix 4.

Materials

Usability testing was performed on LENOVO Yoga 13 computer with Windows 8.1. system. For audio and screen recording I used Free Cam 8.- a free screen recording and video editing software.

Validity & Reliability

Validity of usability testing relies on whether or not, the test measures what it was supposed to measure (Nielsen,1994). As the test was suppose to measure potential/existing user online experience, I involved participants from the right target group. It was the first step towards ensuring validity. I was also available for any questions and clarifications if needed to make sure the exercises were understandable and the end goal of each of them understood.

On the other hand, the fact that I was assisting during the process might have lower the validity as participants might feel tense or use my guidance in different stages of the test. However, completion of the task was not the main focus of the session.

Reliability is harder to define since there are many individual differences between participants (Nielsen, 1994). In case of larger group of users, one could use standard statistical test to determine the significance of the results, however, with only five subjects, it was impossible.

4.5.2 SURVEYS

Development

I developed five separate surveys, each collecting data on different parts of the test. Pre test survey; three scenario specific surveys and a post test survey. I created this distinction to minimise potential bias. As participants were presented with the third senario, they would be more like to judge the online experience, the most recent interaction will be the most vivi in their memory. There was a risiko that intentional negative experience might skew the final outcome. All questionnaires were created using Google Form (Appendix 5)

Questions were divided based on topics they covered. With the exception of pre-test survey, each of them contained questions regarding overall online experience and chatbot. They were presented in the form of question matrix, requiring users to rate individual factors on 5 point Likert scale where 5 stand for "*Strongly Agree*" and 1 stand for "*Strongly Disagree*"; 3 indicated neutral attitude. I choose the factors based on my literature review and theory chapters.

In the table below I present an overview of factors splitted between pre-defined concepts:

Online customer experience	Chatbot evaluation	Technology acceptance theory
Information quality	*Efficiency*	*Preceived usefulness*
Creadibility	*Effectivness*	*Preceived easy of use*
Telepresence	*Performance*	*Pleasure*
Concentration	*Affect*	*Arousal*
Enjoyement	*Satisfaction*	
Engagement		
Web Skills		
Challenge		
Intteractivity		
Interactive speed		
Control		

Table 6 Overview of factors devided by the source. Source: Own creation

To create questions that correspond with each factors I used already existing inquiries from papers published in those fields also used in in my desk research. Questions regarding online experience were taken from Novak et. al (2000) and questions for technology acceptance form Kulvivat et. al (2007). Questions regarding chatbot evaluation factors were created based on a combination of different papers from the field. I then transformed those questions into statements in order to fit the Likert scale.

During the process I quickly discover that many factors from chatbot evaluation and technology acceptance theory were either very similar or identical , thus number of questions does not match the number of factors from the table.

Additionally I asked participants about they web usage, online shopping habits and familiarity with the VELUX brand/website. Those questions were designed to investigate their previous experience level.

After the pilot test, the order of some statements was change to create a better flow, for example statement about information quality was placed at the end together with statement about the brand credibility. Wording also has been adjusted. Word arousal was exchange with excitement.

As the post test survey inquire about the session as a whole, more specific questions about chatbot were asked (conversational flow, esthetics). Those were not present in the other, scenario specific surveys. Being the last survey in the test, I wanted to get a more detailed overview of participants attitudes and satisfaction.

Materials

Survey were presented to subjects using Ipad MIni with IOS 11.3.1 operating system.

Validity & Reliability

In order for surveys to be reliable, they need consistency within the answers (Bryman, 2010). All the questions related to usability testing were given the same 5 point measurement scale in order for results to be comparable. Furthermore, all scenario specific surveys had the same questions to create unison and once again secure comparability of the results. Using questions created and used in other research within relevant fields, strengthen my validity as well. Chosen measurement were already employed and shown their relevance to the topics, thus ensuring that what needs to be measured, was in fact measured.

4.6 QUALITATIVE METHODS

4.6.1. INTERVIEWS

Development

I conducted an interview a short interview with all the participants in order to obtain a better understanding of their experience, which help me to formulate better response for the question posed in the main problem statement.

I asked about their familiarity with the chatbot and previous experiences with that type of tool. How do they see it affecting the overall online experience and what factors are important for them personally when using a website. Lastly, I gave them the definition of Flow and once I made sure they understood the concept, probe about the personal experience with the phenomena. All questions are available in interview guide (Appendix 6)

I decide to include the flow question in the interview only as the precise definition was quite wordy and might discourage participants, causing them to lose focus while answering further questions. Furthermore, Flow refers to a session as a whole thus there was no reason for asking about it after each scenario.

Additionally if any particular answer from the survey caught my attention I encouraged an explanation from the interviewee.

Although interview guide has a particular order, during the pilot testing it was apparent that grouping questions according to themes is easier, less confusing to participant as he or she does not have to switch rapidly between topics.

Bryman (2010) advise to avoid questions that can be answered with “yes” or “no”. I did not manage to bypass it completely. However every time I was met with such answer I asked additional probing questions.

Materials

For recording audio from the interviews, I used Sound Recorder that was a built-in feature in my laptop.

4.7. ETHICAL CONSIDERATIONS

Ethical issues can arise at any stage of the research. To avoid any potential problems I follow principles introduced in The Belmont Report. The report describes three principles: respect for the person, beneficence, and justice (Belmont Report, 1979 as cited in Bordens & Abbott, 2002).

Respect refers to the participant being a volunteer that upon taking part in the research was fully informed. To ensure fulfilment of this principle, I informed subjects about the nature of the study and what data will be collected and shared during each step of the research. I made an overall introduction and a quick summary before conducting each method.

Beneficence principle was created to make participants feel comfortable by minimizing hazards and maximizing the benefit. In this study each volunteer has been anonymised. Although I knew all of them, names were omitted in the audio recording as well as in the surveys. Furthermore the personal data gathering was limited only to general information, ensuring that the person cannot be recognised.

The principle of **justice** divides the burden and the costs of potential benefits to the research equally between participant and the researcher. In my study, users were offered an immediate assistance in case they have some doubts at any point of the research. I encouraged them to ask for help when they get stuck or feel confused. I also schedule a mini break in between different methods that could give the subject ability to relax and give me opportunity to save any data from previous tests.

05. ANALYSIS AND RESULTS

In this chapter I describe my process for analysing gathered data. Chapter is divided based on the methods I choose and each process description is followed by presentation of the corresponding results.

5.1 USABILITY TESTING

5.1.1. ANALYSIS

As mentioned before, the usability testing was created to give subjects aa chance to familiarize themselves with the website and use the chatbot tool as opposed to identifying potential usability issues. Since the goal of the testing is different, I will not strictly follow all the common measures present in usability testing. Instead, I decided to adjust those metrics to fit my test design. My analysis will focus on the following aspects:

How long was each session?

How many times did they ask for guidance/ seem confused

For how many tasks did they turn to the chatbot for help?

The duration of the session will serve as an indicator of the confusion rate with the assumption that the longer the session, the more complicated the site proved to be. This measurement excludes times they took to fill out appropriate questionnaires. Some of the recordings were paused for this period but so,e of the kept running.

Number of instances they ask for guidance will give me insights into the number of potential failures they could have had otherwise which would ultimately spoil the customer experience

Lastly, I am interested in the number of tasks during which they used or attempt to use chatbot tool. This could help me asses how and if they are likely to use the tool again.

Even without the interest actual task completion, usability analysis includes collecting, organising observation in order to extract potential emerging trends. As a moderator I made initial notes, to built upon after re watching each interview a couple of times.

It is important to mention that data of the second user is incomplete due to technical difficulties encountered during the testing. After many attempts I managed to partially recover initial twenty minutes of video recording, therefore information provided in the following result section describes only the beginning of the testing.

Recorded materials available in Appendix 7.

5.1.2. RESULTS

I present the results of each usability testing in a form of comment summary highlighting the three previously mentioned aspects and additional observations.

User 1

User took advantage of chatbot tool total of three times throughout the 29 minutes long session. She did not go back to the tool for all the remaining task after chatbot discovery. Additionally she left the page through redirecting links, partially because of web design and partially because she looked for the answers in the wrong area of the main site. At some point I needed to help her come back on track.

When looking for information, she used an in-page search engine and FAQ section.

Examples of venturing into website's areas unrelated with topics of the task include: going to replacement section for the price of a new window and searching for installation guides in the support section.

User 2

His session took took 53 minutes in total and the analysed part was exactly 20 min and 46 seconds. User interact with chatbot only once during that period of time and having an opportunity to use it again he instead turn to in-page search engine. Reason being that he did not realise that he can come back. When I went to explain that that's the case he explained that the outline of the screen recording used prevent him from seeing that that was a possibility.

User 3

During the 33 minute of the session he used chatbot three times. Aside from pointing out the existence of VELUX Virtual Assistant I did not have to help him anymore. As a first impulse, similarly to user number 2 looked for in-page search engines. Unlike others he found delivery information among content of the web shop , instead of the tool. He did however use it to look for a price of specific type of blind.

In search for installation guide, he ended up in the professional part of the website. He was an exception when it comes to scenario as he found all the information without guidance either from me or chatbot support.

User 4

It took 41 minutes for this user to finalise the usability testing engaging with the chatbot four times throughout. The first task of the first scenario took 14 minutes to accomplish as she misread the

instructions. I needed to clarify any confusion on the spot and make sure the final goal of each task is understood.

When wanted to use chatbot for the first time she looked for it in a contact section. She justify her choices, saying that its natural for her as she prefers to get in touch via telephone or email. This applies also when looking for the price information as normally she would message the company asking for prices estimate for specific projects. She did not avoided the tool all together. As she used it for example to search for information on black out blinds.

User 5

In this case, session ended after approximately 37 minutes with four instances of using chatbot.

The search started with dealer search which she performed using in-page search engine as she did not recognize tasks that came before it.

While looking for information she ended up outside of VELUX website. Once hatbot was introduced she did not come back to it with the next task and instead attempted to find it on her own.

When she got frustrated she returned to it after some difficulty of finding it again.

She asked for information using a combination of individual words (Keywords) treating chatbot tool as an online search engine rather than AI tool.

5.2. SURVEY

5.2.1 ANALYSIS

Data collected through surveys in Google Forms provides an overview of the results in forms of graphs and pie charts. For the sake of better readability, and understanding I transformed these visualisations into appropriate data tables which served as a basis for result descriptions. Some of those tables I converted into better, easier to comprehend graphs for improved reading flow through better visualisation. Although I planned to convert data using Tableau program, due to technicall difficulties connect with the software constantly freezing during the process, I change to Microsoft Excel.

Initially, I hoped to use descriptive statistics however with only five participants I decided that numerical summary description would be more beneficial. My aim is to get an overview of combined attitude towards OCX on VELUX website. Aside from the examination of the results across all the subjects, I tried to detect potential changes in participant perception via individual user data comparison however changes were minor do I opted for pointing to a specific user when needed

5.2.2 RESULTS

	User 1	User 2	User 3	User 4	User 5
How often do you use the internet	*Daily*	*Daily*	*Daily*	*Daily*	*Daily*
Do you shop online?	*Regularly*	*Regularly*	*Regularly*	*Sometimes*	*Sometimes*
Have you ever shop online for products related to building/repairing the house?	*Sometimes*	*Sometimes*	*Sometimes*	*Sometimes*	*Sometimes*
Do you know the VELUX brand?	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Have you previously visited www.velux.co.uk or any VELUX site?	*No*	*No*	*Yes*	*No*	*Yes*

Table 7 a table summary of the Pre-Test survey showcasing participants online habits and familiarity with VELUX brand. Source: Own creation

All of the participants use the internet on a daily basis and tend to shop online either regularly or from time to time. Each of them bought a home renovation related product at least once. When it comes to brand awareness, all of them are familiar with VELUX and two out of five visited VELUX website in the past.

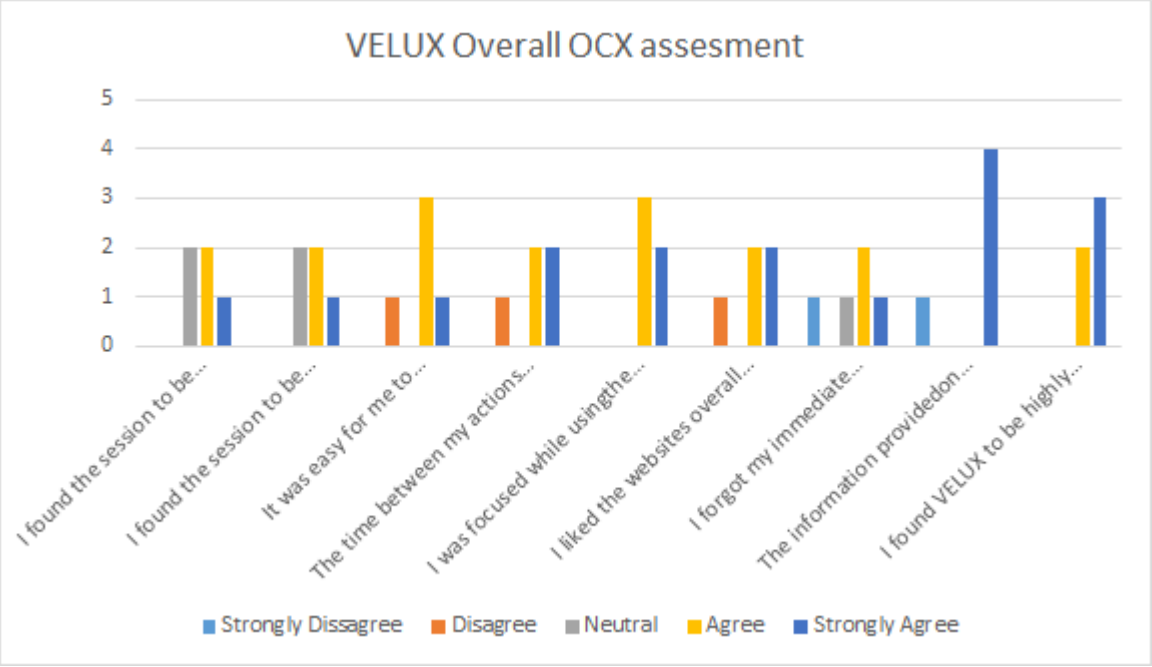

Figure 8 Overview of overall online customer experience assessment of VELUX site. Source Own creation

When it comes to overall online experience assessment, statements regarding session being stimulating and enjoyable have some neutral ratings. Only one participant rated site performance (ease of navigation, responsiveness) and aesthetics as not satisfactory, unlike others who rated the same factors as satisfactory.

Almost all of them (with an exception of one) found information quality being at a good level. Similarly, they all think of VELUX as a credible brand and were focused while using the website. Focus, however,

did not translate into the feeling of forgetting the immediate surroundings. Here the opinions were divided the most with two of the participants assessing it at neutral or very poor (User 5) level.

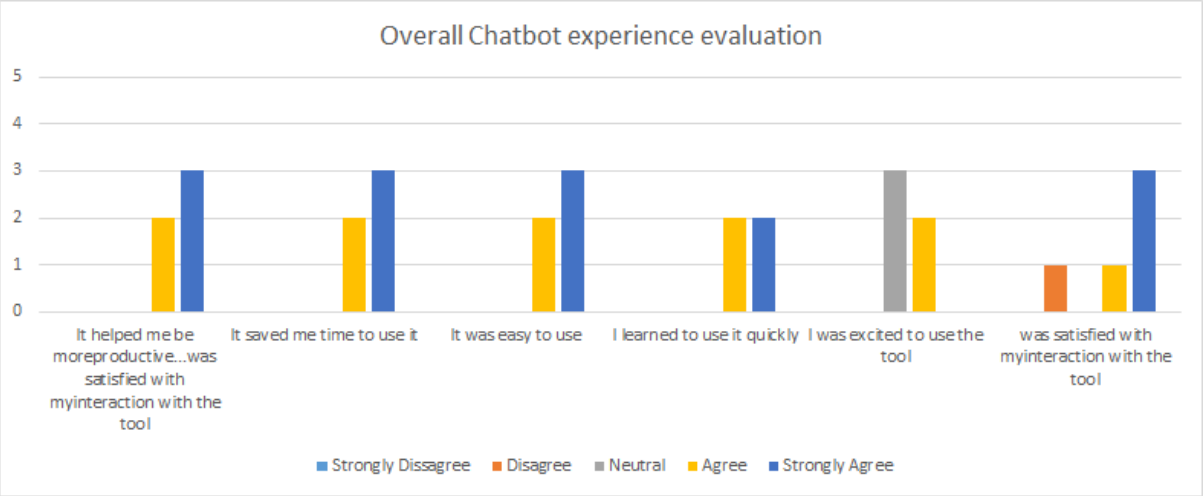

Figure 9 overall assessment of chatbot experience on VELUX site. Source Own creation

Opinions regarding Velux Digital Assistant are more consistent. Subjects agree that the chatbot helps them with finding the right information faster, was easy to learn and to use. Most of them, however, were neutral when it comes to excitement connected with using the tool.

Finally, when it comes to the overall satisfaction of interaction with the Chatbot, User 5 was not satisfied at all as opposed to the majority of the group that rated it on the highest level. However, if we divide the interaction further into satisfaction with responses, conversational flow and design none of them was rated lower than neutral see the table below

	User 1	User 2	User 3	User 4	User 5
Response satisfaction	*4*	*4*	*5*	*5*	*4*
Conversational Flow	*5*	*5*	*3*	*5*	*3*
Chatbot design	*5*	*5*	*4*	*5*	*4*
Was it easy to find?	*Yes*	*Yes*	*No*	*Yes*	*Yes*

Table 8 Assessment of VELUX Digital Assistance's characteristics on 5 point scale where 1= Very Bad; 5= Very good. Source: Own creation

Above data describes participant's opinion of their overall experience about the session, including all three scenarios. Below, I present graphs specific for each scenario individually.

Scenario 1: Buying a roof window

During this scenario, participants had to fulfil four individual tasks two of which could be accomplished only via VELUX Digital assistant. Overall users found all the information they were looking for, they were satisfied with its quality and were focused during task performance. Perception of challenges was mostly neutral with two exceptions, one of which completely disagree with this and one that claims the opposite. A similar pattern can be noticed when talking about enjoyement. Majority of users agreed with the statement with two exceptions, one of each disagreed and one remained neutral with their opinion (see figures below).

Chatbot performance in this particular scenario was asses positively with almost all of the statements. One of the participants did not find it easy to use (see figures below).

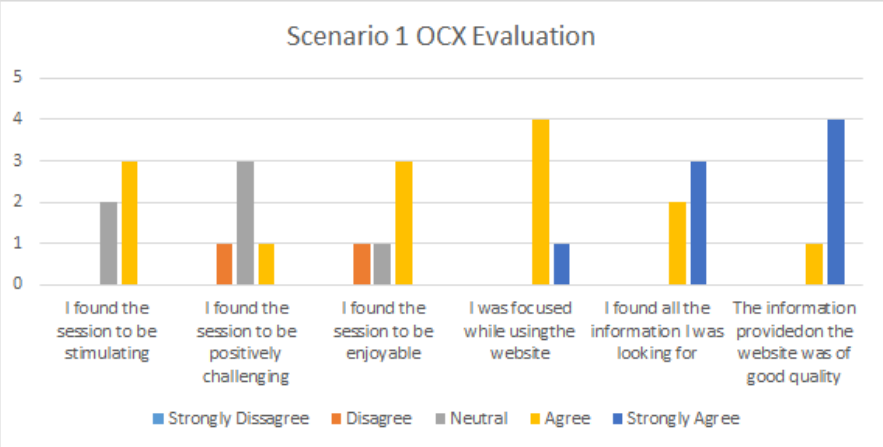

Figure 10 Summary of Customer Experience during the first scenario. Source: Own creation

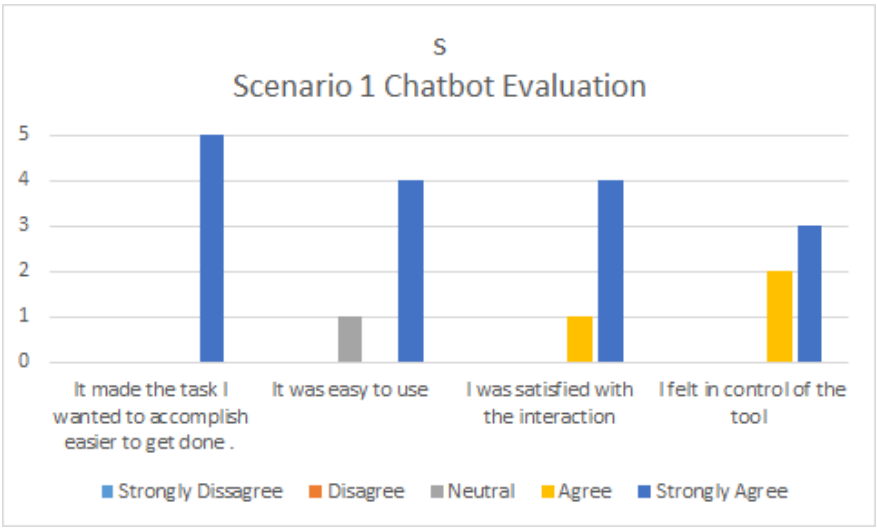

Figure 11 Summary of chatbot performance evaluation during the first scenario. Source: Own creation

Scenario 2: Buying a blind

The scenario included three tasks from which one could be fulfilled only through using the chatbot. Here level of stimulation and enjoyment were spread out across the whole scale ranging from clear disagreement to strong support. Focus and quality of information available remain the best-rated factors of online customer experience. Assessment of challenge level was still strongly neutral and this instance information quality also earned a neutral reaction (see figures below).
Chatbot evaluation, on the other hand, presented absolute unison among the subjects. Almost of chatbot related statements were rated at the highest level with one user giving them the second-best score (see figures below).

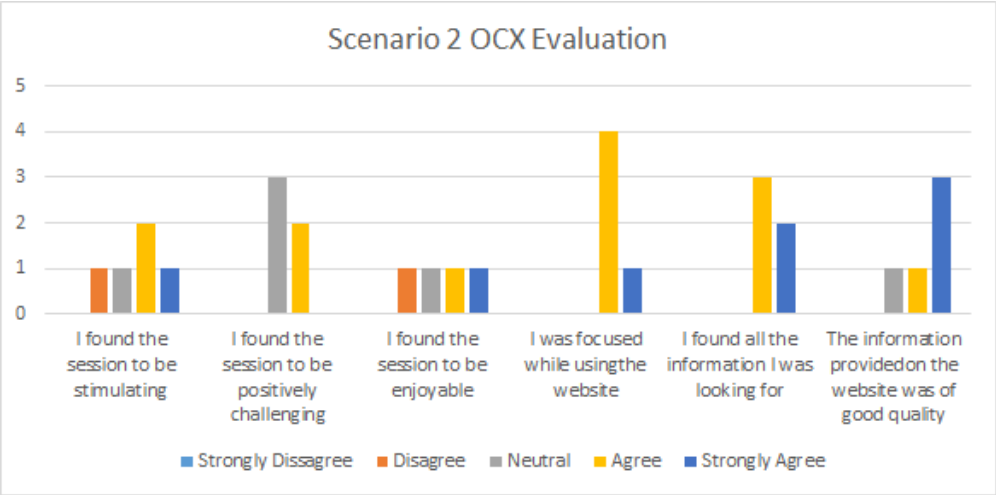

Figure 12 Summary of Customer Experience during the second scenario. Source: Own creation

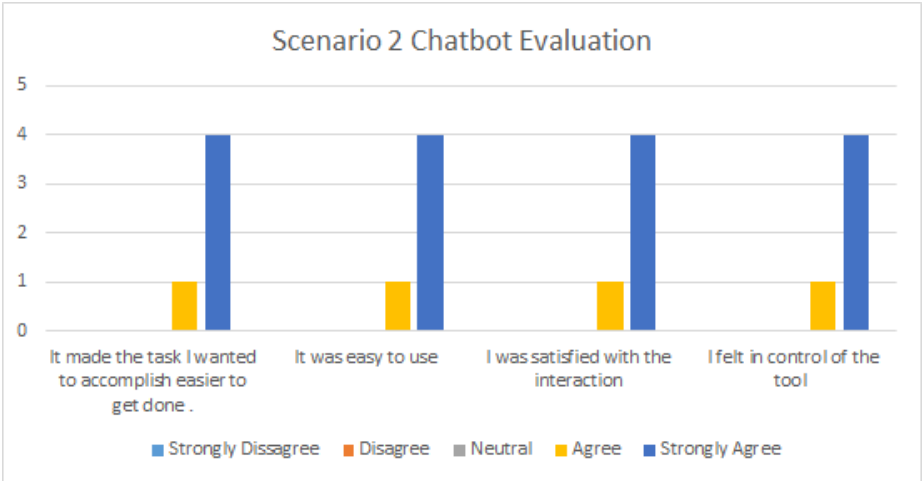

Figure 13 Summary of chatbot performance evaluation during the second scenario. Source: Own creation

Scenario 3: Converting the window

The last scenario was designed to expose participants to the less favourable website experience. Answers for all three tasks could not be found through chatbot conversation and were harder to find (but not impossible) while searching directly on the web.

Although in the end each task was completed there is a shift in the rate proportion. Here most participants only agreed with the statement as opposed to “Strongly agree” like in the previous scenarios. At the same time while asking for the perceived level of positive challenge it drifts towards negative response being the first and only survey where the lowest score was assigned. Perceived quality of information also declined. Although the neutral rating of this factor appeared in the second scenario, it was accompanied by the majority of the users giving it the highest score; whereas here, other scores are dispatched equally between “Agree” and “Strongly Agree” (see figures below).

When it comes to chatbot performance, there is a big change in score proportions. Starting with the shift of the opinion towards the second-best score, from the highest ratings in previous scenarios. Going further, there is a clear indication of dissatisfaction as two ratings lower than “Neutral” appeared. One of the participants does not feel that the tool make hs search easier or that they have control over the tool at all (see figures below).

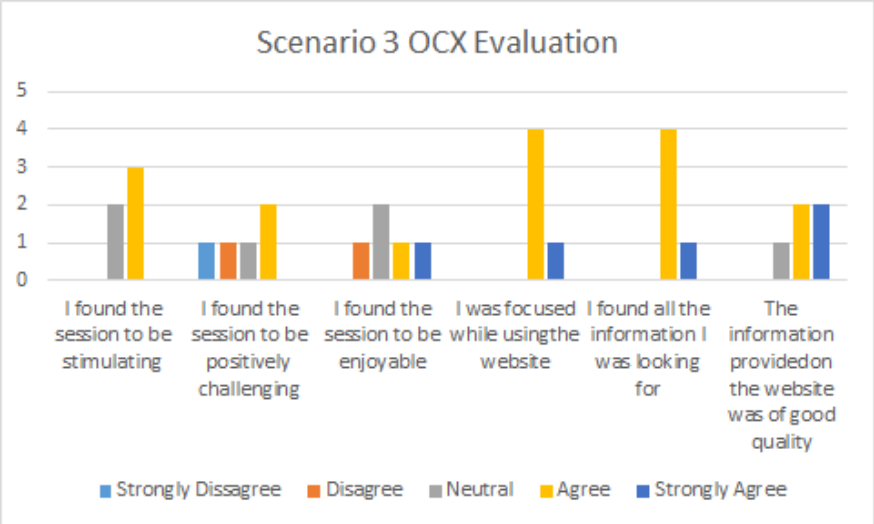

Figure 15 Summary of Customer Experience during the third scenario. Source: Own creation

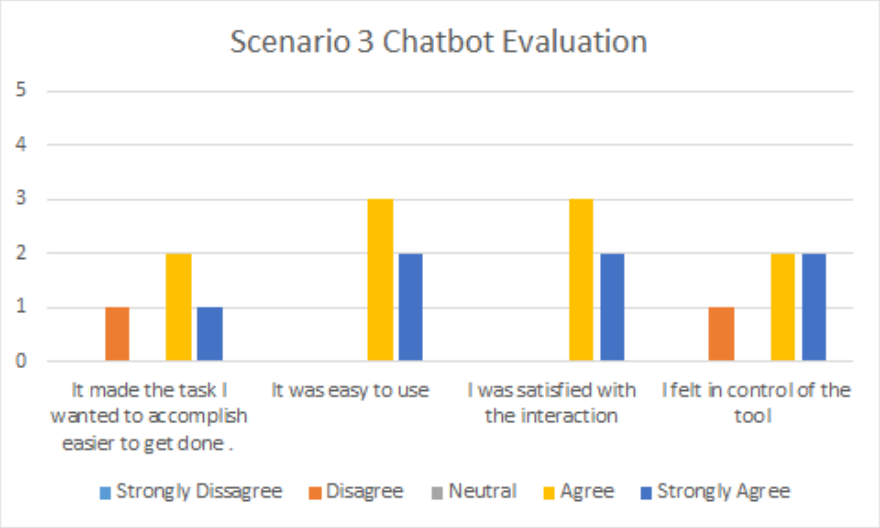

Figure 16 Summary of chatbot performance evaluation during the third scenario. Source: Own creation

5.3 INTERVIEW

5.3.1 ANALYSIS

To analyse the interviews, I followed the Framework format proposed in Bryman (2010).

Interviews were first transcribed with the assistance of a free web app "*oTranscribe*". I identified five individual themes that I further divided into the corresponding sub-themes (see table below).

Theme	Sub-theme
DETERMINANTS OF GOOD ONLINE EXPERENCE	*Navigation*
	Overall design/Esthetics
	Content
CHATBOT USEFULNESS	*Performance*
	Efficiency
	Quality of the response
ATTITUDES TOWARDS CHATBOT PERFORMANCE	*Positive performance*
	Negative performance
PRIOR CHATBOT EXPERIENCES	*Situations*
	Purpose
	Experience assesment
FLOW	*Determinnants*
	Situational occurance

Table 9 Overview of identified themes and sub-themes from the interview's transcripts.

Determinants of good experience refer to what users see as an important factor for creating a good online experience. Answers from that category were further deconstructed into navigation-related factors, overall design characteristics and content which concerns kind of information participants expect to see on the page.

Chatbot usefulness contains statements about the benefits of having the chatbot on the website. Here performance describes whether or not the chatbot fulfils its purpose, efficiency refers to a speed with which these purposes were achieved, information quality refers to user satisfaction with the answer.

Attitudes towards chatbot performance include snippets participants view on chatbot influence on the OCX based on its performance as well as examples of bad and positive chatbot performance.

Prior chatbot experiences were divided into type of situation subject used the chatbot tool before for what purpose and was he or she satisfied.

Lastly, I identified the flow category and decided to split it into what are the kind of situations that participants feel themselves n the zone and did it occur at any point during testing.

In order to create the distinction, I listen and reread transcripts multiple times and adjust them accordingly. To increase the validity of the analysis I invited an outside person, not familiar with the project asking for the second opinion. After the co-creator proposed their solution we compare it against mine. Overall the topics, although phase differently mostly overlapped. However, "Chatbot usefulness" theme was redefined as the coding of the other researcher was better organised.

Once I decided my categories I classified snippets of the interviews accordingly (See appendix 9 and 8 dor Audio)

5.3.2. RESULTS

Only one participant was not at all familiar with the chatbot concept at all. Additionally one user although familiar with a tool as such could not identify whether or not they ever used one. Others used it for banking or mobile services or simply to look for information.

User	Timestamp	Citation
User 1	00:21	*"particularly on my banking and like mobile "*
User 2	00:24 00:55	"not familiar at all with a chatbot, not that format" "*we've got ALEXA but its more google search(..) more familiar with that."*
User 4	00:17	*"yes but not on that site with I think a live person rather than automatic"*

When asked about online customer experience and its determinants, participants named easy navigation as their first priority other characteristics include overall looks of the website, information presentation. Some were more specific, listing kind of information they would like to see and where.

User	Timestamp	Citation
User 1	*03:30*	*"look of the website writing and colours clear and homepage not to busy and information across the bar to give links with a drop-down, the bottom of the website has web mar and have the contact us section at the bottom"*
User 4	04:41	*"easy navigation (...) somewhere where you can go and look for advice(...)also graphics as it makes it easier to follow"*

User 5	*04:55*	*" finding things easy, having site map (...)not having too much text (...)"*

Another identified theme regards participants attitude towards chatbot performance. Here Participants agreed that as long as its working, it gives an additional benefit to the website and OCX.

User	Timestamp	Citation
User 2	*02:35*	*"it's essential (...)once you got used to it"*
User 4	03:24	*"if you've got a website and have a chatbot and its works(...) it adds you extra benefit"*

On the other hand, if the tool fails to perform, it can become really frustrating. Such situations can include referring the user to the wrong part of the website, feed them irrelevant information. In general not fulfilling user expectations.

User	Timestamp	Citation
User 5	*00:07*	*"(...) but it was a horrible experience (...)because it didn't really seem to satisfy my needs"*

At the same time even when the chatbot has problems with accomplishing given tasks, the impact, although negative, seem to be smaller. Participants point out that if they need something, they could always go back and search for information themselves thus the negative impact is reduced.

User	Timestamp	Citation
User 3	01:10	*"if it doesn't perform then you usually know pretty quickly if it is not gonna perform. Then you jus go back to interact with the page there is a point when I think you will get to. "*

User 4	02:00 03:24	*"(...)think I wouldn't be able to get the information () I would think () negatively impact the website"* *"(…)and it's not good i would say oh they've got a chatbot but its not very good you better just do that "*
User 5	03:06	*"if it's an easy website to navigate then it doesn't matter, but if it's a hard website s this one, then I'm completely dissatisfied"*

Finally we talked about the flow concept. I presented participants with definition and inquired whether or not they ever experience it during web exploration. They give me descriptions of certain situations, emphasising for example the fact that they need to be interested in the topic or that the search process needs to smooth. User 5 provided examples of specific websites

User	Timestamp	Citation
User 2	*05:40*	*"particularly is something interesting me not necessary for buying online shopping "*
User 3	08:13	*(…) website (...) leads me down the path...fairly quickly, i dont have time to focus o anything else.*
User 5	08:10	*" i have that flow with facebook, with gmail otherwise ikea "*

As a parting question I asked users if at any point during their session with VELUX website they could recognise themselves as being in a state of Flow. Almost all, identified feeling the flow at least during some parts of the session.

User	Timestamp	Citation
User 1	*06:36*	*"yes I think so I think when you have focused coz all of the info that you need to look for ...it was easy to just get cought"*
User 2	-	-
User 3	*09:51*	*"the only part was a black out blinds I suppose and the last one (...) which was where I expected it to be. (...) I was in the zone"*
User 4	*09:00*	*" probably not as much probably because I know I was being watched but if i had to do it myself then I might get into it a little bit more"*
User 5	*08:49*	*"a bit in the first and a little second as well"*

06. DISCUSSION AND CONCLUSION

This paper investigates the impact of a chatbot implementation on the overall online customer experience. Results gathered throughout the report suggest that chatbot can have a strong positive impact on OCX. In order to facilitate a positive experience, chatbot should be able to provide users with adequate information in a fast and efficient manner. If those conditions are not met, there is a risk of negative effect, influencing visitor's perception of the webpage.

Nevertheless, negative experience is not equal in weight with a positive one. All participants agreed that if the chatbot does not meet their expectations in terms of information delivery, they would simply return to navigating website themselves. Thus the perception of a disappointing interaction is easier to forget.

When it comes to choice of data, this study relies heavily on Flow and Technology acceptance theories. Some of the factors introduced in both models proved to be more influential than others. Previous experience mentioned in Venkatesh and Davis (2000) as a moderator between *social norms* and *intention to use* (TAM2 model) appear to have an active impact on User's 5 perception of chatbot

performance who already had a very negative experience using the tool. When confronted with chatbot's poor performance she rated it harsher than other participants. Additionally her overall introduction with VELUX Virtual Assistant seemed a little apprehensive from the beginning based on her rushing typing and lack of focused on what she was writing.

Another example of factor importance is the case of perceived usefulness and perceived ease of use. Both aspects are present in all variations of Technology Acceptance model. In the study chatbot usefulness, as a quick, easy to learn an alternative to in-page search, was highlighted by all of the participants.

It is important to remember however when talking about subjective perception there are a lot of additional, hidden factors that influence each of the mentioned components. Bagozzi (2007) warns that with so many unknown elements, we would never be able to fit them into one, ultimatum model.

Overall neutral level of excitement over new tool as well as neutral attitude towards the level of existing challenge and enjoyment could be explained using the work of Hoffman et al. (2002). In his paper, he describes how the level of task importance to oneself influence leads to greater attention and that leads to achieving the sate of flow. Putting it simply, even if participant enrolment is voluntary, the inability of choosing your own task will decrease the overall enjoyment amongst subjects. Alternatively, handed out tasks are not challenging enough.

Quality of gathered data, its reliability and validity should also be taken into consideration while discussing the outcomes of one's paper. Starting with the size of my sample, this research was built upon five participants (User 2) with one of them not fitting fully into the description of VELUX's main target group. Although it is said that five people are enough to expose most of the major issues hindering users from fully enjoy their experience (Nielsen as cited in Rubin and Chisnell, 2008), I entertained the thought that it might be too little. As part of the data from the second user is missing the data might have decreased because of that.

While creating the research design for this paper, due to the time constraint I decided to gather all types of my data simultaneously from one participant at a time. This resulted in research ´where gathering process of each type of data is strictly dependable from one another. The problem with it can be demonstrated using my usability testing scenarios. At the time of designing scenarios, I already planned to use it as an opening for suggesting chatbot usage. In the scenario, right after that participants, were assumed to come back to the tool with a list of tasks chatbot had full capacity to answer. At this point users expectations rise and when presented with something that is hard to find and VELUX assistant do not provide expected support, potential flustration can rise higher than expected. Hence, the evaluation unison on the second scenario.

If not for the time limitation, this paper could also have been more coherent in its research design. Knowing that I have no time in between one method and the other, surveys and interview guide were created simultaneously risking disconnect between one area and the other. In the ideal world, the interview guide would have been created on a survey result of a bigger scale.

Another potential issue is employing usability testing as a basis for creating an experience for further evaluation. From the time perspective, an easier, more relaxed settings would be better to facilitate more natural behaviour resulting in a more accurate depiction of the everyday user.

Implementation of the Likert scale to unify the measurements across different sources was not a bad experience at all however if I was to redo the study I would introduce a wider scale to ensure bigger variability in options. I found it challenging at times describing results as change between Agreed and strongly agreed can prove hard to explain.

Despite its challenges participants were mostly able to achieve flow at some point during their interaction with VELUX website. Chatbot contributes to it by making the search process easier faster and more accurate. Creating navigation paths that could be convenient for users to follow. Seeing as potential mishaps will not have equally strong effects, my recommendation is to look into your user needs and expectations as a basis for future chatbot development.

07 REFERENCES

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. MIS quarterly, 665-694.

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. Journal of the association for information systems, 8(4), 3.

Benbasat, I., & Barki, H. (2007). Quo vadis TAM?. Journal of the association for information systems, 8(4), 7.

Bevan, N. (2009, August). What is the difference between the purpose of usability and user experience evaluation methods. In Proceedings of the Workshop UXEM (Vol. 9, pp. 1-4).

Bordens, K. S., & Abbott, B. B. (2002). Research design and methods: A process approach. McGraw-Hill.(pp. 202)

Brandtzaeg, P. B., & Følstad, A. (2017, November). Why people use chatbots. In International Conference on Internet Science (pp. 377-392). Springer, Cham

Brandtzaeg, P. B., & Følstad, A. (2017, November). Why people use chatbots. In International Conference on Internet Science (pp. 377-392). Springer, Cham.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: changing user needs and motivations. interactions, 25(5), 38-43.

Bryman, A. (2012). Social Research Methods. 4th ed. New York: Oxford University Press. Chapters 3, 7-8, 10, 13, 15, 17-18, 20, 24 and 27

Cahn, J. (2017). CHATBOT: Architecture, design, & development. University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science (pp.5-7).

Chakrabarti, C., & Luger, G. F. (2013, May). A framework for simulating and evaluating artificial chatter bot conversations. In The Twenty-Sixth International FLAIRS Conference

Choi, D. H., Kim, J., & Kim, S. H. (2007). ERP training with a web-based electronic learning system: The flow theory perspective. International Journal of Human-Computer Studies, 65(3), 223-243.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in society, 1(1), 104

Creswell, John W., and J. David Creswell. Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications, 2017.

Cronin, P., Ryan, F., & Coughlan, M. (2008). Undertaking a literature review: a step-by-step approach. British journal of nursing, 17(1), 38-43.

Csikszentmihalyi, M. (2014). Flow and the foundations of positive psychology: the collected works of Mihaly Csikszentmihalyi. New York: Springer Science Business Media Dordrecht pp (209-279).

Dale, R. (2016). The return of the chatbots. Natural Language Engineering, 22(5), 811-817.

Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: Theory and results (Doctoral dissertation, Massachusetts Institute of Technology) (PP.1-115)

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly, 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. Management science, 35(8), 982-1003.

Fowler Jr, F. J. (2009). Survey research methods. Sage publications.

Geisen, E., & Bergstrom, J. R. (2017).Usability testing for survey research. Morgan Kaufmann.

Gentile, C., Spiller, N., & Noci, G. (2007). How to sustain the customer experience:: An overview of experience components that co-create value with the customer. European management journal, 25(5), 395-410.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In The Mind's Eye (pp. 493-516). North-Holland

Goodman, E., Kuniavsky, M., & Moed, A. (2012). Observing the user experience: A practitioner's guide to user research. Elsevier. Chapter 2 and 11

Grewal, D., Levy, M., & Kumar, V. (2009). Customer experience management in retailing: an organizing framework. Journal of retailing, 85(1), 1-14

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environments: Conceptual foundations. Journal of marketing, 60(3), 50-68.

Hoffman, D. L., & Novak, T. P. (2009). Flow online: lessons learned and future prospects. Journal of interactive marketing, 23(1), 23-34.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

Hsu, C. L., & Lu, H. P. (2004). Why do people play on-line games? An extended TAM with social influences and flow experience. Information & management, 41(7), 853-868.

Huang, M. H. (2003). Designing website attributes to induce experiential encounters. Computers in human behavior, 19(4), 425-442.

Hung, V., Elvir, M., Gonzalez, A., & DeMara, R. (2009, October). Towards a method for evaluating naturalness in conversational dialog systems. In 2009 IEEE International Conference on Systems, Man and Cybernetics (pp. 1236-1241). IEEE.

Jwalapuram, P. (2017, September). Evaluating Dialogs based on Grice's Maxims. In Proceedings of the Student Research Workshop associated with RANLP (pp. 17-24).

Kaid, L. L., & Wadsworth, A. J. (1989). Content analysis. Measurement of communication behavior, 197-217.

Khan, R., & Das, A. (2018). Introduction to chatbots. In Build Better Chatbots (pp. 1-13). Apress, Berkeley, CA

Klaus, P. P., & Maklan, S. (2013). Towards a better measure of customer experience. International Journal of Market Research, 55(2), 227-246.

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017, June). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In Proceedings of the 2017 Conference on Designing Interactive Systems (pp. 555-565). ACM.

Kondracki, N. L., Wellman, N. S., & Amundson, D. R. (2002). Content analysis: Review of methods and their applications in nutrition education. Journal of nutrition education and behavior, 34(4), 224-230

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. Information systems research, 13(2), 205-223.

Kuligowska, K. (2015). Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. Professionals Center for Business Research, 2.

Kulviwat, S., Bruner II, G. C., Kumar, A., Nasco, S. A., & Clark, T. (2007). Toward a unified theory of consumer acceptance technology.Psychology & Marketing,24(12), 1059-1084.

Kvale, S. (2008).Doing interviews. Sage.

Mackenzie, N., & Knipe, S. (2006). Research dilemmas: Paradigms, methods and methodology. Issues in educational research, 16(2), 193-205.

Mathieson, K. (1991). Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. Information systems research, 2(3), 173-191.

Maurer, F., & Ghanam, Y. (2008). Discount Usability Testing. University of Calgary.

Meyer, C., & Schwager, A. (2007). Understanding customer experience. Harvard business review, 85(2), 116.

Novak, T. P., Hoffman, D. L., & Duhachek, A. (2003). The influence of goal-directed and experiential activities on online flow experiences. Journal of consumer psychology, 13(1-2), 3-16.

Novak, T. P., Hoffman, D. L., & Yung, Y. F. (2000). Measuring the customer experience in online environments: A structural modeling approach. Marketing science, 19(1), 22-42

Opdenakker, R. (2006, September). Advantages and disadvantages of four interview techniques in qualitative research. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 7, No. 4).

Peras, Dijana. "Chatbot evaluation metrics." Economic and Social Development: Book of Proceedings (2018): 89-97.

Poole, A., & Ball, L. J. (2006). Eye Tracking in HCI and Usability Research. In C. Ghaoui (Ed.), Encyclopedia of Human Computer Interaction (pp. 211-219).

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis.

Randolph, J. J. (2009). A guide to writing the dissertation literature review. Practical assessment, research & evaluation, 14(13), 1-13

Rose, S., Clark, M., Samouel, P., & Hair, N. (2012). Online customer experience in e-retailing: an empirical model of antecedents and outcomes. Journal of Retailing, 88(2), 308-322.

Rowley, J., & Slack, F. (2004). Conducting a literature review. Management research news, 27(6), 31-39.

Rubin, J., & Chisnell, D. (2008). Handbook of usability testing: how to plan, design and conduct effective tests. John Wiley & Sons. Chapter 2,3 and 13

Sanchez-Franco, M. J. (2006). Exploring the influence of gender on the web usage via partial least squares. Behaviour & Information Technology, 25(1), 19-36.

Shawar, B. A., & Atwell, E. (2007, April). Different measurements metrics to evaluate a chatbot system. In Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies (pp. 89-96). Association for Computational Linguistics.

Shawar, B. A., & Atwell, E. (2007, January). Chatbots: are they really useful?. In Ldv forum (Vol. 22, No. 1, pp. 29-49)

Title

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing & health sciences, 15(3), 398-405.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., ... & Goel, R. (2018). On evaluating and comparing conversational agents. arXiv preprint arXiv:1801.03625, 4, 60-68

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management science, 46(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS quarterly, 425-478.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS quarterly, 36(1), 157-178.

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. Journal of retailing, 85(1), 31-41

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. arXiv preprint cmp-lg/9704004.

Wang, Y. J., Hernandez, M. D., & Minor, M. S. (2010). Web aesthetics effects on perceived online service quality and satisfaction in an e-tail environment: The moderating role of purchase task. Journal of Business Research, 63(9-10), 935-942

Wilks, Y. (2010, July). Is a Companion a distinctive kind of relationship with a machine?. In Proceedings of the 2010 Workshop on Companionable Dialogue Systems (pp. 13-18). Association for Computational Linguistics.

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on Facebook. Cyberpsychology, Behavior, and Social Networking, 21(8), 491-497.

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to a chatbot on Facebook. Cyberpsychology, Behavior, and Social Networking, 21(8), 491-497.

-