
The User Experience of Chatbots

A Design Science Approach

By

Zeljko Maric

Copenhagen Business School

Thesis submitted in partial fulfilment
of the requirements for the Degree of
Master of Science in
Business Administration and Information Systems
E-business

Supervised by: Torkil Clemmensen

March 15, 2018

Number of pages: 66

Abstract

Even though Chatbots have been a research subject since the early 1960s when Joseph Weizenbaum developed the first text-based bot named ELIZA there is little research on the user experience of chatbots. Possible reasons are the overemphasis of HCI scholars on graphical user interfaces (Følstad & Brandtzaeg, 2017) and the fact that most of the mainstream messaging platforms like Facebook Messenger opened their platforms for developers just in the last two years. The question what a good chatbot user experience entails remains mostly unanswered. This thesis wants to make a first contribution in answering this question by following a Design-Science Research (DSR) approach. A chatbot prototype was build and then user tested with 10 students (age 22-29) through formative usability testing to explore factors that influence a good user experience. Additionally debriefing interviews were led to explore unvoiced aspects during the user test. Based on the insights from both user testing and debriefing interviews, user experience guidelines were formulated. First results indicate the users' overall preference for human-like characteristics in a bot, but users also reported that some specific bot-like characteristics made the conversation preferable to a human-to-human interaction. Among other things participants highlighted the non-judgemental space provided by a bot that made them feel safe to voice otherwise unvoiced thoughts.

Table of Content

Abstract	1
Table of Content	2
List of Tables	4
List of Figures	5
1 Introduction	6
1.1 Background	6
1.2 Motivation	8
1.3 Scope	8
1.4 Objective	10
1.5 Research Question	11
1.6 Outline	11
2. Theory	12
2.1 Paradigmatic Assumptions	12
2.2 Literature Review	13
2.3 Theoretical Framework	16
2.3.1 User Experience vs. Usability	16
2.3.2 Usability Testing & Formative Evaluation	19
2.3.3 User Experience Guidelines	21
3 System Design	24
3.1 Technical and Content Design	25
3.2 Persona Design	27
4 Method	28
4.1 Research Design and Strategy	28
4.2 Research Method	30
4.3 Data Collection Setup	32
4.4 Data Source & Sample	35
4.5 Data Analysis	36
4.6 Reliability and Validity Considerations	37
5 Analysis	39
5.1 User Experience Factors	40
5.1.1 Pre-Test Questionnaire	40
5.1.2 Usability Test	42

5.1.3 Debriefing Interview	45
5.1.4 Insights	50
5.2 User Experience Guidelines for Chatbots	52
5.3 Usefulness	53
5.4 Concluding Remarks	55
6 Discussion	56
6.1 Good User Experience: Human-like but not Human	56
6.2 Contribution to Theory and Practice	61
6.3 Limitations	62
6.4 Future Research	63
7 Conclusion	64
8 References	66
9 Appendices	75
Appendix 1: Usability Test Checklist	75
Appendix 2: Consent Form	76
Appendix 3: User Scenario	77
Appendix 4: Pre-Test Questionnaire	78
Appendix 5: Post-test Questionnaire	79
Appendix 6: Usability Test Notes	80
Appendix 7: Usability Test + Debriefing Interview Summary Spreadsheet	81
Appendix 8: Pre- and Post-Test Questionnaire Summary	82
Appendix 9: Usability Recordings	89

List of Tables

Table 1. Formative and summative usability testing	21
Table 2. Design Science Research Guidelines (Hevner et al., 2004, p. 83)	30
Table 3. Data sample	36
Table 4. Negative remarks during usability testing	45
Table 5. Positive remarks during usability testing	45
Table 6. Errors and bugs during usability testing	46
Table 7. Insights gained from usability testing and debriefing interviews	51
Table 8. User experience guidelines	53

List of Figures

Figure 1. High level conversational decision tree.	27
Figure 2. Chatbot Gustav using language and slang common among students.	30
Figure 3. Usability Lab setup.	34
Figure 4. Pre-Test Questionnaire.	34
Figure 5. Usability Test evaluation.	35
Figure 6. User test summary spreadsheet loosely based on Tomer (2013).	38
Figure 7. Easy-of-use evaluation.	41
Figure 8. Prior Experience with Chatbots.	42
Figure 9. Smartphone devices used in the usability test.	43
Figure 10. Quick Replies too similar were negatively remarked (NR3)	44
Figure 11. Free writing input on chatbot Gustav.	47
Figure 12. Thematic Map “What would you improve about Gustav?”.	50
Figure 13. Thematic Map “What was the best thing about your experience with Gustav?”.	51
Figure 14. Pre-Test usefulness expectation.	55
Figure 15. Post-Test evaluation of usefulness.	55
Figure 16. Establishing interaction pattern while onboarding.	60
Figure 17. Provide variety in default messages to catch errors.	61

1 Introduction

Even though Chatbots have been a research subject since the early 1960s when Joseph Weizenbaum developed the first text-based bot named ELIZA there is little research on the user experience of chatbots. This is mostly due to the overemphasis of HCI scholars on graphical user interfaces (Følstad & Brandtzaeg, 2017) and due to the fact that most of the mainstream messaging platforms like Facebook Messenger opened their platforms for developers just in the last two years. The question what a good chatbot user experience entails remains mostly unanswered. This thesis wants to make a first contribution in answering this question by following a Design-Science Research (DSR) approach. A chatbot prototype will be build and then user tested to explore factors that make good user experience. Based on those insights User Experience Guidelines will be formulated that can help build chatbots with a good user experience.

1.1 Background

How do we want bots to chat with us? This is the question that lies at the heart of this thesis. Ever since Joseph Weizenbaum's creation of the first text-based chatbot named ELIZA at MIT in the early 1960s bots elicit a seemingly inherent human desire to be able to communicate with a non-human entity. Ironically when Weizenbaum programmed the bot he wanted to demonstrate the exact opposite. He expected to show the superficiality and impossibility of a meaningful conversation between humans and machines. In the most famous script DOCTOR, Weizenbaum attempted to replicate a conversation between a psychotherapist and a patient by using simple rules that made ELIZA answer the user's input with non-directional questions. Even though the underlying algorithms only employed simple pattern-matching and substitution methods without being able to understand the contextual frame of the input provided, Weizenbaum was surprised to see the openness of responses and how much time participants invested.

In the decades that followed chatbot creators strived for greater levels of computational abilities, that are considered a proxy for intelligence (Coniam 2008). The Turing Test became the defining benchmark for determining whether a chatbot could

exhibit intelligent behavior that could match that of a human. The test consists of a human evaluator chatting with a human and a bot without seeing which is which. Based on the conversation the evaluator has to decide for who the human is and who the bot. The exchange takes places through a text-only interface which lets control for other factors that could influence the evaluation like appearance, bodily movement, gesture, mimics, etc.. Even though the Turing Test has received a lot of criticism it remains the benchmark for evaluating chatbots in regard to their conversational capabilities. Fueled by the evolution of modern technologies like Natural Language Processing, Machine Learning and Speech Recognition chatbots moved beyond the text-only interface and span a larger portfolio of bots that are often called *automated conversational agents*. Intelligent personal assistants like Siri (Apple), Google Assistant and Amazon Alexa are among the most popular ones. Increasingly the focus of the work moved toward the capabilities of the AI engine and less on the user experience. Especially text based conversational agents deployed on popular social network messaging apps like Facebook Messenger did not receive much research attention. This is also due to the early development stage of the platform itself. For example Facebook Messenger introduced chatbots in 2016.

Aside from Weizenbaum's ELIZA anecdote there is research that points at the potential of chatbots in regard to educational and therapeutic processes. Ly et al. (2017) were able show the efficacy of a text-based chatbot in promoting well-being by providing simple educational content from positive psychology and self-help literature. Similarly the efficacy of a chatbot delivering Cognitive Behavioral Therapy (CBT) to students with symptoms of depression and anxiety was proven in a controlled trial (Fitzpatrick et al., 2017). Other research focused on measuring the efficacy of chatbots in regard to learning outcomes. Atwell's (1999) research about the the impact of speech and language technologies in English language teaching indicates the potential of a chatbot as a conversation practice partner. Coniam (2008) too emphasises the potential of chatbots for teaching english as a second language. It is only Fitzpatrick et al. (2017) and Ly et al. who (2017) touch upon the user experience in passing, but do not provide any guidelines or concrete findings of what a good user experience on educational text-based chatbots actually entails.

Hence, this thesis will build a prototype that will be user-tested in order to explore the factors influencing the user experience.

1.2 Motivation

Aside from little prior research there are different reason for researching the user experience of chatbots.

First, the main four messaging apps (WhatsApp, Facebook Messenger, WeChat, Viber) combined have more monthly active users than the main four social network platforms (Facebook, Instagram, Twitter, LinkedIn). There are two billion messages being exchanged between people and businesses each month on Facebook Messenger alone. William Meisel (2016) projects that specialised chatbots will generate global revenues of \$623 billion by 2020. Using messaging platforms like Facebook Messenger to reach their customers is becoming a standard for businesses. With a growing user base a human to human service will become increasingly expensive. Providing parts of its services by employing specialised chatbots will thus become increasingly important. But because messaging platforms come with certain user interface and user experience constraints and opportunities it is important for practitioners to understand how to build chatbots that provide a good user experience.

Second, Følstad and Brandtzaeg (2017) attest that the HCI scholars will need to adjust their theories, methods and tools with growing demand for conversational interfaces. It is especially important due to the aforementioned grow of messaging apps on which chatbots will be deployed. After HCI's long standing focus on graphical user interfaces (GUIs) this could prove more important than other technological evolutions so far. Følstad and Brandtzaeg end their paper with a call to action for HCI to start researching conversational interfaces.

1.3 Scope

On Facebook Messenger alone there are over 200.000 active chatbots according to Facebook (2018b). Given the number of different technologies (Machine Learning, Voice Recognition, NLP, NLU, etc.), platforms (Facebook Messenger, Slack, Kik, etc.) and use

cases (ecommerce, news, travel, etc.) there are many different preconditions in regard to a good user experience. Amir Shevat (2016) proposes a simple taxonomy across different axes:

- **Voice vs. text:** A voice bot works through verbal speech, whereas a text-based bot only through written text in and output.
- **Super bot vs. domain specific:** A super bot exposes multiple services (e.g. Google Assistant, Siri). A domain specific bot provides one service/product/brand (e.g. airline travel bot).
- **Business vs. consumer:** A business bot aims at providing a business process in an easy and reliable way. Communication is short and transactional. Consumer bots on the other hand provide all kinds of services that are not directly driven by business transactions, like information, news, etc.

Further bots are either flow-oriented, AI-powered, a hybrid of those two or human-supported (Tarazi, 2017). In a flow-oriented chatbot the user communicates along a predefined path that allows the user to navigate through a set of questions, options and conditions that are based on a logic tree. In an AI-powered chatbot there are no predefined paths and the user navigates through the conversation by free text writing. In a chatbot with hybrid functionality the user still navigates mainly through a set of questions and options like in a flow-oriented one, but the bot is able to handle free writing input to a certain degree. And lastly in a human-supported chatbot there are people monitoring the conversation between bot and user and jump in where needed.

This thesis will employ a *text based domain specific consumer chatbot on Facebook Messenger with hybrid functionality*. Going forward the term chatbot will refer to this definition if not stated otherwise.

Not within the scope of this thesis are several aspects: AI-powered Super Bots will not be studied. They face different user experience challenges and are according to Shevat (2016) still not developed enough as to provide valuable user experiences. Besides building a super bot is at this point not viable. Further research around the linguistic accuracy and

intricacies of chatbot messaging is out of scope and done by Coniam (2008). Also out of scope are learning theories applied, that aim at measuring the efficacy of chatbots on learning outcomes (Fryer et al., 2017; Griol & Callejas, 2012; Jia, 2009).

Given that prior research indicates that chatbots can help provide educational content that has measurable effect on mental well-being (Fitzpatrick et al., 2017; Ly et al., 2017) and that there is not enough prior research in regard to user experience factors of such chatbots, this thesis will research the user experience of a chatbot providing simple self-help content to students (20-29 years of age). The following sections will define the research objectives and main research question.

1.4 Objective

This thesis was conducted in the tradition of the Design Science research paradigm, thus the assumption was that “knowledge and understanding of a problem domain and its solution are achieved in the building and application of the designed artifact” (Hevner et al., 2004, p. 75). The following research objectives guided the research:

1. Design and build an educational chatbot prototype on Facebook Messenger for self-help that aims at teaching concepts that help with procrastination and productivity.
2. Explore the main factors influencing the user experience of an educational chatbot.
3. Evaluate the usefulness of the chatbot in regard to teaching self-help concepts.
4. Outline a set of usability guidelines that can help create better educational chatbot experiences.

For the Information Systems (IS) research community, this thesis represents a case study for how to build and research the user experience of chatbots using usability testing. The approach could serve both researchers and practitioners for how to evaluate user experiences of text based chatbots. Additionally, the usability guidelines resulting from this thesis can help practitioners create chatbots that provide a good user experience to consumers.

1.5 Research Question

This thesis will aim to answer the following research question:

What makes a good user experience for students using a text-based educational chatbot aiming at teaching self-help concepts on Facebook Messenger?

1.6 Outline

In order to be able to answer the above research question this thesis will follow this outline: Chapter 2 will set a theoretical foundation, that will show that the literature review indicates a need for HCI to rethink its theories, methods and processes when it comes to conversational interfaces like chatbots. Further, a theoretical framework will be given that will help research the user experience of chatbots. In order to be able to achieve the latter it is necessary to define what definition of *user experience* is used. After this is done, an overview of usability evaluation methods (UEM) will be presented as to define which method is appropriate to formulate user experience guidelines for chatbots. In Chapter 3 – System Design the chatbot prototype named “Gustav” will be described. This entails the content as well as the persona design. Chapter 4 will outline the methodological approach taken in this thesis. It will follow the Design Science Research approach that will employ a *formative user-testing* method. Formative user testing is mostly employed in the early stages of the product development cycle and aims at eliciting qualitative findings that can help improve the user experience (Rubin & Chisnell, 2009).

Chapter 5 and 6 will present and analyse the results of the user research phase and, based on that, try to formulate user experience guidelines for text-based chatbots providing educational content.

Chapter 7 and 8 will critically assess and discuss the results given in the chapters before. It will also take a look ahead and suggest aspects that need further research as the research around conversational interfaces generally and chatbots specifically is still in its early stages.

2. Theory

2.1 Paradigmatic Assumptions

The term “paradigm” is ubiquitously used, especially in social sciences, which can lead to confusion. For this thesis I will refer to Saunders et al.’s (2009) definition that a paradigm “is a way of examining social phenomena from which particular understandings of these phenomena can be gained and explanations attempted” (p. 118). In other words: it is a way of looking at the world and trying to make sense of what is happening in it in a particular way.

As this thesis follows a Design-Science Research (DSR) approach (see Chapter 4.1 – Research Design and Strategy) a *pragmatist philosophy* is underlying the research. As the main aim of DSR is to create design artifacts that are useful, the research should not restrict itself along strict lines of research philosophy beliefs, rather be guided by the research question itself. Depending on the question asked a multi-method approach can be most suitable to find a valid answer (Hevner et al., 2004, p. 83).

From an ontological perspective this thesis assumes both a objectivist and subjectivist stance. As will be discussed in the following chapters the concepts of *user experience* and *usability* can be understood from both viewpoints. Finding out what a good user experience is subjective and dependent on the evaluation and interpretation of the participant using the chatbot. User experience is thus an emergent factor of participants using the chatbot. Usability on the other hand is a more objective factor that emerges through observable behaviour and is thus to a certain degree measurable. Although aspects of both perspectives – objectivist and subjectivist – are relevant to this research, the main focus will lie on the subjectivist viewpoint due to the exploratory nature of the research that aims to find out what users consider a good user experience with chatbots.

From an epistemological perspective this thesis takes a interpretivist stance. The methods employed in this research are aiming at generating insights and detailed descriptions of the phenomena observed. Law-like generalisation based on statistical significance (i.e. positivist stance) will not be obtained. It also follows that the findings are

dependent on the interpretation of the researcher, which can be criticized in regard to validity and reliability aspects (see Chapter 4.6).

2.2 Literature Review

There is little prior research about the user experience of chatbots. Følstad & Brandtzaeg's (2017) paper "Chatbots and the New World of HCI" calls out for the importance of chatbots and the need for HCI researchers to adjust and further develop its methods and theories to properly cover the increasing importance of conversational interfaces after decades of focusing on graphical user interfaces (GUIs). (p. 38) They see three implications for HCI due to this development:

1. **Conversations as the object of design:** Moving away from the GUI toward conversational interfaces greatly reduces the graphical possibilities for designers. We thus need to move "from seeing design as an *explanatory task* – that is, a task of explaining to the user which content and features are available and which steps to take to reach the desired goal – to an *interpretational task* – that is, as task of understanding what the user needs and how she may best be served" (Følstad & Brandtzaeg, 2017, p. 41)
2. **The need to move from user interface design to service design:** HCI will need to move its focus away from the design of specific interfaces (e.g. different websites) towards the overall user experience across different conversations within the same messenger platform. A conversation with a friend or with a chatbot is happening within in the same platform. (Følstad & Brandtzaeg, 2017, p. 42)
3. **The need to design for interaction in networks of human and intelligent machine actors:** Conversation threads can be populated by multiple actors, thus designing for interaction in networks becomes a prominent challenge. (Følstad & Brandtzaeg, 2017, p. 43)

Most research before the uptake of mainstream platforms like Facebook Messenger took place on either website chatbots or on chatbots that were programmed with a specific graphical user interface (Xiao et al., 2008; Chang et al., 2008).

Hill et al. (2015) have compared 100 human-to-human and human-to-bot conversations along seven dimensions: words per message, words per conversation, messages per conversation, word uniqueness, use of profanity, shorthand, and emoticons. A multivariate analysis of these conversations showed that people communicated with a chatbot longer, but with shorter messages. Further most of the messages sent to bots lacked the richness of vocabulary of human-to-human messages and they showed a greater level of profanity. The results suggest that there are notable differences in the way humans communicate with chatbots, but leave out a description of how a bot should be programmed in order to provide a good user experience. Wilcox and Wilcox's (2013) chatbots have won the Loebner Prize for three consecutive years. However, the paper focused on describing the design process of building a super bot that is AI powered and supports free writing. Because designing a super bot is out of scope (see Chapter 1.3) these findings do not provide insights for designing a chatbot based on predefined conversation paths as employed in this thesis.

The potential of conversational agents for educational and self-help purposes has received more research attention. Atay et al.'s (2016) research showed the possibility of employing a smartphone-based chatbot to engage older community group members and help against age related mental diseases like dementia. Coniam (2008) evaluated the language-teaching potential of chatbots. The focus lied on evaluating the linguistic accuracy of a number of chatbots available online. The paper concluded that although progress has been made in terms of handling natural language inquiries, a robust "conversation practice machine" (Atwell, 1999, p. 24) is still not within reach. Crutzen et al. (2010) looked at how adolescents use and evaluate a chatbot that answers questions about sex, drugs and alcohol. The evaluation focused mainly on quantitative measures in regard to efficacy. Usability constructs like ease-of-use were also measured, but with no further insight or recommendation on how to design a chatbot for greater efficacy. Newer research in that regard was not found.

The research about the potential of chatbots for psychological purposes has gained more importance over the last two years. The most cited and comprehensive study about the efficacy of chatbots in treating depression and anxiety was conducted by Fitzpatrick et al. (2017). The primary aim of the research was to determine the efficacy of delivering Cognitive Behavioral Therapy (CBT) based self-help content through a chatbot (Woebot) on Facebook Messenger. Beyond the efficacy of delivering CBT content via the chatbot, Fitzpatrick et al. (2017) also explored aspects of the user experience through open-ended interview questions that were analyzed using thematic analysis as outline by Braun and Clarke (2006). The result were thematic maps sorted by frequencies that describe certain aspects of good and bad experiences with the chatbot. Among the most cited aspects contributing to a good experience were daily check-ins from the bot, that the bot showed empathy and concern and the educational video content shared with participants. Mentioned aspects of a bad experience were the fact that the bot was not able to converse naturally and free writing input threw him off track. Further it was mentioned that the bot was not good at handling errors, thus once off track he repeated himself often. Generally repetition and looping of messages were described as a bad experience. Although Fitzpatrick et al. (2017) did analyze user experience aspects, the main research objective remained the efficacy of delivering CBT content through a chatbot.

The literature review showed that even though chatbot user experience aspects have been studied, the existing research so far falls short on achieving the research objectives laid out for this thesis. There are several reasons for this: First, the research covers chatbots that are not really comparable in regard to user experience factors. Either they focus on GUIs enacting embodiment aspects of human communication or they are AI-driven, thus focusing on natural language processing aspects like linguistic accuracy. Second, with the exception of Fitzpatrick et al. (2017), none of the studies studied chatbots that were deployed on the most popular platforms like Facebook Messenger. Hence, the applicability of the findings are only partly transferrable to other research employing different kinds of chatbots. And lastly, none of the studies analysed the user experience of a Facebook Messenger chatbot in order to deduct user experience guidelines for building chatbots. In order to achieve the latter, a theoretical framework will be given in the

following sections that will enable the analysis of user experience factors of an educational chatbot prototype with the aim of formulating usability guidelines.

2.3 Theoretical Framework

There are several different definitions for both terms *user experience* and *usability*. The same applies to user experience evaluation methods. In order to have a conceptualisation that can be methodologically operationalised, a theoretical definition is needed. This will provide the theoretical framework to evaluate the user experience of the chatbot so as to eventually formulate user experience guidelines in the third part of this chapter.

2.3.1 User Experience vs. Usability

Even though user experience has been defined by the international standard on ergonomics of human system interaction (ISO) in FDIS 9241-210 as “a person’s perceptions and responses that result from the use or anticipated use of a product, system or service”, there exist a range of other definitions. Alben (1996) defines it as “All the aspects of how people use an interactive product: the way it feels in their hands, how well they understand how it works, how they feel about it while they’re using it, how well it serves their purpose, and how well it fits into the entire context in which they are using it” (p. 12). Hassenzahl (2008) sums it up as “a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service” (p. 96). Law et al. (2009) surveyed 275 researchers and practitioners about their definition of user experience and concluded that “[...] the respondents tend to agree on a concept of user experience as dynamic, context-dependent and subjective, which stems from a broad range of potential benefits users may derive from a product” (p. 727). Beyond that there was little consensus on as what user experience should be specifically defined. The smallest common denominator to these definitions is that user experience is something that goes beyond the mere functionality of a product, service or system to more emotional constructs like affects, joy and feelings. Hassenzahl and Tractinsky (2006) make out three general characteristics of user experience:

- *Holistic*: user experience takes a more holistic view, aiming for a balance between task-oriented aspects and other non-task oriented aspects (often called hedonic aspects) of a systems such as beauty, challenge, stimulation and self-expression.
- *Subjective*: user experience is more concerned with users' subjective reactions to a information system, their perceptions of and their interaction with it.
- *Positive*: user experience is more concerned with the positive aspects of a information system use, and how to maximize them, whether those positive aspects be joy, happiness, or engagement.

All these definitions are not necessarily contradictory, but they emphasise different aspects of the term. Generally there are three broad ways of conceptualising the term according to Bevan (2009). User experience can therefore be defined as:

1. An elaboration of the satisfaction component of usability.
2. Distinct from usability, which has a historical emphasis on user performance.
3. An umbrella term for all the user's perceptions and responses, whether measured subjectively or objectively.

Each conceptualisation comes with a variety of accompanying evaluation methods (Bevan, 2009; Cockton, 2011). The second conceptualisation is exclusive in evaluation methods: either usability is measured quantitatively with performance metrics or user experiences are evaluated qualitatively. The third conceptualisation is too broad for a rigorous research design and robust findings. It is thus worth looking at the first conceptualisation as user experience being a subfunction of the concept of usability.

As alluded to above, the concept of *usability* was traditionally focused on human performance and narrower in scope. It focused mainly on efficiency, goal achievement and other quantitative measures (Mifsud, 2011). Etymologically, the term *usability* predated the concept *user experience*. The term originated in the advent of falling prices for personal computers in the 1980s. Up until then, almost all users were highly trained specialists of expensive centralised equipment. When PCs became more affordable they were designed

with the implicit assumption of knowledgeable and competent users, who are familiar with technical interfaces. This made a lot of personal computer systems unusable for a broad user group. Personal computing quickly became associated with constant frustrations and consequent anxieties. Thus *usability* became a key goal in designing computer software and terms like *ease-of-use* and *error rates* among others its main values (Cockton, 2011). The goal was first and foremost to create systems that are quantifiably usable and allowed users to achieve their goals.

In a nutshell it can be said that the aim of designing systems with a traditional usability mindset was to make a system easy to use, i.e. to improve the human performance using the system. On the other hand, the goal of designing systems from a user experience perspective was to make the use of systems enjoyable, thus aiming at more hedonic aspects. Or to put it more sharply: from a usability perspective the main question was “Can the user accomplish their goals?” whilst from a user experience perspective it was “Did the user have as delightful an experience as possible?” (Mifsud, 2011).

The notes accompanying the aforementioned ISO 9241-210 broadened the traditional usability concept stating explicitly that “the concept of usability used in ISO 9241 is broader and [...] can include the kind of perceptual and emotional aspects typically associated with user experience”. In the updated ISO 9241-11 standard, *usability* was further defined as “The extent to which a product can be used by specified users to achieve specified goals, with effectiveness, efficiency and satisfaction in a specified context of use.” *Usability* was hence not just a measure of user performance anymore (i.e. efficiency and effectiveness), but is also concerned with more hedonic measures like user satisfaction. This newer definition of usability with user experience as a sub-function bears two advantages:

1. It includes not only pragmatic aspects, but also hedonic ones usually associated with the term user experience. It follows that creating products that aim for usability improvements not only aim for improved human performance but also for better user experiences.

2. There exist more established usability evaluation methods to work with, which thus provide for a more robust research design.

Thus going forward the term *user experience* will be used in line with the first conceptualisation defined by Bevan (2009) above. That means as part of the concept of *usability*. Furthermore I will use the term *user experience* and *usability* interchangeably and in line with the ISO 9241-11 definition of the term.

2.3.2 Usability Testing & Formative Evaluation

As alluded to in the previous chapter, there are different usability evaluation methods (UEMs) for different purposes of the research. Even though the research question should mainly lead the choice of UEM, there are other approaches to choose the right method. For a broad overview of most common UEMs see Obrist et al. (2009), who identified 35 *UEMs*, or Petrie and Bevan (2009) who group them broadly into 5 categories.

Rohrer (2014) categorises 20 popular UEMs on three dimensions: attitudinal vs. behavioural, qualitative vs. quantitative and context of use. Thinking along those three dimensions can help to identify which evaluation method is most appropriate for the research objective. The difference between attitudinal vs. behavioral methods is the difference between what people think and what they do. Generally, usability studies tend towards a behavioral approach as the goal is to find out how a product is being used. On the other dimension, UEMs either measure data directly and qualitatively, like Ethnographic Field studies or an instrument like surveys or an analytics tool is used to measure data indirectly and predominantly quantitatively. Rohrer's (2014) framework should help to choose which method to employ when.

A more pragmatic approach in choosing an UEM is to consider the product development stage. Not every of the methods mentioned by Rohrer (2014) could render valuable data on a prototype. Usability lab studies or *usability testing* is considered the most popular UEM (Nielsen, 1993) and can be used throughout the entire product life cycle (Norman & Panizzi, 2006). It is well established since the 1980s and is widely accepted by practitioners (Hartson et al., 2001). According to Nielsen (2012) there are three main characteristics to usability testing:

1. The participants are potential real end users.
2. The participants are asked to perform representative tasks with the design.
3. The evaluator observes and records what the participant is doing and saying with least possible interference by the evaluator.

Usability testing has initially evolved from the traditional usability concept with its emphasis on performance measurement and metrics analysis. By now usability testing has become more open to qualitative aspects and even some interference by the evaluator especially in the early stages of product development. This is due to software development systems like Agile, LEAN or Extreme Programming that gave the product development lifecycle a more iterative spin.

Rubin and Chisnell (2008) consider usability testing the most effective UEM for all stages of the product life cycle. Within usability testing there are two main techniques to employ depending on the purpose of the test.

Summative evaluation aims for more rigorous quantitative data analysis and is only applicable once a design is reasonably complete. Often the aim is to measure the product usability or accessibility with emphasis on constructs such as effectiveness, efficiency and satisfaction. Each type of measure is usually regarded as a separate numerical factor with a relative importance that depends on the context of use (Cockton, 2011).

Formative evaluation on the other hand is employed before a product or system's design is considered final and accepted for release (Hartson et al., 2001). Formative tests "focus on understanding the user's behavior, intentions and expectations" (Cockton, 2011) in order to improve the final usability. Typically formative tests employ a "think-aloud" protocol, that encourages the participants to continuously voice their thoughts as they use the product. The results can be less formal than in summative evaluation depending on the needs of designers, developers, project managers, and other project participants.

Table 1. Formative and summative usability testing.

Evaluation Type	Testing subject	Method	Purpose
Formative evaluation	Prototype	Mostly thinking aloud usability testing	Find usability problems in order to improve prototype
Summative evaluation	Final product	Performance measurement	Assess the overall quality of the product

It is worth to note that the techniques of usability testing have undergone a similar development as the definition of usability itself. From a rather narrow scope in *summative testing*, focusing mainly on quantitative performance measures employed in later stages of the product development cycle, to *formative testing* to help to find usability problems and explore what it means to have a good user experience in general. This kind of collaborative exploration in formative testing can help to identify, formulate and shape user experience guidelines which will be introduced in the following section.

2.3.3 User Experience Guidelines

As defined in Chapter 2.3.1 user experience is considered a subfunction of usability in this thesis. Because usability guidelines are more established than user experience guidelines this has the advantage that well known usability guideline concepts can be taken into account.

Usability guidelines, standards or principles for informations systems have been developed for many years. The early usability guidelines were based on the traditional concept of usability, hence guidelines were more technocentric in their approach. The prevailing conviction was that user interfaces would be inherently usable if they conformed with predefined guidelines (Cockton, 2011). Over time early and detailed usability guidelines like Smith and Mosier's "Guidelines for Designing User Interface Software" (1986) were distilled into rather high-level guidelines such as Shneiderman and Plaisant's (2016) "8 Golden Principles of good interface design". One of the most popular guidelines are Nielsen and Molich's ten "Usability Heuristics" (Nielsen & Molich, 1990) that were

further refined in Nielsen's usability evaluation method "Heuristic Evaluation" (1994), a UEM that employs expert evaluators that examining software products for potential causes of poor usability and mapping them against the ten heuristics. According to Cockton (2011) Heuristic Evaluation became the most popular user-centred design approach in the 1990s, but has become less prominent with the move away from desktop applications. Faster and less expensive UEMs soon overtook Heuristic Evaluation. The same applied to detailed user interface guidelines defined in the ISO standard. Those worked well for referencing, but because they were too time consuming they were rarely used by practitioners.

Furthermore most of the guidelines were focused on Desktop or Web applications. This also means that the those guidelines were less applicable to mobile environments. Generally mobile is "sharpening" (Nielsen, 2011) usability guidelines, which means that guidelines for desktop get more constraints and thus ask for stricter usability guidelines. The main reason are technological constraints and possibilities that come with mobile user interfaces. According to Budiu (2015) those are:

- **Smaller screen:** Due to the smaller screen estate the user is incurring higher interaction costs for the same amount of content as on a desktop interface. This means that the designer needs to think about the opportunity costs of each new element. Only the most important elements should make it on to the screen.
- **Portable = interruptible:** Because mobile devices can be taken anywhere they compete with many more outside influences than a desktop device. Interaction on mobile thus needs to be designed for interruptions, i.e. enabling users to pick up the thread when they get back to their device.
- **Single window:** Mobile devices mainly allow for one single window to be displayed at the same time. The design should be self sufficient and not rely on other applications.
- **Touchscreen:** Gestures represent an alternative (UI), that, when built with the right affordances, can make the interaction fluid and efficient and can save screen real

estate. On the other hand, it is hard to type proficiently on a virtual keyboard and error rates are high.

- **Variable Connectivity:** Because connectivity is not equally well distributed, systems should be designed with as little back-and-forth between client and server as possible.
- **GPS, Camera, Accelerometer, Voice:** Mobile devices come with certain technical possibilities that desktop devices lack either in functionality or user experience. Designers should take advantage of those.

Even though those constraints and opportunities can help think about the user experience of chatbots on mobile devices, they are not guidelines that would help create good user experiences. In his book “Designing Bots” Shevat (2017) describes different bot variations in detail, but misses to define concrete guidelines. Facebook Messenger’s developer documentation lists nine “Design Principles” that should help to design chatbots:

1. **Be Brief:** Interactions should be kept short because of interruptibility (mentioned above).
2. **Avoid Modality:** A chatbot is in a *modal* state when it is expecting a specific set of responses. If a chatbot gets interrupted while in a modal state it can be hard to reestablish context or pick up the thread for the user.
3. **Mix Conversation and UI:** The Messenger Platform offers a range of conversation components, from pure text messages to structured templates to full GUI interactions in the webview. Designers should fully consider what format will create the most straightforward, intuitive experience. Often, the answer will be a combination of conversational and UI interactions.
4. **Observe Conversational Norms:** Chatbot designers should deliberate about language, editorial voice, length of messages, and even how fast a bot responds. Further a chatbot should never pretend to be a human being.
5. **Embrace Structure:** While recognizing free-form typed responses can be valuable, it can also be challenging to implement for people interacting with your

bot. Designers should make use of buttons, quick replies, and the persistent menu to structure user input. This can help streamline interactions and clearly communicate expectations.

6. **Be Predictable:** Designers should use the typing indicator to let people know when the bot is in-progress. Further clear opt-in functionality for subscriptions should be provided and subscriptions should not be changed without consent.
7. **Notify with care:** Users should be notified deliberately.
8. **Fail Gracefully:** If a chatbot does not understand a request, reiterate the capabilities: highlight help functionality, or use buttons, quick replies, and the persistent menu to clarify. Each failure should be treated as feedback.
9. **Do Not Create a Separate Entity:** It is recommended to tie the identity of a Messenger bot to an existing Facebook Page, rather than creating a new one. This ensures people will have an easier time finding it and feeling confident it is really this business.

Even though those principles can help building chatbots, they are according to Facebook “by no means exhaustive” (Facebook 2018).

Although guidelines such as Nielsen’s Heuristics or Facebook’s chatbot design recommendations are to a certain degree applicable for the research objectives, they are only generalizations so that they fall short on certain aspects that are specific to creating an educational chatbot. We face difficulties in applying those general guidelines without having a certain expertise in the application domain like educational self-help content. (Petrie & Bevan, 2009) It thus makes sense to try to formulate new user experience guidelines specifically for educational chatbots on Facebook Messenger.

3 System Design

To find out what a good user experience for chatbots entails, a prototype called *Gustav* was designed that was then user tested. The chatbot was what Rubin & Chisnell (2008) call a full horizontal and partly vertical representation of the final product. In a horizontal representation the user is able to move horizontally within the feature scope of a prototype,

but not all features are fully fleshed out. In a vertical representation on the other hand the user is able to use one feature in its entirety, but not move across different features. In case of this thesis, the prototype had full horizontal and vertical integration in regard to technical features. From a content perspective only two conversation paths were fully fleshed out.

3.1 Technical and Content Design

The chatbot was deployed on Facebook's messenger platform using chatfuel.com as the development and deployment platform. Chatfuel enables the creation of a chatbot based on a decision tree and some rudimentary artificial intelligence features. Developing, testing and deploying all happened within chatfuel.

The content creation on chatfuel was structured mostly in conversation blocks that can be linked together through different links like predefined responses, keywords, user attributes and others. A block entails a conversation path that can be structured with different automatic responses and links to other blocks. A sequence consists of blocks that are separated by a specified timeframe. For example on the first day of starting the conversation the user receives Block 1 and a day later Block 2. Attributes describe certain aspects of a user. For example a user can specify that she is interested in a certain topic and thus has the attribute of "interest: XYZ" from then on. This enables targeting users with specific messages.

As defined in Chapter 1.3 – Scope Gustav was a chatbot with hybrid functionality, which means that the conversation was based on a decision tree that enabled the user to go down certain conversation paths with suggested user responses. It is closer to a "choose your own adventure self-help book" (Fitzpatrick et al., 2017, p. 3) than an AI-powered free conversational agent like Siri. The chatbot was therefore not fully capable of understanding all user inputs. If Gustav did not understand an input, default messages helped the user to get back on track. Figure 1 shows the high-level content architecture with the content group "Procrastination" and the sequence group "Productivity". Each consists of several content blocks that are a full conversation path within the chatbot.

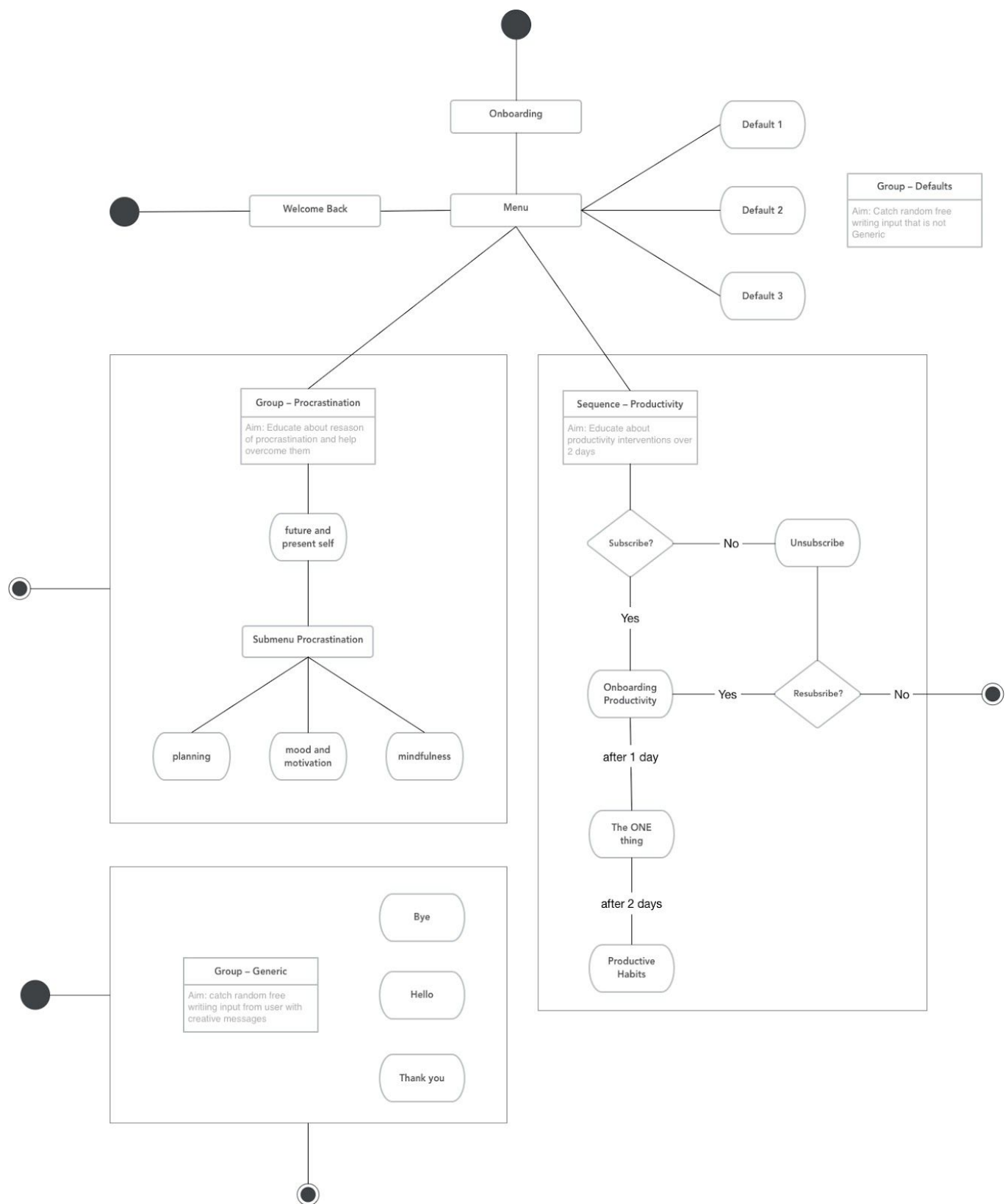


Figure 1. High level conversational decision tree. Each block represents an entire conversation path.

Gustav employed several computational methods depending on the specific section. Furthermore some natural language inputs were processed with simple natural language inputs embedded at specific points in the tree to determine routing to subsequent conversational nodes. This is noted as “Group – Generic” in Figure 1. For the duration of the study, the decision tree structure remained the same for each participant and parameters did not change depending on the participants’ inputs.

3.2 Persona Design

Fitzpatrick et al. (2017) research indicated that the relationship aspect between human and bot is essential for the efficacy of its chatbot helping with mental disorders. Lloyd (2016) describes designing human-bot relationships as the new frontier in user experience design. It was thus an important part in the developing process to create a chatbot persona that is relatable to the targeted user group (i.e. students). The bot’s persona was modeled with a conversational style with a very jovial and uplifting language. Pop cultural references and linguistic idiosyncrasies of students (age 20-29 years) were used.

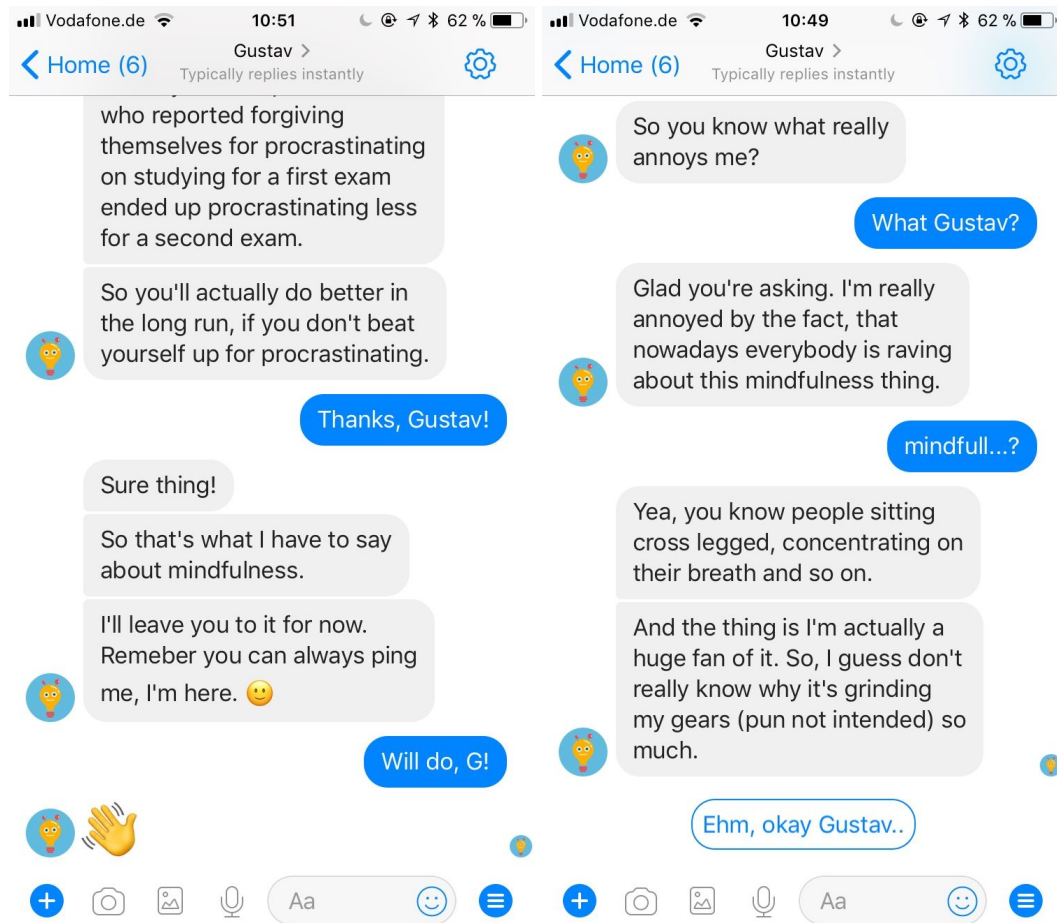


Figure 2. Chatbot Gustav using language and slang common among students.

Psychoeducational content was adapted from self-help articles on the web. Further the bot included self-help exercises like Goal setting or mindfulness meditation instructions.

4 Method

4.1 Research Design and Strategy

As laid out in the introduction chapter, this study followed a Design Science Research (DSR) Strategy, which is essentially a pragmatic research approach that aims at creating artifacts that can solve relevant business problems. Hence the main question DSR tries to answer is what is useful (Hevner et al., 2004). Nevertheless DSR is aiming for scientific knowledge generation. The assumption is that in creating and applying the designed

artifact, knowledge about a problem domain and its solution can be achieved. In other words: the designed artifacts are knowledge containing. Truth and utility are inseparable as “[a] justified theory that is not useful for the environment contributes as little to the IS literature as an artifact that solves a non existent problem.” (Hevner et al. 2004, p. 81).

Broadly speaking, there are four different kinds of design artifacts in DSR: *constructs* (vocabulary and symbols), *models* (abstractions and representations), *methods* (algorithms and practices) and *instantiations* (implemented and prototype systems). (Hevner et al., 2004, p. 98) The artifact as well as the construction of the artifact must be rigorously evaluated. Hevner et al. (2004) formulated seven guidelines that should help guide the building and evaluation process in DSR:

Table 2. Design Science Research Guidelines (Hevner et al., 2004, p. 83)

Guideline	Description
Guideline 1: Design as an artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
Guideline 2: Problem relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
Guideline 3: Design evaluation	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
Guideline 4: Research Contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
Guideline 5: Research rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
Guideline 6: Design as a search process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Guideline 7: Communication of Research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

These guidelines will be covered in Chapter 6 discussing the results of the study. Summing up, even though DSR lacks a unified standard in regard to conducting and evaluating studies (Prat et al., 2014), there is consensus on to what the goal of DSR should concretely be: to produce design theories (Walls et al., 1992; Gregor and Jones, 2007). Especially design principles (e.g. guidelines) and requirements often constitute the central components of a design theory (Prat et al., 2014). Hence this thesis' research objectives are aiming at producing two kinds of DSR artifacts that can inform a broader design theory: the chatbot itself as an *instantiation* (prototype) and usability guidelines for chatbots as a *method* (practises). The following section will cover how the artifacts as well as the research design will be evaluated.

4.2 Research Method

The thesis is following a mixed-method approach by combining different qualitative research methods. As mentioned in Chapter 2.3.2 *formative usability testing* is the UEM that enables to not only find usability problems, but also explore a user's design preferences and attitudes towards the designed artifact. These can then help formulate user experience guidelines. There are a variety of guidelines and definitions on how to conduct usability tests, this thesis will mainly follow Rubin and Chisnell's (2008) guidelines on formative or exploratory user testing.

Usually in usability testing there is an emphasis on task completion and user performance. In case of formative evaluation it is also common to simply employ a "walk through" (Rubin & Chisnell, 2008, p. 18) approach. The walk through approach has been adopted in this thesis in order to not only find higher level usability problems, but also explore the user's expectations and solutions in regard to certain features and user experience problems. Participants were encouraged to voice their ideas on how to improve confusing aspects. Hence the emphasis on understanding *why* a participant performed as she did and *how* that could be improved. Rubin & Chisnell (2008) summarise formative or exploratory testing:

“The testing process for an exploratory test is usually quite informal and almost a collaboration between participant and test moderator, with much interaction between the two. Because so much of what you need to know is cognitive in nature, an exploration of the user’s thought process is vital. The test moderator and participant might explore the product together, with the test moderator conducting an almost ongoing interview or encouraging the participant to “think aloud” about his or her thought process as much as possible.” (p. 31)

One of the disadvantages of a classical *Think Aloud* protocol like Ericsson and Simon (1993) or Boren and Ramey (2000) is that its strict protocol can lead to very little verbalized insights depending on the participant (Shi, 2010). Thus a Relaxed Thinking Aloud (RTA) (Hertzum, 2016) was employed, which allowed for more open prompts like “What did you expect here?” or “How would you improve this?”. These probes are what Bergstrom (2013) calls Concurrent Probing (CP), i.e. probing that was voiced while the user test was taking place. The rich verbalization resulting from a RTA in combination with CP were more relevant for extracting insights in regard to redesign proposals, explanations of behaviour and the overall user experience.

Even though a RTA can provide the insights needed, often the participant is overwhelmed to use the product and verbalize her thinking at the same time. It therefore made sense to conduct a Debriefing Session after the usability test. The scope of a Debriefing Session can vary from a fully fledged interview to one or two open questions that allow the participant to structure and reflect on her user experience with the product. It is often that in those debriefing sessions participants are able to verbalize their thoughts and ideas properly (Rubin & Chisnell, 2008). Hence, a short debriefing session was employed in this research asking the participant two questions: what she liked about using “Gustav” and what she would improve.

4.3 Data Collection Setup

To make sure the prototype worked and was bug-free two informal pilot tests were conducted. These were loosely based on “Guerilla Testing” Methods often used by practitioners. These tests did not aim at scientific rigor, but technical viability. After those guerilla user tests a real pilot test was conducted in order to test the usability test lab and the overall test design. This enabled to correct weaknesses and errors in the research design and helped finalize the Research Plan.

The final usability test took place in a usability test lab or what Rubin & Chisnell (2008) call a “Simple Single-Room Setup” (p. 101). It consisted of two chairs, a desk, a computer with USB cable to record the screen of the testing device. The participants sat in a 90 degree angle to the researcher, who was able to follow in real time what is happening on the participant’s screen through the computer screen.

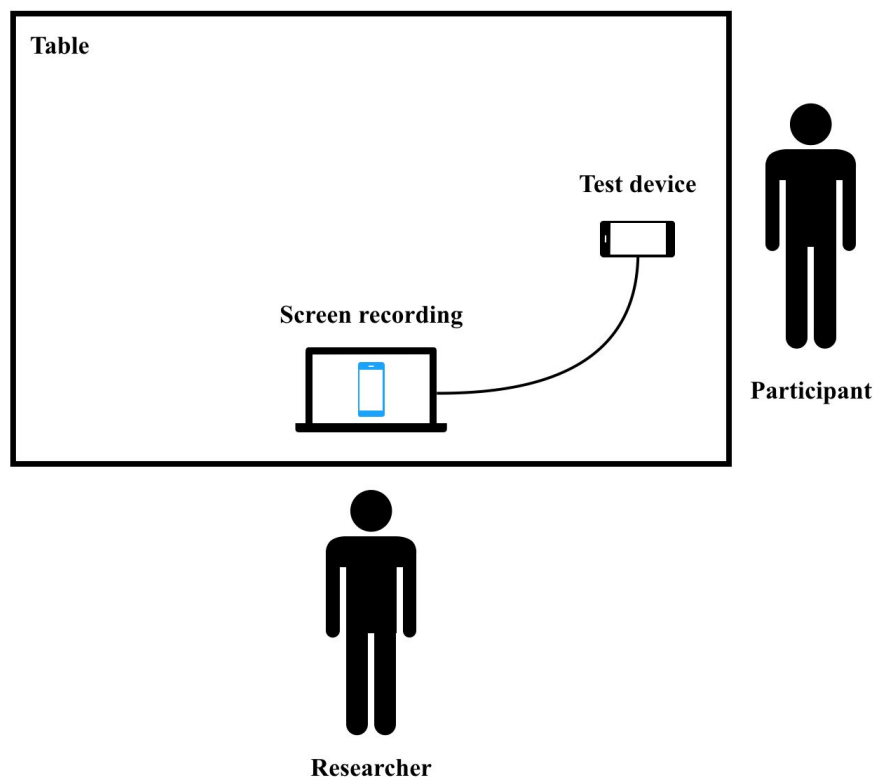
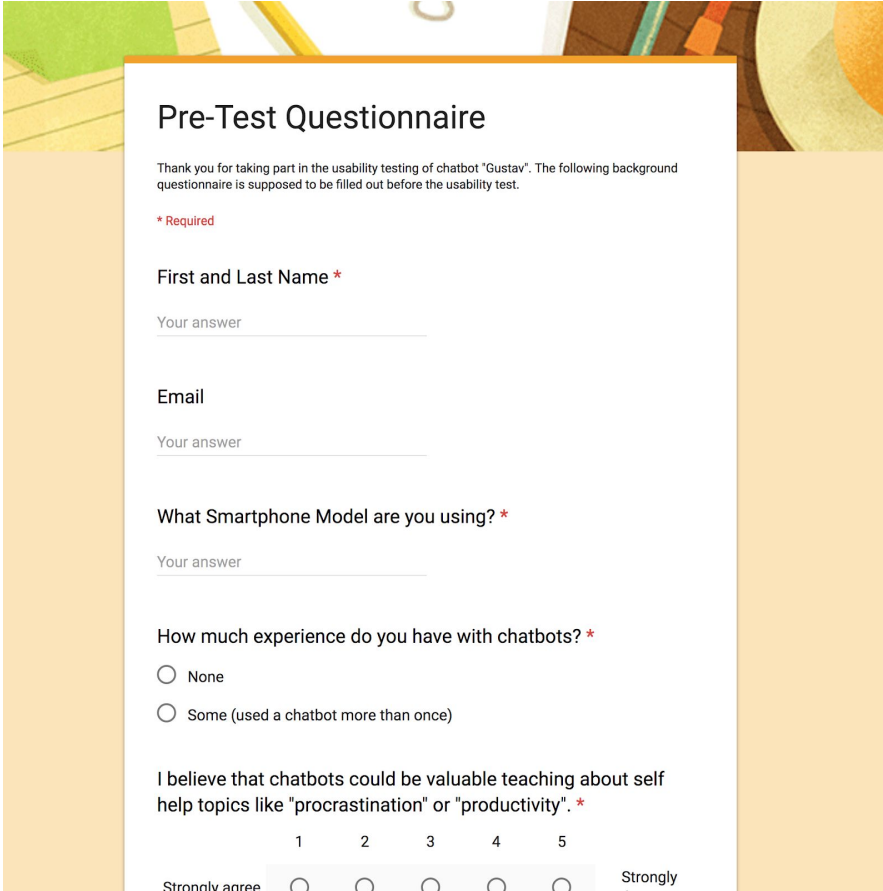


Figure 3. Usability Lab setup

Before a usability test a consent form and a Pre-Test Questionnaire (see Appendix 4 & 5) was administered. The Questionnaire helped to measure the level of expertise or previous experience with chatbots. Also pre-existing opinions and expectations in regard to chatbots were measured. The participants were asked if they believed that chatbots could be valuable in teaching about self-help topics like “procrastination” or “productivity”. This was done in order to potentially see if pre-existing biases in regard to the value of chatbots would influence the final performance. Further simple demographics were measured that helped confirm the sampling criteria. Finally technical conditions like the participant’s smartphone model were asked for, in order to see in how far this influences the performance.

The image shows a digital form titled "Pre-Test Questionnaire" with a white background and orange borders. The form includes a thank-you message, a list of required fields marked with red asterisks, and a Likert scale for a belief statement. The fields are: "First and Last Name", "Email", and "What Smartphone Model are you using?". The experience question has two radio button options: "None" and "Some (used a chatbot more than once)". The belief statement is "I believe that chatbots could be valuable teaching about self help topics like 'procrastination' or 'productivity'." followed by a 5-point scale from "Strongly agree" to "Strongly".

Pre-Test Questionnaire

Thank you for taking part in the usability testing of chatbot "Gustav". The following background questionnaire is supposed to be filled out before the usability test.

*** Required**

First and Last Name *

Your answer

Email

Your answer

What Smartphone Model are you using? *

Your answer

How much experience do you have with chatbots? *

☐ None

☐ Some (used a chatbot more than once)

I believe that chatbots could be valuable teaching about self help topics like "procrastination" or "productivity". *

1 2 3 4 5

Strongly agree Strongly

Figure 4. Pre-Test Questionnaire

After the Pre-Test Questionnaire the participant's phone was connected to the computer and a link to the chatbot was sent to the participant's Facebook account. The task scenario was read aloud and the user was then ready to start on her first task ("To find out more about the topic procrastination"). The RTA protocol allowed for probes and prompts, but the interaction was still held to a minimum. Should the participant get off track, the researcher waited to see if the user will be able to get back on track by herself. During the test the screen and audio was recorded plus the researcher took notes with pre-defined codes in regard to metrics such as successfully finished path, errors, bugs and positive and negative remarks. Also other observations and remarks were noted.

After the participant finished the user test, the debriefing session was administered in situ. The audio was still recording and the participant was asked what she liked about the chatbot and what she would improve. The debriefing session ended with a post-test questionnaire, that aimed at measuring usefulness, ease-of-use and preferences on a 5-item likert scale and asked for optional open-ended questions in regard to improvements and preferences.

After the participant left the usability lab, the video recording of the test was watched and in conjunction with the in-situ notes the test was transcribed in a usability test spreadsheet (see Figure 5).

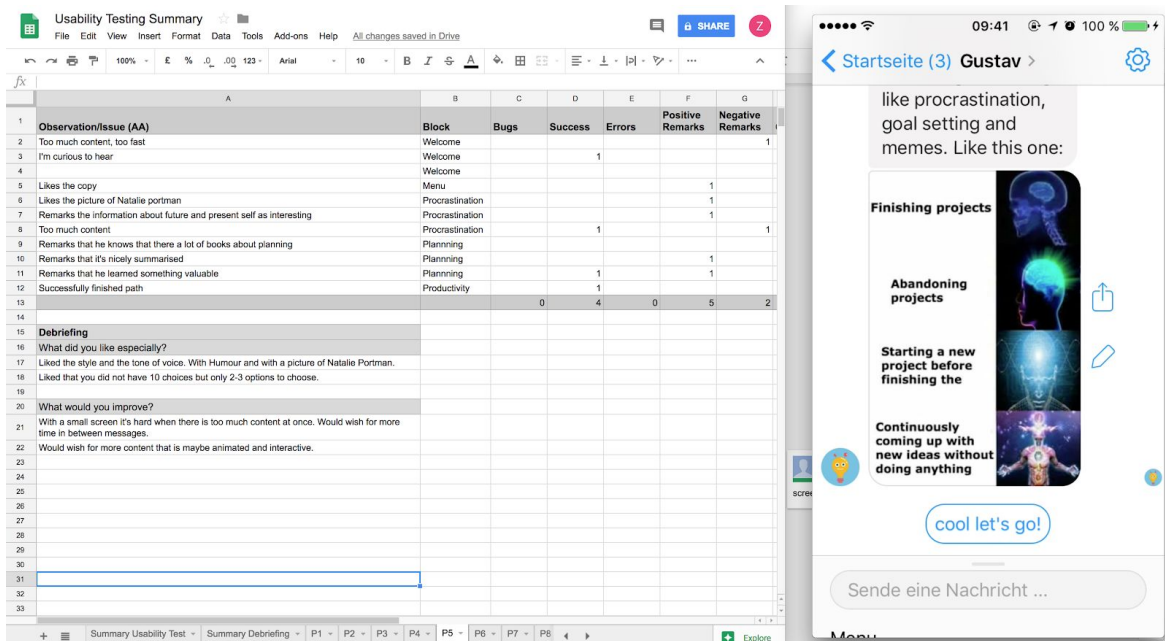


Figure 5. Usability Test evaluation. On the right the recording of the user test, on the left the summary spreadsheet.

4.4 Data Source & Sample

A non-probability sample was employed. The participants were recruited through social media and friend referrals. The advertisement for the study was placed in student social media groups and through an email bulletin of the scholarship foundation of the researcher. The screening followed a purposive sampling approach with ad hoc quotas. Main criteria were that the participant is still a student and between 20-29 years old and currently lives in Berlin. Further the sample was varied in regard to two criteria: 1) gender and 2) experience with chatbots. The research aimed at having an equal distribution along these two dimensions. The aim was to recruit at least 10 people or until data saturation is reached. Data saturation was achieved after about 6 user tests, nevertheless the user test continued a) because participants were invited and b) to make sure that more hidden or less obvious aspects of the user experience got noted. Nielsen's (2012) minimum sample size for usability testing is five participants. Rubin and Chisnell (2008) recommend at least 10-12 participants. Initially 18 people expressed interest and of those, 10 participants took part in the usability test. The sample size thus fulfills both threshold recommendations. The

test took place on two days in Berlin with 5 participants each. See Table 3 for additional sample data:

Table 3. Data sample

	Criteria	Size (n=10)
Age	Mean (SD)	26.3 (2.16)
	Min-max	22-29
Gender, n (%)	Male	5 (50)
	Female	5 (50)
Occupation, n (%)	Student	10 (100)
Experience with chatbot, n (%)	None	5 (50)
	Some	5 (50)
Language, n (%)	German	8 (80)
	English	2 (20)

An equal distribution between male and female was achieved, also the second varying condition “experience with chatbots” was equally distributed. The mean age was 26.3 years (SD 2.16) ranging from 22 to 29 years. Most of the user tests were conducted in German (n=8) and two in English.

4.5 Data Analysis

The data analysis consisted of several steps for each of the usability test phases. The Pre- and Post-Test Questionnaire was administered through Google Forms and the descriptive measurements automatically summarised by the application. The evaluation of the Questionnaire was done as the last step. The full questionnaire evaluation is attached in the Appendix (Appendix 8).

The user tests were evaluated right after each test in order to keep the memory as fresh as possible. After each user test the video recording was replayed and the test was transcoded in a spreadsheet (see Figure 5) that counted the performance metrics but also

noted the path, the issue or observation and important quotes. The hand written notes taken during the user test were used complementary at this stage.

After all tests were evaluated and transcoded, a summary spreadsheet was created that is inspired by Sharon's (2013) rainbow spreadsheet analysis. The spreadsheet helped see patterns by showing the count of certain issues or remarks plus the distribution among the participants. This helped weigh issues and remarks in case certain participants were very outspoken for example.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Code	Observation/Issue/Note	Score	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Quotes	Insights
Negative Remarks														
NR1	Too much content that gets sent too fast	18											AM: "Whoa" AH: "You just get bombarded with those messages"	-> slow down content, make it human spee
NR2	Chatting with bot lacks the freedom of a conversation	3											AH: "It's not really chatting but rather funny way of reading" SK: "Why can't I answer d	-> enable as much free conversation as pc
NR3	Quick replies too similar	3											JF: "What is the difference?"	-> clearly differentiate quick replies
NR4	Does not know how to end the conversation	1											JF: "The only way to end it is to write some gibberish and then get thrown into menu an	-> educate more about conversation flow
NR5	Bot is never reading the last goodbye message	1											JF: "Where is he gone?"	
NR6	Bot cannot pickup where left once you fall out of a path	1											EM: "Feels a bit like going through the motion just to to get back where we left off"	-> enable hooks?
NR7	Copy is sometimes a little too funny	1												-> stick balance between humor and seri
NR8	Content sometime not to the point enough	1												-> be concise
Positive Remarks														
PR1	Humour in copy, replies and media is very good	14											"This is really funny" "He has a good sense of humor"	-> make bot relatable human-like but not
PR2	Content is valuable and well formulated	8											"Very nicely summarised." "Wow that's interesting" "Very nice did learn something no	-> try to find out what move the target audi
PR3	Asking for consent before signing up feel good	2											"Ah that's nice he's asking again"	-> ask for consent, dont automate
PR4	Quick replies are popular	1											"I like the replies"	-> use quick replies more than cards
PR5	Emojis are popular	1											"Ha, I like that"	-> use emojis
PR6	Free writing give feeling of responding	1											"Now it's cool! I have the feeling that Gustav is responding to me"	
PR7	Likes the picture with Natalie Portman	5												
Errors														
E1	Starts free writing and breaks the flow but gets caught by default message	5											"Just wanted to see what he is able to do"	
E2	Clicks of picture received, edits the picture in messenger and sends back, breaks the flow	1											"Didn't think much about it"	-> before sending media think good about
E3	Clicks on message instead of quick reply	1												
Bugs														
B1	After free writing quick reply and keyboard overlap	4												
B2	Short freeze but then continued	2												
Other														
O1	Clicks on picture send tries to read it	7												
O2	Thinks that the menu cards are too big and harder to understand compared to quick replies	1												

Figure 6. User test summary spreadsheet loosely based on Sharon (2013)

The Debriefing interview data was transcoded into the participants (see Figure 5) spreadsheet too and certain representative quotes were translated into english verbatim. The processing of the debriefing session was loosely based on Braun and Clarke's Thematic Analysis (2006). The data was analyzed using an inductive approach. That means all debriefing remarks were collected unaltered. Several readings helped sort and identify patterns. Clusters were built and these clusters were then summarised in overarching themes. (see Chapter 5.1.3 – Debriefing Interview)

4.6 Reliability and Validity Considerations

The main aim of the research was to explore in depth the problem surrounding the research question. More conclusive findings based statistical significance could potentially build on

these findings as discussed in Chapter 6.4 – Further Research. Nevertheless the research findings needed to be reliable and valid. In order to achieve this the research was conducted in line with Saunders et al.'s (2009) reliability and validity guidelines.

Reliability refers to the extent the data collection method and analysis procedures will yield consistent findings. This means:

1. Will the research yield the same results on other occasions?
2. Will similar observations be reached by other observers?
3. Is there transparency in how sense was made from raw data? (Saunders et al. 2009, p. 156)

Generally there are four threats to reliable findings that address those questions. A *Subject or participant error* and *the subject or participant bias* relate to the first question above. In case of this thesis there were a couple of measures taken to counteract those threats: First, the sample was varied equally along those dimensions that could potentially skew the results (e.g. gender and prior experience with chatbots) and controlled for others (age and education). Second, the subjects were informed about their anonymity in this research thus possibly preventing a *social acceptability bias*. The other two threats shed light on the observer and relate to the second question above. Both *observer error* and *observer bias* can impede reliable research finding. In fact according to Saunders et al. (2009) the greatest threat to the reliability of a research conclusion based on participant observation study – i.e. usability testing too – is that of the *observer bias*. (ibid., p. 296) Delbridge and Kirkpatrick (1994) underline this: “because we are part of the social world we are studying we cannot detach ourselves from it, or for that matter avoid relying on our common sense knowledge and life experiences when we try to interpret it” (p. 43). This puts researchers who work alone at a higher risk of misinterpreting the results. A research conducted in a team enables the processing of data by at least two people. Unfortunately this was not possible in this study, thus it is open to critique on this part (see Chapter 6.3 – Limitations). I tried to minimize the observer bias by recording user tests and debriefing interviews. The recorded data was then deliberately listened to several times with the possible observer bias

in mind. (Mortensen, 2017) Furthermore the Pre- and Post-Test Questionnaire with its 5-item Likert scales enabled cross checking conclusions and findings drawn from the usability test and debriefing interview. It should also be noted, that there is research indicating differences in conducting usability tests across cultures (Clemmensen, 2009; Shi, 2010). The main differences were found between western and asian cultures. Because the user tests in this thesis were conducted with both participants and researcher socialised in western cultures (i.e. Germany, Denmark and Canada) it can be expected to be less of a threat to the reliability of the research results. Lastly, the description of the system design (Chapter 3), the research method (Chapter 4) and the analysis of results should address the transparency issues pointed to in the third question above.

Guba and Lincoln (2005) propose to use the terms *credibility* and *transferability* for qualitative research, instead of *internal* and *external validity* usually used in quantitative research. *Credibility* refers to the degree the results are credible or believable from the perspective of the participant in the research. *Transferability* on the other hand refers to the degree to which the results of the research can be generalized or transferred to other contexts or setting. Aside from semantic disagreements, as mentioned above this thesis should work as the blueprint for potential future research in the same regard. I tried to do a thorough job of describing the research context and the assumptions that were central to the research. Further I tried to triangulate methodologically (by using more than one method – usability test, interview and pre- and post-test questionnaire) and theoretically (by shedding light on the theoretical phenomena from different viewpoints). This is supposed to increase the credibility of the researcher. The person trying to "transfer" the results to a different context should be able to make the judgment of how sensible the transfer is.

5 Analysis

In order to find the main factors that make a good user experience both Pre- and Post-Test Questionnaire as well as the usability testing and debriefing interviews were analysed. First the Pre-Test Questionnaire results will be presented (Chapter 5.1.1) followed by grouping the usability test results in negative remarks, positive remarks, errors and bugs (Chapter

5.1.2). After that the debriefing interview results will be presented as a thematic map (Chapter 5.1.3). The results of all three methods will be summarised in insights gained (Chapter 5.1.4) which will work as a the basis to formulate usability guidelines in Chapter 5.2. The overall usefulness of the chatbot in regard to self-help content (Third Research Objective) will be presented in Chapter 5.4. The overall analysis of results will be concluded in the last chapter, which will be discussed in the chapter following.

5.1 User Experience Factors

As a preface to the results of the usability testing, one of the dimensions of the term usability should be quickly touched upon: ease-of-use. Figure 7 shows that the majority of participants considered the chatbot very easy to use. Often this could lead to little verbal remarks as there is not much to negatively remark. It was therefore good to employ a RTA protocol that allowed to elicit more insights from the participants in situations where they were not very voiceful about the bot's user experience.

Overall the chatbot was easy to use.

10 responses

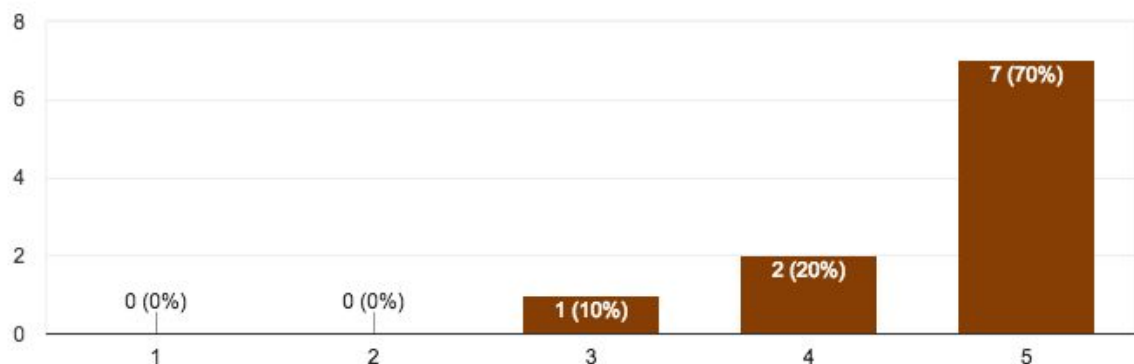


Figure 7. Easy-of-use evaluation

5.1.1 Pre-Test Questionnaire

The Pre-Test Questionnaire had the main goal to measure if the participant is fulfilling the screening and recruiting criteria defined beforehand. Further it also worked as means to cross check interpretations from the user test and debriefing interview. As mentioned in

Chapter 4.4 – “Data Source & Sample” the goal was to control for gender and preexisting experience with chatbots, because it was assumed that both could influence the user performance and thus lead to different insights. Somebody who has a lot of experience with chatbots should perform differently than somebody who never used a chatbot. Half of the sample had no prior experience with chatbots. The other half had used another chatbot at least once. All of them (n=5) have used a chatbot on websites and only two had used a chatbot on Facebook Messenger. There was no perceivable difference between participants with no experience and with some, as both groups performed equally well and often remarked the same things. All of the participant had a Facebook Messenger account (n=10) which could explain why the interaction design was so familiar that there was no difference in performance perceived independent of prior experience with chatbots.

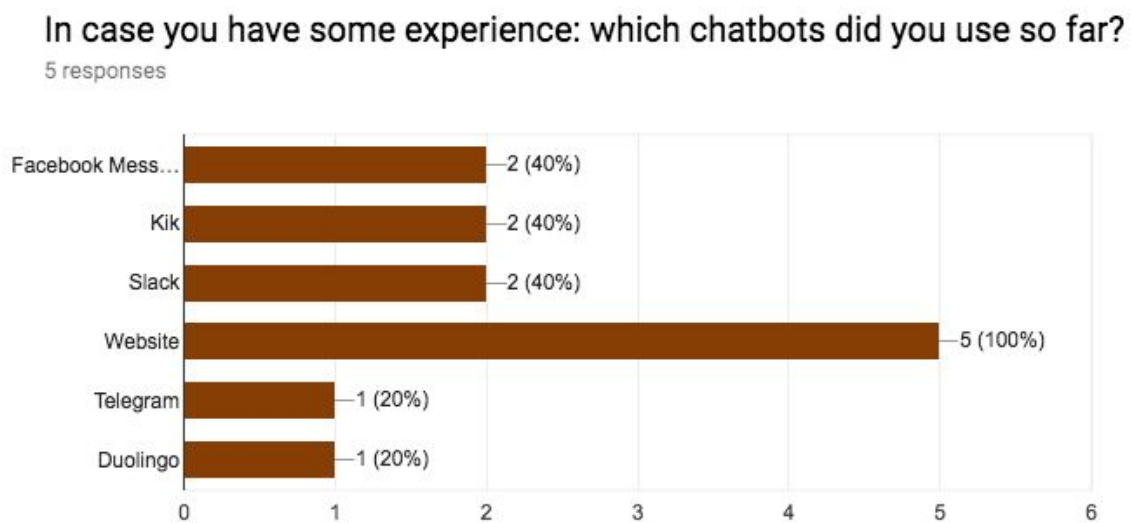


Figure 8. Prior Experience with Chatbots

The same applies for Gender. Except for the fact that four out of five male participants positively remarked about the Natalie Portman picture whereas only one female participant did, there was no performance difference across Gender recorded. The only thing that did seem to matter was the screen size of the smartphone device used. The sample was somewhat equally distributed between a 4 inch screen (n=5) with iPhone 5 and SE and a 4.7 inch screen (n=4) on iPhone 6s and iPhone 7. There was only one 5.5 inch screen on an iPhone 6 Plus (see Figure 9).

What Smartphone Model are you using?

10 responses

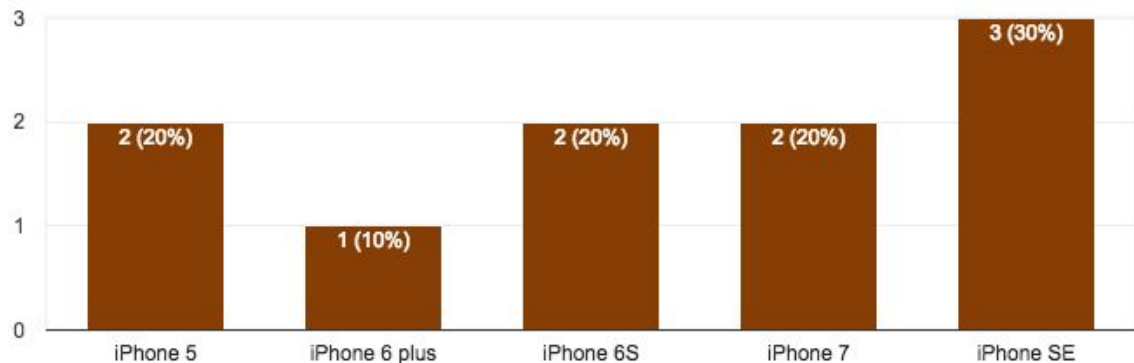


Figure 9. Smartphone devices used in the usability test

The participant with the iPhone 6 Plus (5.5. Inch screen) remarked less about the length of the messages as her screen was able to fit more content on the screen. The participants using the 4 inch screen (n=5) remarked most that the content was sent too fast (see following chapter).

5.1.2 Usability Test

The main objective of the usability testing was to find factors influencing the user experience of the chatbot. This was done by noting and collecting negative (NR) and positive remarks (PR), the number of errors (E) (i.e. the user did something that threw him of path) and bugs (B) (i.e. technical malfunctioning of the chatbot). Further it was noted if the user was able to fulfill the user path on his own or if the researcher needed to intervene and help. If the user was able to do it on her own it was considered a successful user test.

All user tests conducted were finished successfully (n=10). That means in none of the tests the researcher needed to step in to recover the user from either an error she committed that did not allow her to finish the conversation with chatbot Gustav or a bug that did not allow for a successful conversation. This also means that the chatbot would be feature ready to launch publicly for a wider audience.

Negative Remarks

The main negative remark (NR1) was that the chatbot was sending too many messages too quickly, thus the user felt overwhelmed. This was especially the case if other media like pictures or GIFs were used that took even more screen estate. Eight out of ten users remarked this and some several times with a total number of 18 remarks. This was by far the most negatively remarked aspect as seen in Table 4. The second most remarked (NR2) point was that the chat missed a real conversational element. This was remarked by two users and later reiterated by others during the debriefing interview (see following chapter). Followed by the remark that some of the quick reply answers were too similar (NR3). For example on the mindfulness conversation block the chatbot offered two replies “how?” and “in what way?” (see Figure 10).

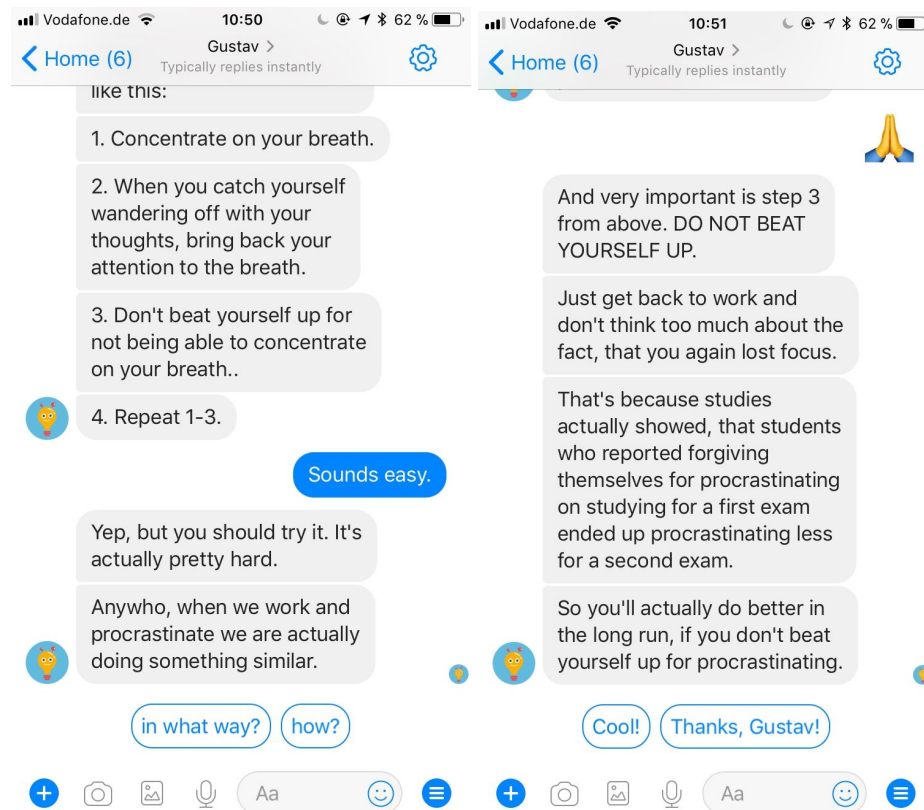


Figure 10. Quick Replies too similar were negatively remarked (NR3)

It was not clear to the user what the difference is. The other negative remarks were mentioned only once and are summarised in Table 4 below.

Table 4. Negative remarks

Code	Observation/Issue/Note	Score	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Negative Remarks												
NR1	Too much content that gets send too fast	18										
NR2	Chatting with bot lacks the freedom of a conversation	3										
NR3	Quick replies too similar	3										
NR4	Does not know how to end the conversation	1										
NR5	Bot is never reading the last goodbye message	1										
NR6	Bot cannot pickup where left once you fall out of a path	1										
NR7	Copy is sometimes a little too funny	1										
NR8	Content sometime not to the point enough	1										

Positive Remarks

The majority of users (n=7) remarked positively about the tone of voice specifically and about the humour of chatbot Gustav generally (PR1). Also the usage of pictures, GIFs, Memes and Emojis was considered positive (PR5, PR7). Further the second most remarked point was the value of the content provided and the way it's formulated. The participants liked the easy approach to the topic and the fact that chatbot Gustav was able to simplify otherwise more complex topics. (PR2) Users also positively remarked the chatbot's is transparency in its communication and asking for consent if for example the user is about to subscribe to a sequence. (PR3) In one conversation block where the user was able to use free writing, this was positively remarked and is in line with the wish to be able to write more freely expressed above (PR6 and NR2).

Table 5. Positive Remarks

Code	Observation/Issue/Note	Score	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Positive Remarks												
PR1	Humour in copy, replies and media is very good	14										
PR2	Content is valuable and well formulated	8										
PR3	Asking for consent before signing up feel good	2										
PR4	Quick replies are popular	1										
PR5	Emojis are popular	1										
PR6	Free writing give feeling of responding	1										
PR7	Likes the picture with Natalie Portman	5										

Errors

An error was noted if the user undertook an action that was either breaking the conversation flow or was not intended when programming the chatbot. An error does not mean that the usability test was unsuccessful. For example a user could fail at a certain

conversation block, but recovered through the menu to finish the usability test. The main error was that the participant started writing freely to the chatbot and thus fell out of the conversation path. (E1) As said before the participants were able to recover quickly. Interestingly only two participants made that error and those two were the users with most prior experience with chatbots. Arguably this could indicate that the more experience one has with chatbots the more one expects to be able to have a free conversation.

Bugs

A bug in contrast to an error is not committed by the participant, but is a technical malfunctioning. This means that potentially this could prevent a user from finishing a conversation successfully. In case of chatbot Gustav two kinds of bugs were reported, but none of them caused an abortion of the conversation. All participants were able to finish their conversation successfully. The main bug recorded was after the free writing exercise in conversation block “planning ahead”. After the user typed in her answer the keyboard interface overlapped with the sequential quick reply that followed. (B1) This is a bug caused by the chatbot engine in conjunction with a 4 inch screen (iPhone 5 and iPhone SE). After trying to debug the issue it still persisted, thus it is arguably a bug due to the Facebook Messenger platform.

Table 6. Errors and bugs

Code	Observation/Issue/Note	Score	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Errors												
E1	Starts free writing and breaks the flow but gets caught by default message	5										
E2	Clicks of picture received, edits the picture in messenger and sends back, breaks the flow	1										
E3	Clicks on message instead of quick reply	1										
Bugs												
B1	After free writing quick reply and keyboard overlap	4										
B2	Short freeze but then continued	2										

5.1.3 Debriefing Interview

The debriefing interview took place as an in-situ interview right after the usability test and helped to highlight certain points that participants did not mention while they were busy operating/conversing with the chatbot. The break between usability test and interview gave them the opportunity to reflect on their user experience and contextualise aspects unsaid or

unnoticed. There were two main questions asked: “What was the best thing about your experience with Gustav?” and “What would you improve about Gustav?”.

“What would you improve about Gustav?”

Figure 12 shows the thematic map for the question what the participant would improve about the chatbot. Three themes emerged: Experience, content and functionality. In the experience theme the main subtheme that emerged was that the messages should slow down in order to enable the user to process them in time, which was a known issue from the usability test (see NR1 in Chapter 5.1.2).

The second subtheme shed light on five subthemes that could improve the content: the content should be more concise (noted by four participants) and dynamic media should be leveraged. Further using social proof or referring to scientific research could improve the content, plus the messages should have a positive spin and lastly they should be motivational. See Figure 12 for quotes illustrating these points.

The third subtheme was functional and aimed at features and the overall functional possibilities of the chatbot. The main subtheme here echoed NR2 from the usability test that participants wished for a more open conversation without the narrow bounds of pre-formulated answers. While only two participants noted this during the user test overall five participants noted it in the debriefing interview, which shows the value of a debriefing interview to unearth unspoken factors from the usability test. The second most mentioned improvement was to have more interactive exercises like the free writing exercise in the “planning” block (noted by 3 participants).

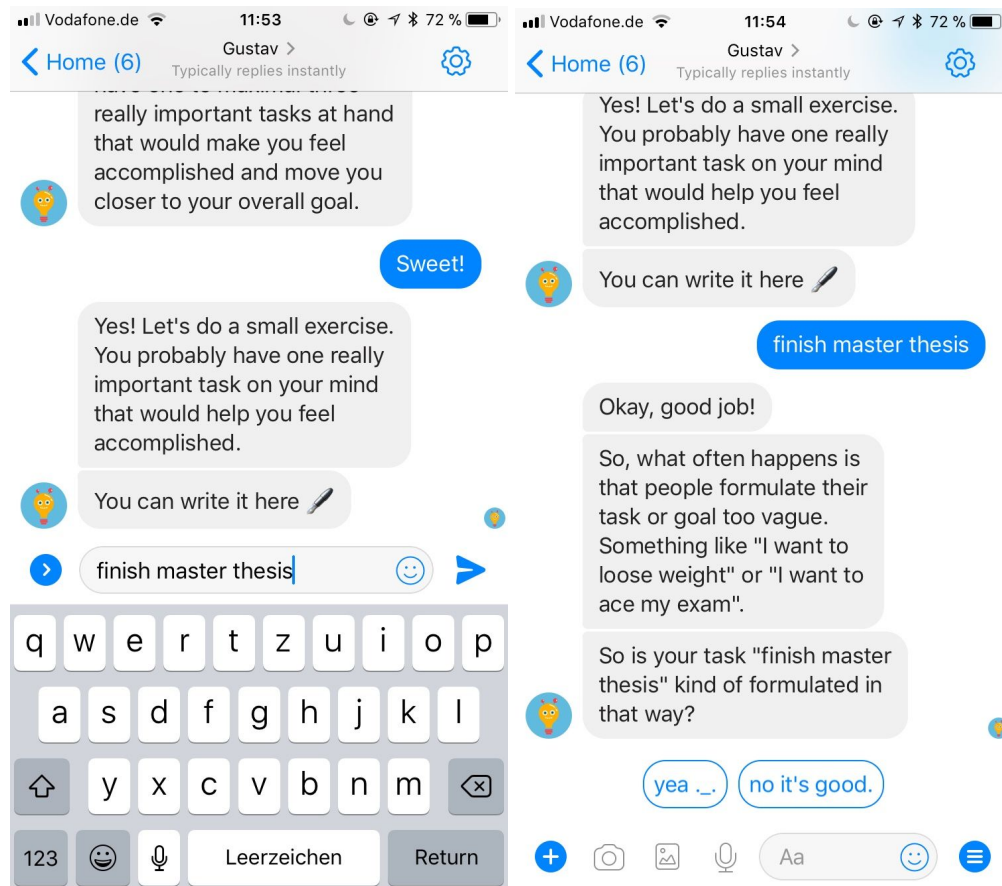


Figure 11. Free writing input on the chatbot

Further aspects that could be improved were learning capability, i.e. that the chatbot should be able to learn over time about the participant and adjusts its content and exercises. Two participants wished for regular checkins to see how she is progressing on her goals. Lastly two participants noted that the navigation is tricky within a conversation as you do not know where you are and how you can navigate within the interface.

“What was the best thing about your experience with Gustav?”

Figure 13 shows the thematic map for what participants considered best about their experience with the chatbot. Three themes emerged: Persona, content and functionality.

The Persona theme entailed two main aspects or subthemes a) the tone of voice and humor and b) the relationship aspect of the interaction. Similarly to the usability test the tone of voice and the humour of the chatbot was highlighted by most of the participants (n=7). The second subtheme relationship is the most interesting one as this was not

mentioned during the usability test, but has an arguably deep insight into the user experience with chatbots. Five participants noted that they felt having a kind of relationship to the chatbot. One participant noted “Even if it sounds stupid, but you really have the feeling that there is somebody who is trying to help you. Even though you know it's a bot.” (Participant 9). Further quotes were “He is somehow nice and you forget that it's a chatbot. I sometimes had to laugh, which is really remarkable.” (Participant 7) and “Interesting how easy it is to say very personal and tricky thoughts, because there is no person to judge you.” (Participant 10). This highlights the importance of being relatable and human-like in the approach to the conversation.

In the content theme two subthemes emerged. Three participants noted that they liked the simplification of the content and the possibility to learn through the chat conversation. The aspect of usefulness in regard to learning will be revisited in Chapter 5.4.

The third theme was functionality and similarly to above it revolved around what features and possibilities were especially appreciated. Two participants noted that the rhythm and the way of the conversation by moving through the conversation by tapping on quick replies was very pleasant. One participant noted: “It's interesting that only the fact that you press buttons keeps you at it. It works strangely well.” (Participant 8). Even though some participants criticized the fact that the conversation lacked the freedom of a real conversation, the user experience of conversing with pre formulated quick replies was noted as pleasant too. The fact that the *productivity* sequence block was checking in once a day was noted as positive. And lastly one participant noted that she appreciated the fact that one can always come back to the content provided within the messenger, which makes it more valuable for learning things as it worked like a notebook.

To sum up, the debriefing interview proved valuable to elicit further insights into the perception and thought process of the participants. There were no contradictions to what came up during the user test, but certain things were further illuminated that went unnoticed during the test. Thus the debriefing interviews complemented the findings from the usability test. The next chapter will summarise the findings as insights gained from

both usability tests and debriefing interviews. These insights will then help formulate Usability Guidelines in the chapter following.

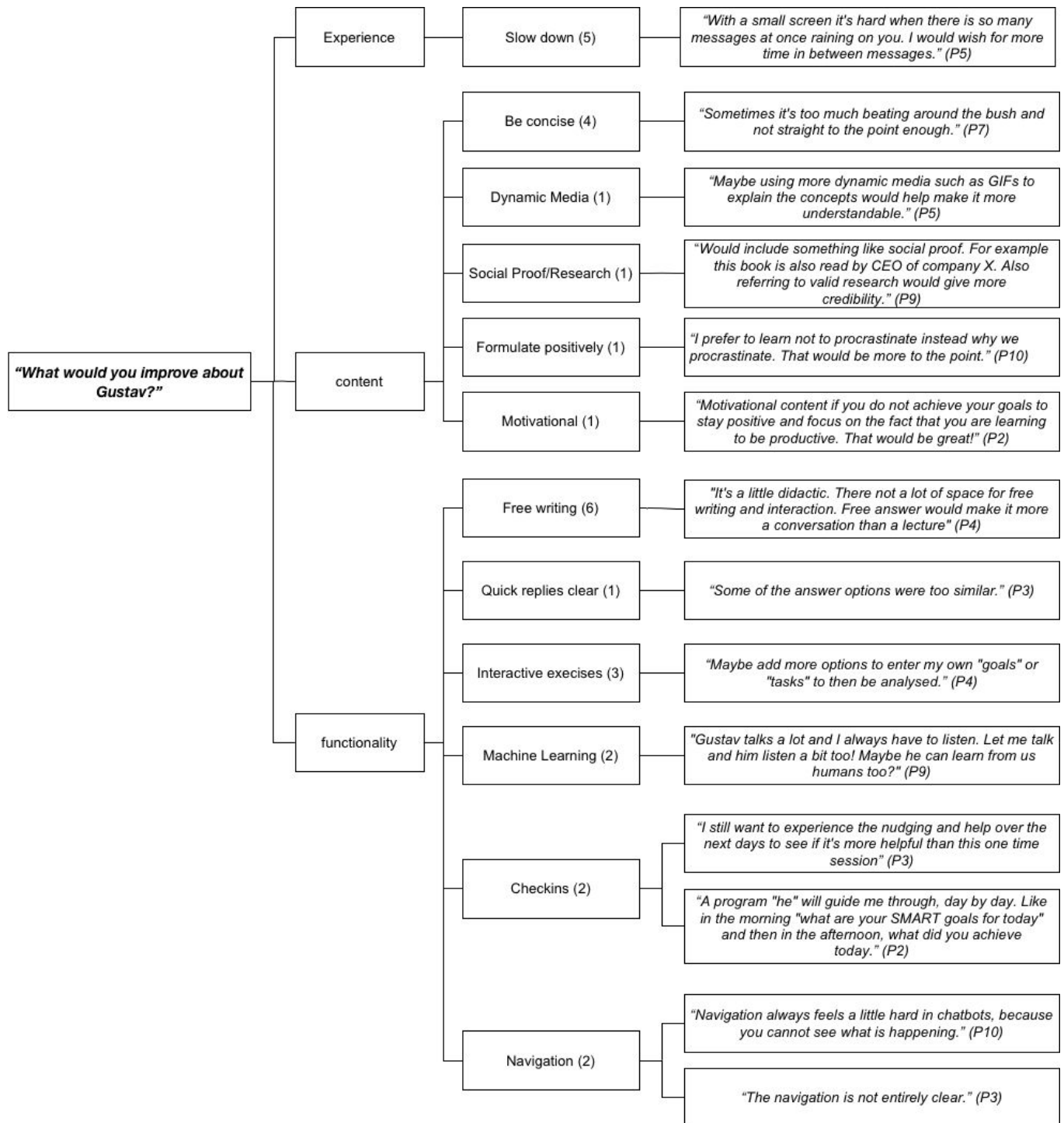


Figure 12. Thematic Map "What would you improve about Gustav?".

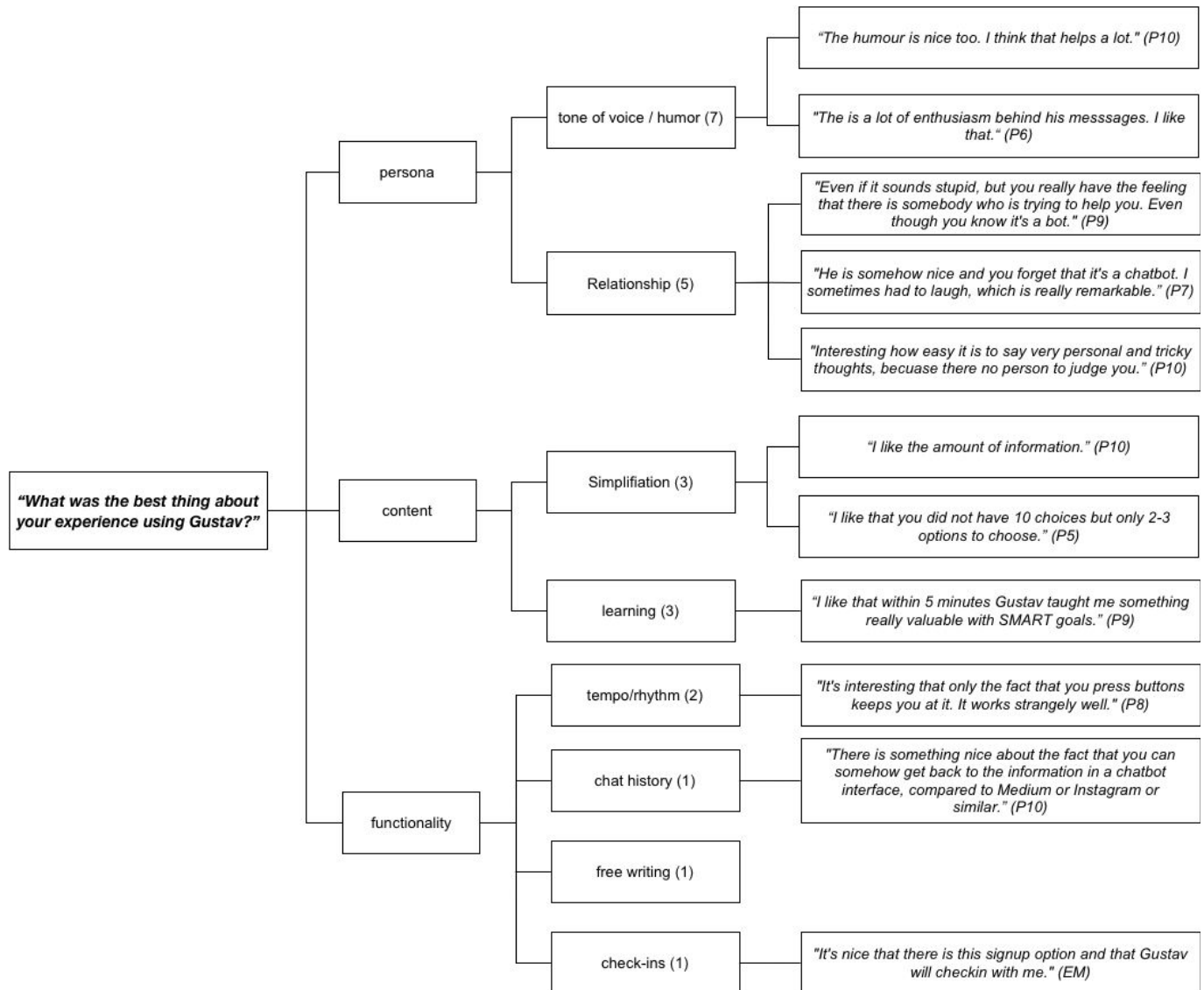


Figure 13. Thematic Map “What was the best thing about your experience with Gustav?”.

5.1.4 Insights

Based on the the usability test and the debriefing interviews the following insights were generated. These insight will serve as the basis to formulate usability guidelines in the following chapter.

Table 7. Insights gained from usability testing and debriefing interviews

Code	Insight	Remarks addressed
IN1	Messages should not be send quicker than the user can process them. This means the conversational tempo should be closer to a human conversation where longer messages need more time to be formulated. Especially if other media like pictures and GIFs are sent, there should be slower cadence between messages.	NR1
IN2	Even though the chatbot clearly tried to manage interaction expectations that are in line with a hybrid functionality chatbot, i.e. decision tree logic, almost all users initially expected the same conversational freedom as in a human to human conversation. Pre-formulated replies sometimes felt too restricting. Also generally a more human like nature of the conversation was expected like reading the last message or picking up the conversation thread where it was left off.	NR2, NR5, E1, E2
IN3	Users expected a clear and concise communication. This applied to both messages received or quick replies provided. Ambiguity in the conversation was considered a bad user experience.	NR3, NR8, PR4, E1, E2
IN4	Navigation was challenging within a conversational interface as there are no visual orientation markers. The underlying conversational architecture was not clear from within a conversation. Falling out of a path brought you back to the menu instead of where the conversation was left off. This frustrated the user.	NR2, NR4, NR6
IN5	The chatbot's persona was considered very important. It enabled the user to establish some kind of relationship that helped to open up, laugh and feel engaged. Feeling engaged in return helped to learn things quicker as seen in thematic map about the positive aspects of the user experience.	PR1, PR5, NR5
IN6	Users expected radical transparency in all regards. The chatbot openly communicated, if the user is being signed up for more messages. This was noted positively.	PR3
IN7	Other media than text was appreciated as long it is in coherent with the chatbot's persona. Emojis, GIFs, Memes, Pictures, etc. were all valued and brought variety into a text only conversational interface. Further they helped further engage the user.	PR5

5.2 User Experience Guidelines for Chatbots

The following section will try to define seven user experience guidelines based on the insights from the previous chapter. As mentioned in the theory section existing guidelines either missed the pragmatism and brevity of guidelines to be practically applicable or they fall short on the depth to be applicable to a certain use-case. Also the majority of user experience guidelines aimed at improving usability in terms of performance measures and accessibility, i.e. in the traditional sense of the term (see Chapter 2.3.1). User experience constructs like pleasure and joy were often not part of this definition. As this thesis understands user experience as a subfunction of usability (see Chapter 2.3.1) the following user experience guidelines aim at improving the overall user experience. The following list should be considered a first version and does not claim to be complete and definite. An adjusted product and further user testing could help refine these guidelines.

Table 8. User experience guidelines

01. Guideline 1: Be human-like, but do not pretend to be human.	Insight
A chatbot needs a persona that users can relate too. Humour helps establish a relationship quickly and engage users. Engagement helps increase the learning effectiveness. But this also means that users often bring the interaction expectation of a free conversation of a human to human conversation.	IN5
02. Use language that resonates with your userbase and is concise.	
The relationship between bot and human is mainly driven by language, thus the language should be appropriate to the userbase. Also messages should be brief and to the point. For educational purposes this means that content needs to be summarised as well as possible. Users do not want to read too much. Built in pauses between messages so the user is able to process the information.	IN3, IN1
03. Be transparent about what happens behind the conversational interface.	IN4, IN6
Because navigating a conversational interface is often a challenge especially in the beginning, everything that happens in the background should be transparently communicated. If a user signs up for a message sequence ask for consent. Start a new conversation by declaring that this is	

a chatbot and not a human being.

04. Keep the user engaged by providing as many interactive moments as possible.

The interaction model should strive for a dialog as much as possible. If free writing and AI-powered NLP is not viable (as it still is not), quick replies and buttons can help keep the user engaged. Try to create moments of conversational reciprocity. IN2, IN5

05. Do not repeat yourself.

In a conversational interface there are only so many elements one can use. Try to not repeat yourself neither in design elements (quick replies, buttons, menu cards) nor in content elements (text, emojis, GIFs, pictures, etc.). Use all elements and provide variation. Default messages like error messages are very important to help navigate a conversational interface. Beyond that they are also an opportunity to show the human-like nature of the chatbot. IN7

06. Guardrail navigation.

As mentioned before navigating conversational interfaces is challenging. It is thus important to guardrail the navigation by interface elements like quick replies. The goal is for the user to build a mental map by chatting with the bot. Leverage menus and submenus.

07. Build narrative loops and recycle content blocks for other paths.

When creating a content block formulate the path as generic as necessary and as precise as possible. Every content block should be accessible from other blocks without breaking the feeling of conversational coherence.

5.3 Usefulness

As stated in Chapter 1.4 the third research objective was to evaluate the usefulness of the chatbot in regard to teaching self-help topics. All methods employed (i.e. Questionnaires, usability tests and debriefing interviews) paid into reaching this research objective.

In order to have a valid measure of the usefulness of the chatbot, the participants were asked before using the chatbot, if they think that chatbots could be valuable teaching about self-help topics like how not to procrastinate or productivity methods. Figure 14 shows that the overall sample tended towards a rather skeptical view on the usefulness of a chatbot providing this kind of content.

I believe that chatbots could be valuable teaching about self help topics like "procrastination" or "productivity".

10 responses

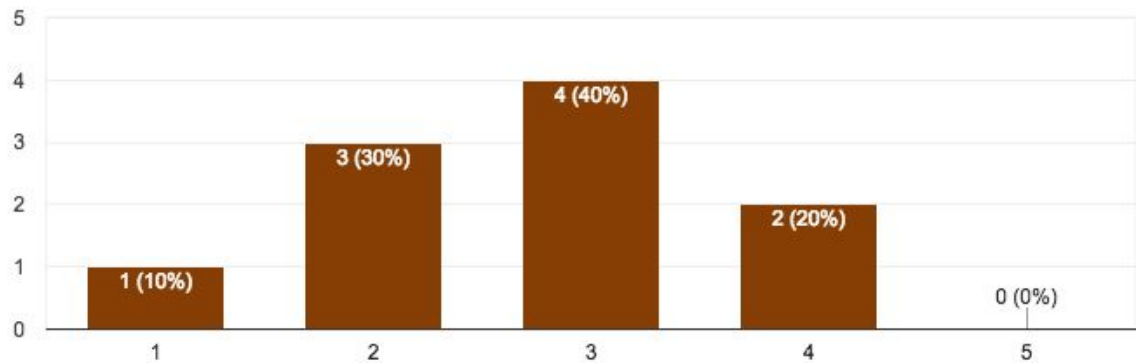


Figure 14. Pre-Test usefulness expectation.

After using the chatbot the participants were asked, if they think that the information they got from the conversation with Gustav was useful. Figure 15 shows that the majority found the content useful.

I thought the information I got from the chatbot was useful.

10 responses

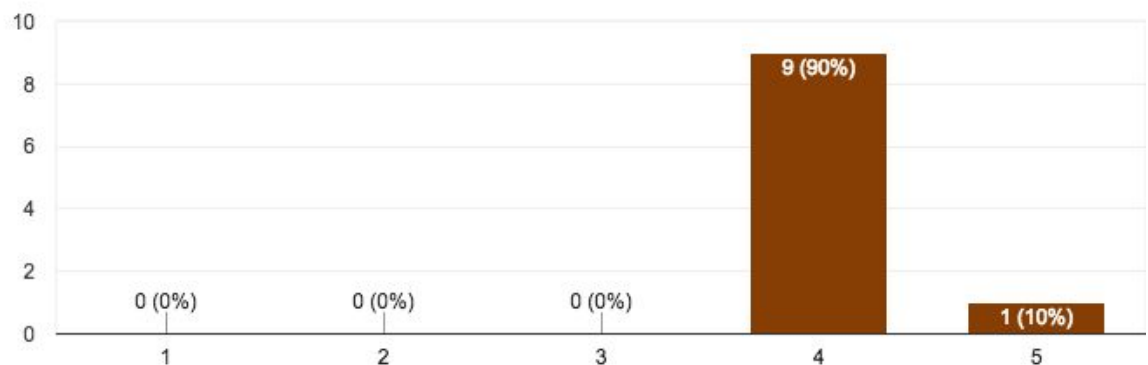


Figure 15. Post-Test evaluation of usefulness.

The Post-Test Questionnaire asked to elaborate more on this question and the following responses were given:

- *“was a good combination of fun and helpful stuff...”*
- *“Through learning about procrastination, I actually was reminded to be more productive in the imminent moment-helping me utilize the tips Gustav gave me, or plan a structure in how to tackle a plan I had prepared.”*
- *“The SMART acronym and the information about future-self with was new”*
- *“I got taught a concept not known before in a blink. Gustav read my mind, when I suggested to wrap up.”*
- *“I think I will actually go home and break my list of tasks into smaller bits. It's not that that is new knowledge - just easy to forget, and good to be reminded.”*

The Pre- and Post-Test questionnaire showed, that almost all participants considered the content valuable and thus chatbots able to help educate about self-help topics. Even though these results bear no statistical significance the overall trend points toward a valid finding that chatbots can be useful in an educational context. A larger sample could help measure the efficacy of the chatbot on people's productivity and mental well-being.

5.4 Concluding Remarks

Beyond the usefulness of chatbots for educational purposes (3rd research objective) the 5th chapter tried to outline the main factors influencing the user experience of chatbots (2nd research objective) and formulate user experience guidelines for an educational chatbot on Facebook Messenger (4th research objective). A multi-method approach with Pre- and Post-Test Questionnaire, user tests and short debriefing interviews were able to help elicit those insights necessary to be able to formulate aforementioned user experience guidelines. Taken all data into account two seemingly contradicting findings came to light:

1. Users expected the bot conversation to strongly resemble a human to human conversation with a human-like persona and free writing functionality. A

good conversation experience should not be restricted by pre-formulated dialog paths.

2. Yet at the same time participants appreciated bot-like communication characteristics like high brevity of messages, concise formulation and high transparency of communication. Further they hinted at the fact that the bot persona possibly allowed for more radically honest communication.

In the following chapter these findings, the research contribution and the limitations of the research approach will be critically discussed. Further an outlook into possible future research will be provided.

6 Discussion

The research question this thesis set out to answer was *What makes a good user experience for students using a text-based educational chatbot aiming at teaching self-help concepts on Facebook Messenger?* The research results point at two initially seemingly contradicting themes that influence a good user experience with chatbots. On one hand users strongly wish for a interaction pattern that resembles that of a human to human communication. This entails a free writing possibilities and a reciprocity of sending and receiving messages similarly to writing with a human. On the other hand users state that there are aspects that make a bot a better conversational partner. Its non-judgmental persona and its availability were among the characteristics most appreciated by users. These somewhat opposing forces delineate a good user experience on chatbots. Both will be critically discussed in light of the existing theory, research limitations and their contribution to theory and practice.

6.1 Good User Experience: Human-like but not Human

As concluded in the literature review (Chapter 2.2) there is very little prior research on the user experience of text-based chatbots. There is however adjacent research that covers among other dimensions user experience aspects. For that reason Følstad & Brandtzaeg

(2017) call for the HCI research community to pick up the thread and think more about conversational interfaces after years of focusing on mainly graphical user interfaces. The scope of this thesis and the chatbot was specific: to look at the user experience of a *text-based educational consumer chatbot with hybrid functionality aiming at teaching self-help topics for mental well-being to students on Facebook Messenger* (see Chapter 1.3). Aside from Coniam (2008) and Atwell (1999) there was no research found that was specifically focusing on the educational usefulness of chatbots. As unfortunately those papers are comparably old, the technological conditions are too different to be relevant. The Facebook Messenger platform for example deployed bots in 2016 for the first time. It therefore made sense to approach the research question with a exploratory impetus.

As concluded in the analysis, the results point at two different themes that delineate a good user experience. The human to bot conversation should on one hand mirror that of a human to human conversation with no limitations in what the bot can understand. On the other hand some bot characteristics like the non-judgmental space to communicate openly and honestly and the great availability to chat seem to be more appreciated compared to human to human conversations. The research showed clear signs that the participants expect a very conversational experience that only differs in certain aspects from a traditional human to human text-based conversation. In a nutshell a good user experience could be summarised as: *human-like but not entirely human*.

The main aspect emphasised positively was the human-like nature of chatbot *Gustav*. Almost all participants highlighted the humour that Gustav showed when trying to educate the participant about self-help topics. The humour was mainly transported through a language that felt familiar to the target group of students (age 22-29) employing cultural references like certain Memes, GIFs, Slang, etc. All of this helped establish a relationship that in the end was engaging which not only helped information transfer, but could supposedly also result in better adherence rates, although the latter was not part of the study. Similarly Fitzpatrick et al. (2017) found that participants noted that chatbot *Woebot* had a good copy and humour. Generally their study also pointed toward the importance of the relationship aspect between participants and chatbot. Although with *Woebot* participants especially mentioned the empathy and attention aspect as very important and

not so much the lightheartedness emphasised in Gustav. This is possibly due to the fact that Fitzpatrick et al. (2017) were studying the efficacy of Woebot in regard to Depression and Anxiety symptoms in students, thus a clinical condition, whereas the sample and the focus of Gustav was on a non-clinical aspect. It would nevertheless be interesting to see if employing more empathetic conversation blocks would improve the overall user experience by deepening the relationship between participant and bot. The fact that participant wished for more daily check-ins from Gustav in order to remind them of the lessons learned point toward that claim.

The finding that users want text based communication with the bot to be as close to human to human conversation as possible is supported by the observation that most of the participants missed the ability to write and converse freely with the chatbot. The negative remarks about the speed of messages sent by Gustav were further emphasising that point. Although it stands to reason that the main factor is not that Gustav should write as slow as a human being, but write as slow as necessary for a human being to process the information. It became clear that it seems hard for users to switch interaction patterns from human to human communication to human to bot communication. A possible explanation seems to be, that this is especially due to the fact the conversation is taking place on Facebook Messenger which most of the participants only used for communication with other people so far. Other research echoes this sentiment and points towards the general disappointment of humans around the capabilities of chatbots. Fitzpatrick et al.'s (2017) study participants remark that Woebot is not able to to converse naturally as the most negative aspect. One participant states "If I wanted to say something when Woebot expected an automated response (like me choosing an answer option) it seemed to really confuse Woebot" (p. 8). To get users into an interaction pattern of a human to bot conversation as quickly as possible seems to be one of the main design challenges. Gustav attempted to catch this as early as the first message to be clear what the interaction possibilities are as seen in Figure 16. This worked better in the controlled user testing setting, but less well if users were using Gustav on their own (note: as the creator of the bot it is possible to see all conversations that the bot is having).

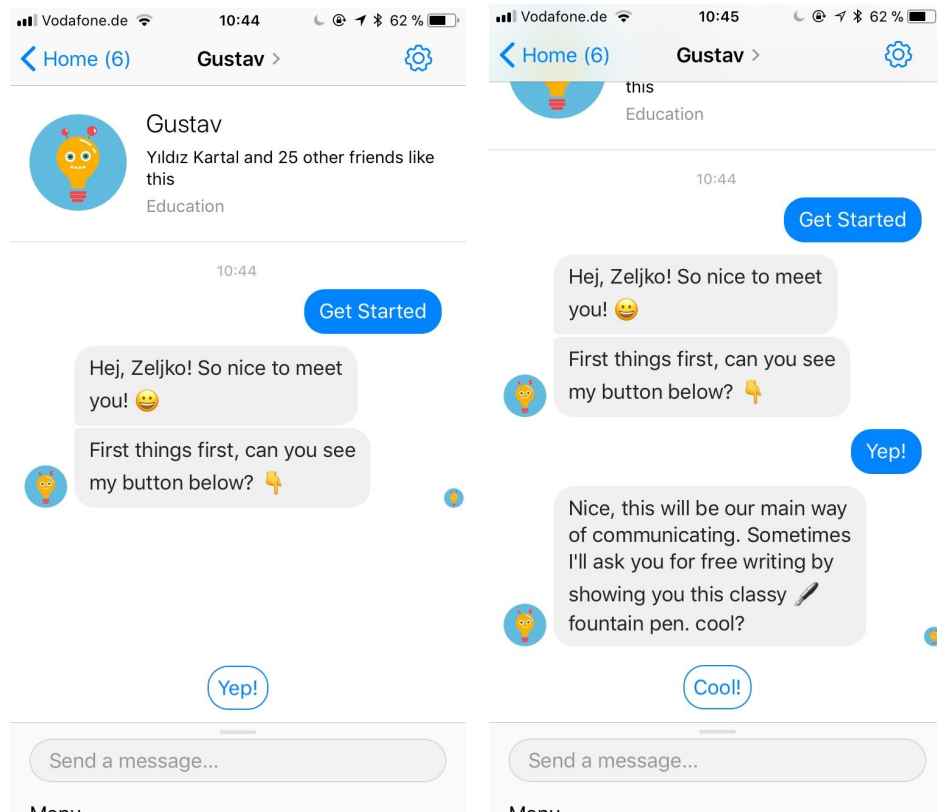


Figure 16. Establishing interaction pattern while onboarding.

Another negatively remarked aspect of the user experience was the repetitiveness of messages, especially if the bot failed to understand the free writing input by the user. Even though there were several different default messages set up, that should have helped the user recover and get back on track, the communication still felt shallow to some participants. Ly et al. (2017) programmed a chatbot called Shim that helped young adolescents with mental well-being similarly to Woebot. The shallowness of the conversation was also remarked by a couple of the participants in that study. One of the participant described it as: “After a couple of times using Shim, I felt I didn't put as much energy to reflect upon the questions. This was a consequence of seeing the same questions again.” (Ly et al., 2017, p. 44). Even though the shallowness aspect was negatively remarked still a majority of participants voiced a positive sentiment in regard to the relationship to Gustav as seen in the paragraph above. This could be interpreted as a certain tolerance in regard to repetitive messages as long as the relationship is carried by other factors like good entertainment and valuable content. This could prove valuable for future

research. Either way “Don’t repeat yourself” was thus formulated as a user experience guideline as seen in Chapter 5.3.

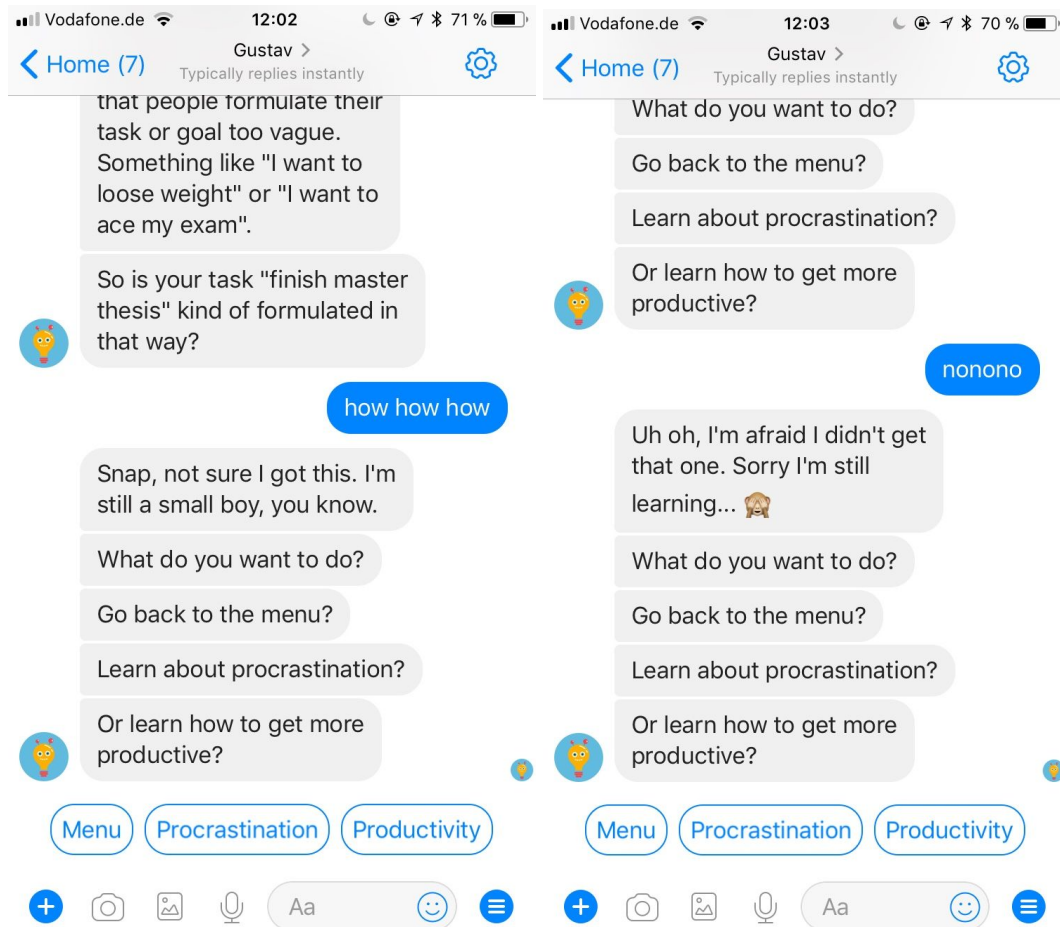


Figure 17. Provide variety in default messages to catch errors.

Even though the research showed a clear preference for human communication characteristics there were also aspects that were uniquely bot-like that were appreciated. First and foremost the precise and fast communication aspect. A chatbot is always available and this is perceived positively. Positively mentioned was also the fact that, because there is no person behind it to judge, participants are more likely to feel free to write whatever they think. This stands in contrast to human to human conversation where the social acceptability bias is more prevalent. The bias describes the fact, that people are more likely to say things that they will not be judged negatively for by others. Even though some characteristics of Gustav like transparent communication and quick reply time point

toward a in some aspects improved communication style with humans, it would be interesting to see how this would change once users were able to use more free writing capabilities.

Overall it can be said that what makes a good user experience on chatbots is very similar to what makes a good conversation with another human being. Unfortunately the technology is yet not developed enough in order to process a human to bot conversation as freely as a human to human conversation. That being said, the research also showed that with small persona characteristics of the bot – like humour and language – relationship aspects can be established that indicate a better user experience and better engagement. Beyond that the research also indicated that certain typical bot characteristics were perceived as positive like a non-judgemental attitude and availability. Overall a good user experience can be described human-like but not entirely human.

6.2 Contribution to Theory and Practice

The goal of a Design-Science Research project is to create a design artifact that addresses a relevant problem. (Hevner et al., 2004, p. 83) Walls et al. (1992) and Gregor and Jones (2007) further emphasize that DSR should produce design theories, which Gregor and Hevner (2013) consider meta design artifacts. Design principles (e.g. guidelines) and requirements are often considered the central component of a design theory. (Prat et al. 2014) Hence, this thesis' contribution from a theoretical perspective are the user experience guidelines formulated in Chapter 5.2 based on the usability testing and debriefing interviews. These should only be considered preliminary, because a DSR project is a search process driven by a continuous iteration process. Beyond the DSR contribution from a design theory perspective, the research indicated to be useful within the problem domain of procrastination problems within a student target group. Fitzpatrick et al. (2017) and Ly et al. (2017) were able to show that chatbots can help improve students mental well-being by employing a chatbot providing small psychological interventions based on cognitive behavioural therapy and self-help interventions from the positive psychology field. Self-efficacy, productivity and procrastination are relevant problems among graduate students. (Flett et al., 2012; Katz et al., 2014) Hence, a chatbot aiming at helping

with procrastination problems and other self-help topics can possibly contribute to a solution within this problem domain.

Even though the research results indicate to be useful from a theoretical perspective, further quantitative and confirmatory research needs to be conducted in order to gain generalizable findings beyond the case researched in this thesis. However, from a practical perspective this thesis should prove valuable for developing new chatbots within the educational domain. As discussed in Chapter 2.3.3 there are no guidelines to how chatbots should be programmed in order to provide a good user experience on Facebook Messenger. Existing guidelines are either too broad and general (e.g. Nielsen's Usability Heuristics) or too specific and outdated for a Facebook Messenger mobile platform (e.g. ISO 9241 standard). Hence formulating user experience guidelines specifically for educational chatbots on Facebook Messenger mobile is arguably relevant for practitioners. Furthermore this thesis could prove valuable to user researchers conducting user research on chatbots. Chapter 3 and 4 can be used as blueprint for how to build a first vertically integrated educational chatbot prototype and then conduct user research to refine the user experience.

6.3 Limitations

Given the exploratory nature of the research the thesis faced a couple of limitations. Even though the results aim at providing a guide on creating good user experiences for other chatbots on Facebook Messenger, it is scientifically not possible to posit generalizability beyond the case of this thesis.

Beyond the research design limitations there were also theoretical limitations in regard to prior research about the user experience of chatbots. Even though chatbots are a research subject since Weizenbaum's ELIZA in the early 1960s, there is very little research about the user experience of chatbots on Facebook Messenger. To the best of my knowledge only Fitzpatrick et al. (2017) touched upon this topic. Reflecting and discussing the findings gained in this thesis was challenging as there was little prior theoretical work to hold it against. I tried to look broader and include findings from other research about chatbots, but often the transferability of findings was not given due to a different subject of

research like chatbots that employ a graphical user interface that aims at mimicking the body language of humans.

From a methodological perspective the greatest limitations was also the most common in qualitative research: observer bias (Saunders et al., 2009). I tried to counteract this by triangulating with a multi-method approach. By having a Pre- and Post-Test Questionnaire and by rewatching each user test recording at least three times, I was hopefully able to contain some of the possible error in interpretation. However, it would have been better, if a second observer were present to interpret the results too.

Lastly as Petrie & Bevan (2009) pointed out design guidelines generally face certain limitations. In order to be able to formulate guidelines that are generalizable one would need to evaluate every page, every screen or in this case every conversation path against every applicable guideline. This of course is impractical and thus selecting certain representative conversation paths can lead to missing some issues or over exaggerating others.

Summing up, even though the research faced limitations especially on a theoretical (not enough prior research) and methodological basis (observer bias), the research design tried to approach the case from different angles in order to provide an accurate description of the case.

6.4 Future Research

Aside from trying to replicate the research with other observers (to counteract the observer bias) and new participants, there are a couple of aspects that are worth researching further.

It would be worthwhile to invest more time and effort in the underlying message database, so that more free writing queries are answered properly, which could in return stimulate a more human-like conversation. This could result in even more positively evaluated user experiences as this research indicated. A positively evaluated user experience should ultimately result in an increase in use of the chatbot.

To take this preliminary evaluation of the user experience further it would make sense to measure the user experience with a more robust user experience framework like attrakdiff (Hassenzahl & Monk, 2010). This would be especially valuable, if the product is

supposed to be marketed further at a later stage as the framework measures and positions the product's user experience on a hedonic and pragmatic quality diagramm.

Aside from a quantitative approach it could make sense to further investigate the user experience from a qualitative perspective. A more comprehensive user research project could link use metrics to individual user experience variables like pre formulated answer formats (e.g. words, numbers, emojis). This could help establish what are the individual design elements that drive a good user experience.

Besides the user experience of the chatbot, its efficacy in regard to mental well-being and productivity also warrants further investigation. Similarly to the quantitative research proposed above this would make sense once the research finishes the exploratory stage. It would be interesting to investigate in how far users perceive their lives improved by the content and exercises provided by Gustav. This could be done in controlled trial measuring the efficacy with the Flourishing Scale (previously The Psychological Well-Being scale). The Flourishing Scale measures the respondent's self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism (Diener et al., 2009). The scale provides a single psychological well-being score and is widely used in well-being intervention studies because of its brevity, simplicity and comprehensiveness (Schotanus-Dijkstra et al., 2016).

7 Conclusion

Even though the number of new chatbots created is consistently growing and chatbots play an important role in Information Systems research since Weizenbaum's ELIZA in the 1960s, there is little research on the user experience of chatbots. This point is emphasised by Følstad and Brandtzaeg's (2017) paper "Chatbots and the New World of HCI", that calls for the research community to shift their focus increasingly to conversational UIs as those will profoundly change the HCI field.

This thesis set out to make a first contribution by exploring factors influencing the user experience of chatbots on Facebook Messenger. A chatbot prototype was developed that delivered educational self-help content and was user tested through formative usability

testing and short debriefing interviews. An initial set of guidelines was formulated based on findings and insights gained from aforementioned research methods.

The research indicates that users want chatbots to have predominantly human-like conversational characteristics like the ability to understand free text input and equal conversational reciprocity between human and bot. Further a human-like bot persona with its own humour and user group specific slang was highlighted as most positive factor of the user experience. Despite users' overall preference for human-like characteristics in a bot, users reported that some specific bot-like characteristics made the conversation preferable to a human-to-human interaction. Among other things participants highlighted the non-judgemental space provided by a bot that made them feel safe to voice otherwise unvoiced thoughts.

Further the research indicates that chatbots can prove useful in providing educational content. Most of the participants were critical prior to using the chatbot, but evaluated the content and the experience as valuable after the chat. This is in line with other research measuring the efficacy of chatbots delivering Cognitive Behavioral Therapy (Fitzpatrick et al., 2017) or psychological interventions from the field of positive psychology (Ly et al., 2017).

Even though the research faced challenges, like a possible observer bias, that could potentially overemphasise negative or positive remarks, the results indicate a valid finding that could work as a blueprint for replication and further more conclusive research with possibly other user groups or different educational content. It would further be interesting to see how the evaluation of the user experience changes over time with long term usage and with a growing possibility of handling free text input by improving the underlying AI-powered conversation database.

8 References

- Alben, L. (1996). Quality of Experience: Defining the Criteria for Effective Interaction Design. *Interactions*, 3(3), 11–15. <https://doi.org/10.1145/235008.235010>
- Atay, C., Ireland, D., Liddle, J., Wiles, J., Vogel, A., Angus, D., ... Rushin, O. (2016). Can a smartphone-based chatbot engage older community group members? The impact of specialised content. *Alzheimer's & Dementia*, 12(7), 1005–1006. <https://doi.org/10.1016/j.jalz.2016.06.2070>
- Atwell, E. (1999). The Language Machine: The impact of speech and language technologies on English language teaching. *London: British Council*.
- Balderas, A., Ruiz-Rube, I., Mota, J. M., Doderio, J. M., & Palomo-Duarte, M. (2016). A Development Environment to Customize Assessment Through Students Interaction with Multimodal Applications. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 1043–1048). New York, NY, USA: ACM. <https://doi.org/10.1145/3012430.3012644>
- Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods?
- BI Intelligence, BI. (2016). Messaging apps are now bigger than social networks. Retrieved March 13, 2018, from <http://www.businessinsider.com/the-messaging-app-report-2015-11>

- Boren, T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261–278.
<https://doi.org/10.1109/47.867942>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 2(3), 77–101.
- Budiu, R. (2015). Mobile User Experience: Limitations and Strengths. Retrieved March 13, 2018, from <https://www.nngroup.com/articles/mobile-ux/>
- Chang, Y.-K., Morales-Arroyo, M. A., Chavez, M., & Jimenez-Guzman, J. (2008). Social Interaction with a Conversational Agent: An Exploratory Study. *Journal of Information Technology Research (JITR)*, 1(3), 14–26.
<https://doi.org/10.4018/jitr.2008070102>
- Clemmensen, T. (2009). A comparison of what is part of usability testing in three countries. Copenhagen: Copenhagen Business School.
- Cockton, G. (2011). Usability Evaluation. In *The Encyclopedia of Human-Computer Interaction* (2nd ed.). Interaction Design Foundation. Retrieved from <https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/usability-evaluation>
- Coniam, D. (2008). Evaluating the Language Resources of Chatbots for Their Potential in English As a Second Language. *ReCALL*, 20(1), 98–116.
<https://doi.org/10.1017/S0958344008000815>
- Crutzen, R., Peters, G.-J. Y., Portugal, S. D., Fisser, E. M., & Grolleman, J. J. (2011). An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. *The Journal of Adolescent Health*:

Official Publication of the Society for Adolescent Medicine, 48(5), 514–519.

<https://doi.org/10.1016/j.jadohealth.2010.09.002>

Delbridge, R., & Kirkpatrick, I. (1994). Theory and practice of participant observation. In *Principles and Practice in Business and Management Research*. Aldershot: Dartmouth (pp. 35–62).

Diener, E., Wirtz, D., Biswas-Diener, R., Tov, W., Kim-Prieto, C., Choi, D., & Oishi, S. (2009). New Measures of Well-Being. *Assessing Well-Being*, 247–266.

Facebook. (2018). A Look back on Messenger Platform in 2017. Retrieved March 13, 2018, from

<https://messenger.fb.com/newsroom/a-look-back-on-messenger-platform-in-2017/>

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>

Flett, G. L., Stainton, M., Hewitt, P. L., Sherry, S. B., & Lay, C. (2012). Procrastination Automatic Thoughts as a Personality Construct: An Analysis of the Procrastinatory Cognitions Inventory. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 30(4), 223–236. <https://doi.org/10.1007/s10942-012-0150-z>

Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the New World of HCI. *Interactions*, 24(4), 38–42. <https://doi.org/10.1145/3085558>

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental

- comparison of Chatbot and Human task partners. *Computers in Human Behavior*, 75, 461–468. <https://doi.org/10.1016/j.chb.2017.05.045>
- Gregor, S., & Jones, D. (2007). Anatomy of a Design Theory. *Journal of the AIS*, 8(5), 312–335.
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Q.*, 37(2), 337–356. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Guba, E.G. and Lincoln, Y.S. (2005) Paradigmatic Controversies, Contradictions, and Emerging Confluences. In: Denzin, N.K. and Lincoln, Y.S., Eds., *The Sage Handbook of Qualitative Research*, 3rd Edition, Sage, Thousand Oaks, 191-215.
- Hartson, H. R., Andre, T. S., & Williges, R. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, 13, 373–410.
- Hassenzahl, M. (2008). User Experience (UX): Towards an Experiential Perspective on Product Quality. In *Proceedings of the 20th Conference on L'Interaction Homme-Machine* (pp. 11–15). New York, NY, USA: ACM. <https://doi.org/10.1145/1512714.1512717>
- Hassenzahl, M., & Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, 25(2), 91–97. <https://doi.org/10.1080/01449290500330331>
- Hassenzahl, Marc, & Monk, A. (2010). The Inference of Perceived Usability From Beauty. *HUMAN-COMPUTER INTERACTION*, 25, 235–260.

- Hertzum, M. (2016). A Usability Test is Not an Interview. *Interactions*, 23(2), 82–84.
<https://doi.org/10.1145/2875462>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Q.*, 28(1), 75–105.
- Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245–250.
<https://doi.org/10.1016/j.chb.2015.02.026>
- ISO 9241-11 (1998) Ergonomic requirements for office work with visual display terminals (VDTs) Part 11: Guidance on Usability. ISO.
- ISO FDIS 9241-210 (2009) Human-centred design process for interactive systems. ISO.
- Jelle van Dijk. (2009). Cognition Is Not What It Used To Be: Reconsidering Usability From An Embodied Embedded Cognition Perspective. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, 5(1), 29–46.
<https://doi.org/10.17011/ht/urn.20094141409>
- Jia, J. (2009). CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4), 249–255.
<https://doi.org/10.1016/j.knosys.2008.09.001>
- Katz, I., Eilot, K., & Nevo, N. (2014). “I’ll do it later”: Type of motivation, self-efficacy and homework procrastination. *Motivation and Emotion*, 38(1), 111–119. <https://doi.org/10.1007/s11031-013-9366-1>

- Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P. O. S., & Kort, J. (2009). Understanding, Scoping and Defining User Experience: A Survey Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 719–728). New York, NY, USA: ACM. <https://doi.org/10.1145/1518701.1518813>
- Lloyd, A. (2016, February 25). Our friends, the bots? Retrieved March 13, 2018, from <https://points.datasociety.net/our-friends-the-bots-34eb3276ab6d>
- Ly, K. H., Ly, A.-M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*, 10, 39–46. <https://doi.org/10.1016/j.invent.2017.10.002>
- Meisel, W. (2016). Specialized Digital Assistants and Bots. Vendor Guide and Market Study. Retrieved March 13, 2018, from <http://tmaa.com/specializeddigitalassistantsandbots.html>
- Mifsud, J. (2011, July 11). Difference (and Relationship) Between Usability And User Experience. Retrieved March 13, 2018, from <https://usabilitygeek.com/the-difference-between-usability-and-user-experience/>
- Mortensen, D. (2018). Best Practices for Qualitative User Research. Retrieved March 13, 2018, from <https://www.interaction-design.org/literature/article/best-practices-for-qualitative-user-research>
- Nielsen, J. (1993). *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

- Nielsen, J. (1994). Usability Inspection Methods. In J. Nielsen & R. L. Mack (Eds.) (pp. 25–62). New York, NY, USA: John Wiley & Sons, Inc. Retrieved from <http://dl.acm.org/citation.cfm?id=189200.189209>
- Nielsen, J., & Molich, R. (1990). Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 249–256). New York, NY, USA: ACM. <https://doi.org/10.1145/97243.97281>
- Nielsen, Jakob. (2011). Mobile UX Sharpens Usability Guidelines. Retrieved March 13, 2018, from <https://www.nngroup.com/articles/mobile-sharpens-usability-guidelines/>
- Nielsen, Jakob. (2012). Usability 101: Introduction to Usability. Retrieved March 13, 2018, from <https://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- Norman, K. L., & Panizzi, E. (2006). Levels of Automation and User Participation in Usability Testing. *Interact. Comput.*, 18(2), 246–264. <https://doi.org/10.1016/j.intcom.2005.06.002>
- Obrist, M., Roto, V., & Väänänen-Vainio-Mattila, K. (2009). User Experience Evaluation: Do You Know Which Method to Use? In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (pp. 2763–2766). New York, NY, USA: ACM. <https://doi.org/10.1145/1520340.1520401>
- Petrie, H., & Bevan, N. (2009). The evaluation of accessibility, usability and user experience. *The Universal Access Handbook*.
- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact Evaluation in Information Systems Design Science Research? A Holistic View. In *PACIS 2014 Proceedings*

- *Pacific Asia Conference on Information Systems* (p. Paper 23). X, France.
Retrieved from <https://hal.archives-ouvertes.fr/hal-01126545>
- Rohrer, C. (2014). When to Use Which User-Experience Research Methods. Retrieved February 1, 2018, from <https://www.nngroup.com/articles/which-ux-research-methods/>
- Rubin, J., & Chisnell, D. (2008). *Handbook of Usability Testing* (2nd ed.). Wiley.
Retrieved from <http://shop.oreilly.com/product/9780470185483.do>
- Saunders, M., Lewis, P., & Thornbill, A. (2009). *Research Methods for Business Studies* (5th ed.). Pearson United Kingdom. Retrieved from https://www.researchgate.net/publication/201382148_Research_Methods_for_Business_Studies
- Schotanus-Dijkstra, M., ten Klooster, P. M., Drossaert, C. H. C., Pieterse, M. E., Bolier, L., Walburg, J. A., & Bohlmeijer, E. T. (2016). Validation of the Flourishing Scale in a sample of people with suboptimal levels of mental well-being. *BMC Psychology*, 4, 12. <https://doi.org/10.1186/s40359-016-0116-5>
- Sharon, T. (2013). The Rainbow Spreadsheet: A Collaborative Lean UX Research Tool. Retrieved March 13, 2018, from <https://www.smashingmagazine.com/2013/04/rainbow-spreadsheet-collaborative-ux-research-tool/>
- Shevat, A. (2017). *Designing Bots*. O'Reilly Media. Retrieved from <http://shop.oreilly.com/product/0636920057741.do>
- Shi, Q. (2010). *An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective* (PhD). Copenhagen Business School, Copenhagen.

- Shneiderman, B., Plaisant, C., Cohen, M., & Elmqvist, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson Education.
- Smith, S. L., & Mosier, J. N. (1986). *Guidelines for Designing User Interface Software*. Bedford, MA: The MITRE Corp.
- Tarazi, S. (2017). Exploring the different types of chatbots. Retrieved March 13, 2018, from <https://www.eila.io/single-post/2017/08/31/Exploring-the-different-types-of-chat-bots>
- Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Info. Sys. Research*, 3(1), 36–59. <https://doi.org/10.1287/isre.3.1.36>
- Wilcox, B., & Wilcox, S. (2013). Making It Real: Loebner-Winning Chatbot Design. *Arbor*, 189(764), a086.
- Xiao, J., Stasko, J., & Catrambone, R. (2002). Embodied Conversational Agents as a UI Paradigm: A Framework for Evaluation.

9 Appendices

Appendix 1: Usability Test Checklist

Usability Test Checklist

Pre-test activities

- ☐ Write down test hypothesis
- ☐ Form scenarios and tasks for the test
- ☐ Recruit participants
- ☐ Schedule Sessions

Before each session

- ☐ Make sure you know the name of the participant
- ☐ Print out "Tasks and Scenarios" for participant
- ☐ Check if you recording software works
- ☐ Turn off any unnecessary software and notifications on test device

During each session

- ☐ Welcome the participant and introduce yourself
- ☐ Explain the reason for the session
- ☐ Explain "think aloud" protocol
- ☐ Give participants "Permission for recording" form
- ☐ Turn on screen recorder software
- ☐ Ask, easy-to-answer, introduction questions to ease up situation
- ☐ Read first Task/Scenario and hand it to participant
- ☐ Ask participant about their questions, explain that you won't be able to answer during the test
- ☐ Thank (and compensate) user for participation

After each session

- ☐ Make sure you have all the documents signed
- ☐ Make a backup copy of the recording
- ☐ Analyze/review the recording as soon as possible

Appendix 2: Consent Form

Consent Form – Usability Testing “Gustav”

Thank you for participating in this usability test. I will be recording your session to improve the design of “Gustav” and to allow the evaluators of my Master Thesis to see the results after the fact. This recording will only be used for my Master Thesis and not in any other shape or form other than that.

Please read the statement below and sign where indicated.

I agree to participate an audio, video, and/or digital recording during the study conducted by Zeljko Maric, Student at Copenhagen Business School.

I understand and consent to the use and release of the recording by Zeljko Maric. I understand that the information and recording is for research purposes only and that my name and image will not be used for any other purpose.

I understand that participation in this usability study is voluntary and I agree to immediately raise any concerns or areas of discomfort during the session with the study administrator.

Signature: _____

Print your name: _____

Date: _____





Appendix 3: User Scenario

Scenario Task

You heard from a friend that there is this new Chatbot on Facebook Messenger called Gustav. He sais it's it can teach you a little bit about procrastination and productivity. He sends you a link to the facebook page. Please start chatting with Gustav and try to find out more about "Procrastination".

Appendix 4: Pre-Test Questionnaire

What Smartphone Model are you using? *

Your answer _____

How much experience do you have with chatbots? *

- ☐ None
- ☐ Some (used a chatbot more than once)

I believe that chatbots could be valuable teaching about self help topics like "procrastination" or "productivity". *

	1	2	3	4	5	
Strongly agree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly disagree

In case you have some experience: which chatbots did you use so far?

- ☐ Facebook Messenger
- ☐ Kik
- ☐ Slack
- ☐ Website
- ☐ Telegram
- ☐ Other: _____

How old are you? *

Your answer _____

What is your gender? *

- ☐ Female
- ☐ Male

NEXT

Page 1 of 2

Appendix 5: Post-test Questionnaire

Post-Test Questionnaire

I thought the information I got from the chatbot was useful. *

12345

Strongly disagreeStrongly agree

Why or why not?

Your answer

I can do everything I would expect to be able to do with a chatbot for self-help. *

12345

Strongly disagreeStrongly agree

What was one thing you missed?

Your answer

Overall the chatbot was easy to use. *

12345

Strongly disagreeStrongly agree

What did you like about it?

Your answer

What did you dislike about it?

Your answer

BACKSUBMIT

Page 2 of 2

Appendix 6: Usability Test Notes

Participant:

Date 13.01.18

Time: 17:00

Task

Path/Block

Issue

Observation, comments & notes

1

Welcome

+ how it's called "uninstall"
+ klar und verständlich
- too funny
- behaltene Wort → für und Text
+ beide verständlich
+ schneller v. "offen".
Bug: the thing
- funny
+ home. Behalt. Polster
- too long text →

1

✓

By: the thing is missing

2

Proclamation

gives back up again to reveal
- completed to install

Shorthand Codes

✓	Completed correctly	X	Incorrect Action	E	Error	#?	Probe during debriefing
P	Prompted by Mod	-	Negative sentiment	?	Participant confused	+	Positive sentiment

Appendix 7: Usability Test + Debriefing Interview Summary Spreadsheet

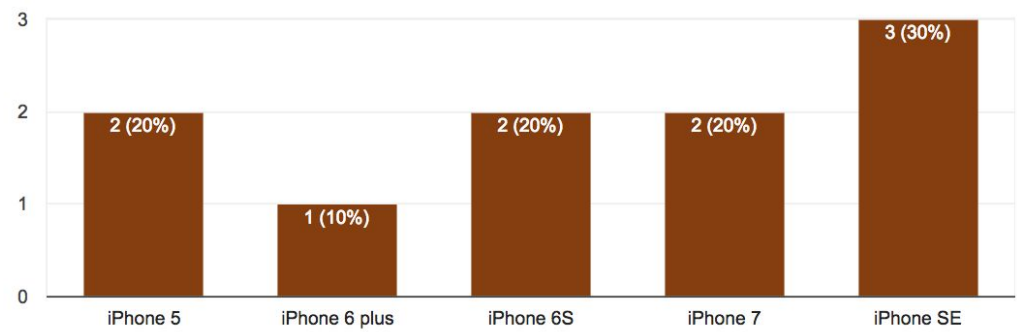
Code	Observation/Issue/Note	Score	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Quotes	Insights
	Negative Remarks													
NR1	Too much content that gets sent too fast	18											P7: "Whoa" P2: "You just get bombarded with those messages"	-> slow down content, make it human spee
NR2	Chatting with bot lacks the freedom of a conversation	3											P2: "It's not really chatting but rather funny way of reading" P4: "Why can't I answer di"	-> enable as much free conversation as pc
NR3	Quick replies too similar	3											P3: "What is the difference?"	-> clearly differentiate quick replies
NR4	Does not know how to end the conversation	1											P3: "The only way to end it is to write some gibberish and then get thrown into menu an"	-> educate more about conversation flow
NR5	Bot is never reading the last goodbye message	1											P10: "Where is he gone?"	
NR6	Bot cannot pickup where left once you fall out of a path	1											P10: "Feels a bit like going through the motion just to to get back where we left off"	-> enable hooks?
NR7	Copy is sometimes a little too funny	1												-> stickle balance between humor and sent
NR8	Content sometime not to the point enough	1												-> be concise
	Positive Remarks													
PR1	Humour in copy, replies and media is very good	14											P2: "This is really funny" P7: "He has a good sense of humor"	-> make bot reliable human-like but not
PR2	Content is valuable and well formulated	8											P1: "Very nicely summarised." P6: "Wow that's interesting" P10: "Very nice did learn"	-> try to find out what move the target audi
PR3	Asking for consent before signing up feel good	2											P10: "Ah that's nice he's asking again"	-> ask for consent, dont automate
PR4	Quick replies are popular	1											P4: "I like the replies"	-> use quick replies more than cards
PR5	Emojis are popular	1											P5: "Ha, I like that"	-> use emojis
PR6	Free writing give feeling of responding	1											P8: "Now it's cool I have the feeling that Gustav is responding to me"	
PR7	Likes the picture with Natalie Portman	5												
	Errors													
E1	Starts free writing and breaks the flow but gets caught by default message	5											P3: "Just wanted to see what he is able to do"	-> before sending media think good about
E2	Clicks of picture received, edits the picture in messenger and sends back, breaks the flow	1											P10: "Didn't think much about it"	
E3	Clicks on message instead of quick reply	1												
	Bugs													
B1	After free writing quick reply and keyboard overlap	4												
B2	Short freeze but then continued	2												
	Other													
O1	Clicks on picture send tries to read it	7												
O2	Thinks that the menu cards are too big and harder to understand compared to quick replies	1												

Note: the entire spreadsheet is attached in the upload.

Appendix 8: Pre- and Post-Test Questionnaire Summary

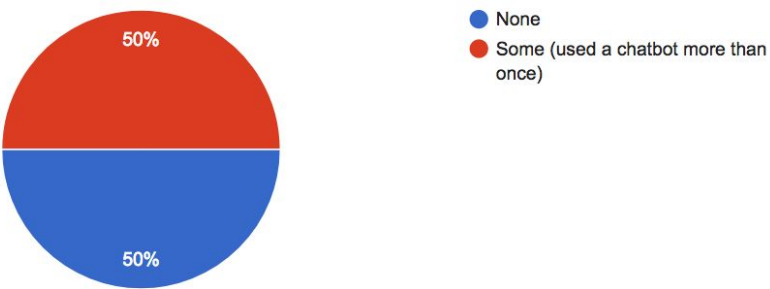
What Smartphone Model are you using?

10 responses



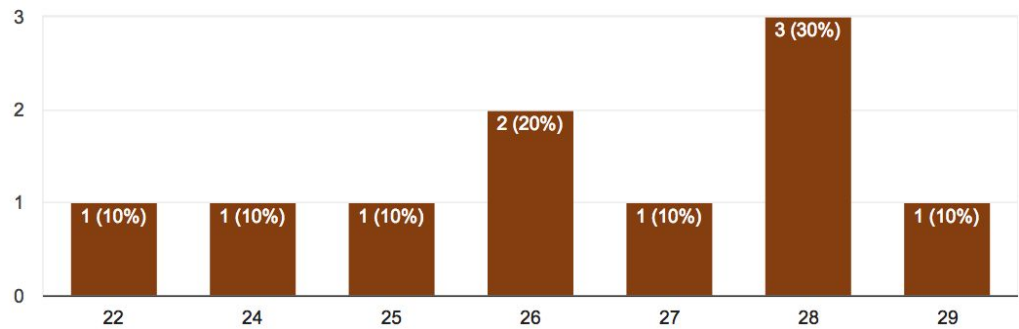
How much experience do you have with chatbots?

10 responses



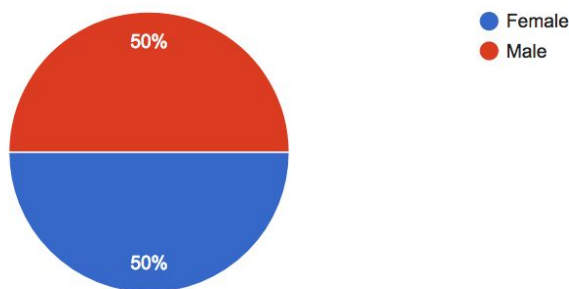
How old are you?

10 responses



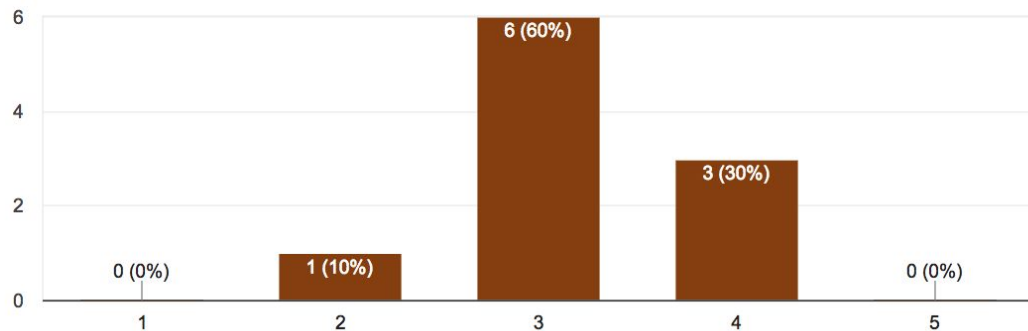
What is your gender?

10 responses



I can do everything I would expect to be able to do with a chatbot for self-help.

10 responses



What was one thing you missed?

10 responses

I still want to experience the nudging and help over the next days to see if it's more helpful than this one time session...

Maybe more options to enter my own "goals" or "tasks" to then be analysed...

Maybe using more dynamic media such as gifs, eg to explain the concept of SMART goals.

Lack of free answers.

Choices

More conversational with free writing + interactive excercises.

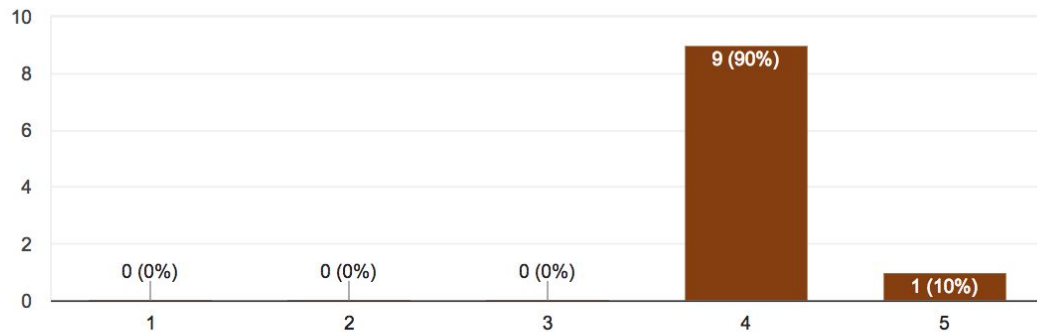
Gustav talks a lot and I always have to listen. Let me talk and him listen a bit too! Maybe he can learn from us humans too?

More "real" interaction. Step by step assignments. A program "he" will guide me through, day by day. Like in the morning "what are your SMART goals for today" and then in the afternoon, what did you achieve today. Also motivation if you do not achieve your goals, but to stay positive and focus on the fact that you are learning to be productive.

Maybe more choices during the conversation.

I thought the information I got from the chatbot was useful.

10 responses



Why or why not?

8 responses

was a good combination of fun and helpful stuff...

Specific answers, no long texts.

Through learning about procrastination, I actually was reminded to be more productive in the imminent moment- helping me utilize the tips Gustav gave me, or plan a structure in how to tackle a plan I had prepared.

The SMART acronym and the information about future-self with was new

I got taught a concept not known before in a blink. Gustav read my mind, when I suggested to wrap up a complex concept in 2-3 main takeaways. How anticipative of him!

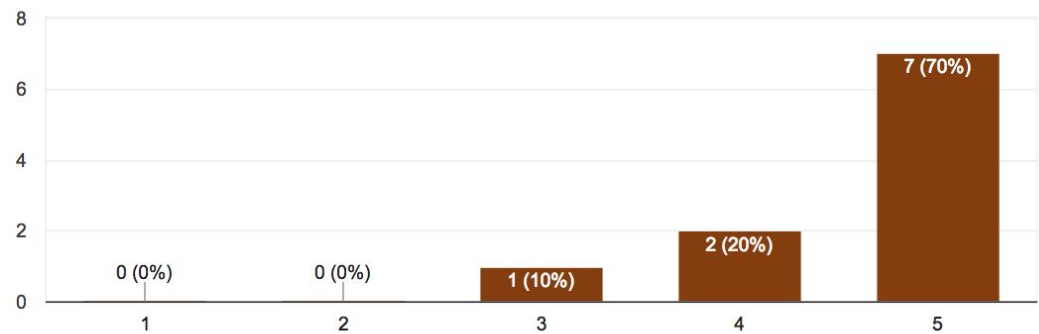
Simple suggestions, that could probably work. The SMART acronym sounds too smart, to business management class.

The conversation was too brief to transport more information.

I think I will actually go home and break my list of tasks into smaller bits. It's not that that is new knowledge - just easy to forget, and good to be reminded.

Overall the chatbot was easy to use.

10 responses



What did you like about it?

10 responses

Copy was great
Language and tone, visual material and text in good balance!
Straight forward options I could choose from
I enjoyed the humour, helped me forget Gustav was a robot. I likewise enjoyed the break-down of an abstract emotion in order to understand it better.
humor
Straight forward and very easy to use.
easy interaction with pre-filled answers
The tone.
The conversational approach really keeps you more attentive to the text.
humour, pathways were clearly laid out,

What did you dislike about it?

10 responses

getting back to the menu / other content was hard and sometimes it was a lot of text at once that made me scroll up and down...

sometimes a bit too fast

2-3 text bits came one after another which was too quick for me.

I found the chatbot provided too many predetermined answers. The content of information afterwards had the tendency to overwhelm the reader slightly.

too much information too quick and a lack of answer choices

Some times it lags and the elements are jumpy. And it's too much text for an iPhone 5 sometimes.

be more straight-forward in your answers

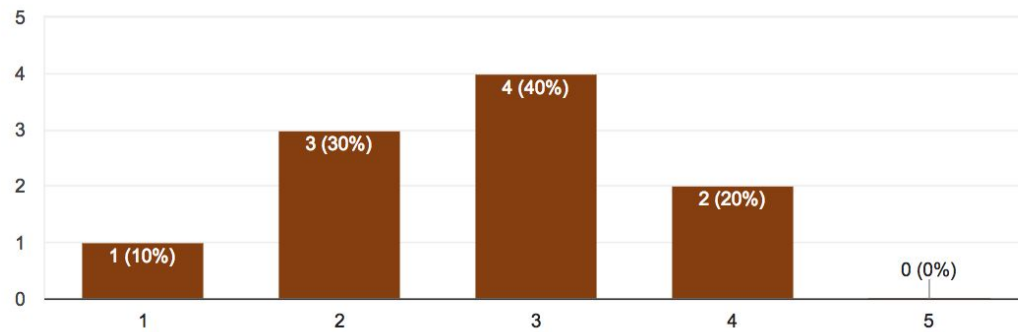
When a lot of messages came at one they came too fast

Sometimes the information pops up too fast, so it feels like reading a longer text.

that there are still constraints that remind oneself that chatbots are not as smart as humans

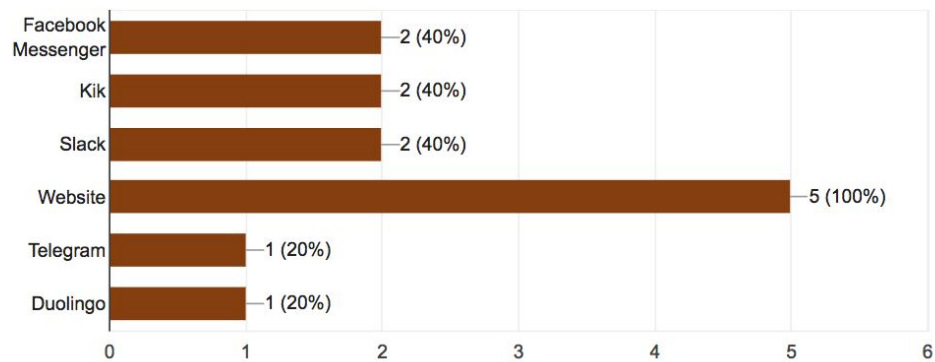
I believe that chatbots could be valuable teaching about self help topics like "procrastination" or "productivity".

10 responses



In case you have some experience: which chatbots did you use so far?

5 responses



Appendix 9: Usability Recordings

All 10 usability tests were recorded and the videos can be watched through the following link

<https://goo.gl/xDMsNc>