Towards Real-Life Adoption of Conversational Interfaces: Exploring the Challenges in Designing Chatbots That Live up to User Expectations

Luc van der Zandt
Drillster BV
De Meern, The Netherlands
luc@drillster.com

Esther van der Stappen
Avans University of Applied Sciences
Breda, The Netherlands
ej.vanderstappen@avans.nl

Koen van Turnhout
HU University of Applied Sciences
Utrecht, The Netherlands
koen.vanturnhout@hu.nl

Chatbots are increasingly popular, but state-of-the-art chatbots still struggle to meet user expectations, limiting their application in many domains. The factors affecting use have been studied extensively in laboratory contexts, resulting in context-independent requirements. However, user expectations and experiences of chat interfaces are affected by the context of use. Research efforts measuring experiences with chat interfaces need to shift from studies in controlled laboratory settings to studies in real-life settings in various domains. This paper explores this field of study by reporting on a small-scale real-life case study on the gap between expectations and experiences with an educational chatbot. More case studies in the wild, such as this one, could contribute to a deeper understanding of factors affecting acceptance and real use. We propose the use of the CIMO logic across these studies to build upon previous results.

Conversational User Interfaces, Educational Chatbots, Expectations–Experience Gap, Real-life, Case Study

1. INTRODUCTION

There is a long-standing tradition of research on natural language interaction with computers, dating back to the development of the famous ELIZA chatbot in the 1960s (Weizenbaum, 1966). This conversational interaction paradigm has had ups and downs over the years but is currently on the rise because of increased maturity of speech and language technology, the availability of speech-based assistants such as Alexa and Google Home on the market, and – particularly for text-based interfaces – the fact that the public is nowadays completely comfortable with communicating with short text messages such as with Facebook Messenger and WhatsApp (Dale, 2016). The design of conversational agents has been researched extensively. Recent studies focus, for example, on challenges like dialogue design and embodiment of chatbots (Foster, 2019; Fischer et al., 2019), specific design requirements for different roles and domains of chatbots such as comedy (Perone and Edwards, 2019), news delivery (Dubiel, Cervone and Riccardi, 2019), therapy and tutoring (George, 2019), and safeguarding qualities in interaction with chatbots

such as trust (Rheu et al., 2020) and engagement (Candello et al., 2019).

However, as conversational technology is maturing and applied more and more in real-life contexts, it has become evident that a mismatch between user expectations and experience exists, limiting user adoption. This notion became mainstream in the field of study to conversational agents since the 2016 publication by Luger and Sellen (2016), which mentions a ‘gulf of evaluation’ for conversational agents. Luger and Sellen found that users constructed poor mental models about the workings of the interface and, as such, set unrealistic expectations for it. This mismatch between the user expectations and the actual system may be particularly urgent in language interfaces. Unlike graphical user interfaces, they do not convey much information about their action possibilities in their appearance.

Later studies in controlled laboratory settings have focused on discovering more concrete factors affecting use and generic requirements for developing chat interfaces. Weber and Ludwig (2020), for example, identified user needs for chatbots. Among others, they elaborate on the need for initial explanation and guidance, natural

conversation, the ability of the agent to remember the context of the conversation, customisation of the interaction, alternative ways of contact, technical robustness, and keeping in control of if, and when, the agent proactively contacts the user through (push) notifications. In turn, Følstad et al. (2018a) focused on the factors specifically influencing trust in conversational agents for customer service. This study found that users value a chatbot's ability to understand the user, human-likeness, self-preservation, professional appearance, and security and privacy. As the authors of these studies acknowledge, such lab studies are suitable for uncovering generic requirements for conversational interfaces, but they lack sensitivity for needs that arise out of the context of use. This usage context is expected to be an important determinant of user expectation which warrants case studies that particularly explore those contexts. Mapping expectations and experiences of chat interfaces in those contexts may bring forward different requirements and priorities than those previously conceived. As such, this paper argues that research efforts measuring experiences with chat interfaces need to shift from studies in controlled laboratory settings to studies in real-life settings from a variety of domains.

The application of chat interfaces in real-life settings in the educational domain is a particularly unexplored area in the existing literature which may hold perspectives for future design efforts within and outside this domain. This paper explores this field of study by reporting on a small-scale real-life case study on the gap between expectations and experiences with an educational chatbot. In the following sections, we describe the used methods and our findings. We end the paper with a call for further real-life case studies and propose a method for knowledge-building across these cases.

2. METHODS

In the case study described in this paper, we realised and evaluated a chat interface to support learning using an existing adaptive learning system called Drillster. Drillster is a question-based adaptive learning tool. It uses a proprietary algorithm that incorporates elements of graduated interval recall (Pimsleur, 1967) and the Leitner system (Leitner, 1972) to battle the 'forgetting curve', a term coined by Hermann Ebbinghaus (1885) to describe the gradual decline of memory retention over time. Users can create, share and do exercises called 'drills' to gain and retain knowledge. The company states that "incorrectly answered questions are repeated more often than correctly answered ones" (Drillster BV, 2019). The platform predicts the forgetting curve of the user and, based on this data, brushes up the knowledge that is likely to be forgotten soon. A chat interface would support

users by actively reminding and involving them to gain and retain knowledge via a familiar, highly used interface. After its realisation, we introduced the interface in a real-life educational setting to 33 Dutch high school students aged 14 to 18, who used and evaluated it in their daily educational tasks in the school subjects Greek, Latin and biology over two weeks. Prior to the intervention, the participants already used Drillster for their education via the conventional interfaces, i.e., web and mobile.

2.1 Gathering User Expectations

To determine what these potential users expected from a chat interface for the Drillster platform, we provided our subjects with a survey of four open questions asking how the subjects thought about a chatbot for the Drillster platform, what positives and negatives they expected from it, and what effect they expected that the interface would have on their frequency of use. This approach provided us with insight into the expectations and wishes of the subjects but did not establish specific functional requirements, as the functional outline was determined by the learning platform and were beyond the scope of this study. We asked the following four questions (translated from Dutch):

- What are your thoughts on a chatbot to practise drills? Please write down what first comes to mind.
- What are your hopes on what a chatbot can do for you?
- What are your hopes on what a chatbot will certainly not do?
- Do you think you will use Drillster more active, less active, or with an equal frequency because of a chatbot? Why?

Using deductive thematic analysis as described by Braun and Clarke (2012), we coded the responses to these questions and categorised them into themes. The approach describes six phases: familiarisation with the data, generating initial codes, searching for themes, reviewing potential themes, defining and naming themes, and producing the report.

2.2 Gathering User Experiences

After initial familiarisation with the users and their expectations, a large part of the study consisted of designing and realising the prototype. The prototype evolved in iterations with the study's knowledge acquisition, using an action design-based approach (Sein et al., 2011). When the prototype reached a usable state, we asked the participants to use the prototype for two weeks. This comprised practising drills for the concerning subject that were pre-made by the teacher. Consequently, we evaluated it using a questionnaire constructed with the modular

extension of the User Experience Questionnaire (UEQ+) (Laugwitz et al., 2008, Schrepp and Thomaschewski, 2019). We could map the user expectations discovered by our thematic analysis to the actual experience found by analysing responses to the UEQ+. The evaluation used the following nine scales: attractiveness, efficiency, perspicuity, dependability, stimulation, novelty, trust, usefulness, and intuitive use, which could be evaluated on a scale from -3 to +3. We chose the UEQ+ in favour of other models such as TAM (Davis, 1989), TAM2 Venkatesh and Davis, 2000), and UTAUT (Venkatesh et al., 2003), because these instruments include external (organisational) factors influencing actual usage. The present study mainly focuses on intrinsic user intentions and motivations, making a questionnaire solely focusing on these factors more appropriate. Therefore, we advocate using the User Experience Questionnaire (UEQ) and its modular extension, the UEQ+, in the present study.

3. RESULTS

To evaluate our prototype, we asked students to use the prototype for two weeks to learn for school subjects that already incorporated Drillster – Greek, Latin and Biology. After these two weeks, we asked the students to evaluate the prototype using the UEQ+. Twenty-four responses came in ($N=24$). We detected major inconsistencies among the answers to the items in the scale 'trust' ($\alpha=.1$). Because of this, we omitted the scale 'trust' from the results. Careful comparison between the pre-test and post-test results provided insight into the prototype's attributes on which the expectations and experience differed the most.

3.1 An Efficient, Fast and Stimulating Interface

The results indicate a distinct difference between expectations and experience on efficiency and stimulation. In the general question about expectations in the pre-test, respondents named these factors as a positive, more so than the other themes. Furthermore, a total of 22 out of 39 discovered codes on the question *What are your hopes on what a chatbot for the Drillster platform can do for you?* were categorised as positive expectations regarding 'efficiency' and 'stimulation'. However, the efficiency and stimulation scales received the two lowest mean evaluations in the post-test ($\bar{x}=.48$ and $\bar{x}=.47$). Participants particularly leaned toward rating the application as slow ($\bar{x}=.25$), while the pre-test showed they hoped the chatbot would help them learn more efficiently (6/16 codes categorised in 'stimulation') and feared it would be slow (6/12 codes categorised in 'dependability'). These results indicate that students had high expectations regarding the chatbot's effects on their learning efficiency and stimulation, but the chatbot could not live up to them.

3.2 Trust and Transparency

Looking at the pre-test results, a big concern of respondents was that the chatbot would send spam messages or unsolicited notifications (15/36 codes on the question *What are your hopes on what a chatbot will certainly not do?*). Furthermore, some feared it would invade their privacy (2/36 codes) and be slow or dumb (12/36 codes). We could not accurately measure the 'trust' scale in the post-test, but the pre-test results indicate that our implementation should comfort the user regarding spam messages and data access and answer the user's queries as fast as possible.

3.3 Straightforward Navigation and Overview of Capabilities

Furthermore, the pre-test responses indicated that respondents hoped the chatbot would enable easy and accessible learning (10/39 codes on the question *What are your hopes on what a chatbot can do for you?*). While the chatbot's perspicuity scored relatively well in the post-test ($\bar{x}=1.15$), the fact that it was such a prominent positive expectation for the chatbot in the pre-test indicates that perspicuity is an important factor influencing the user experience of our prototype. Improvements regarding perspicuity may be beneficial to the user experience and acceptance.

3.4 Interface Flexibility

Apart from the fear that a chatbot would be slow, respondents also indicated that they feared a chatbot would need specific input. We can categorise responses into two categories: (1) needing specific input for chatbot instructions and (2) needing specific answers to practice questions. The need for user error correction seems particularly apparent for text-based interfaces, such as chatbots, where typos can easily be made. While respondents positively evaluated the chatbot's dependability ($\bar{x}=1.35$) in the post-test evaluation, anticipating user errors by correcting their input to the extent that the learning effect is not affected may improve the acceptance of our prototype.

3.5 An Authentic Messaging Experience

Lastly, our results indicate that students open and use messaging apps more often than a Drillster client application and, thus, would be motivated to use a Drillster interface on such apps more often than the usual Drillster client applications (18/33 responses on the question *Do you think you will use Drillster more active, less active, or with an equal frequency because of a chatbot?*). However, the post-test evaluation results indicate that the use of the chatbot did not feel intuitive, as the mean evaluation of the scale 'intuitive use' is the fourth-lowest rated scale in the post-test ($\bar{x}=.9$). This

suggests the interface did not quite feel like an authentic messaging experience, and the integration with the messaging platform could have been more seamless.

4. CONCLUSIONS AND DISCUSSION

The present study reports on a small-scale real-life case study of an educational chatbot. It aims to contribute to a deeper understanding of the expectations and experiences of chatbot users in real-life settings. We elaborate on the scarce existing work regarding motivations to use chatbots in the wild and draw conclusions regarding the gap between expectations and experience observed in this case study in relation to controlled laboratory studies and previous studies regarding general-use and customer service chatbots. For this, we realised a chatbot for an e-learning platform and asked a group of 33 high school students about their expectations and asked them to evaluate the application after using it, for which 24 responses came in.

4.1 Discussion of Results

Our findings align well with existing work. We found that respondents did not really know what to expect but generally had an overall positive attitude towards the idea of a chatbot. Next to stimulation to learn, respondents indicated they hoped the chatbot would benefit their efficiency by being able to learn easier and faster. This sentiment of a chatbot aiding general task efficiency is shared by Følstad et al. (2018b, p. 7) and Luger and Sellen (2016). However, these studies did not find motivations regarding stimulation of use, which is a big motivator in the present study. We could explain this difference because our case is in the educational domain, where task motivation can generally not be taken for granted. The cases of Følstad et al. and Luger and Sellen focus on chat assistants in the general domain (i.e., to manage everyday tasks) and customer service, respectively. In these cases, task motivation may be more present or even be taken for granted.

We found that respondents feared the chatbot would send spam messages, a finding also reported by Weber and Ludwig (2020, p. 325). Respondents indicated they feared the chatbot would not understand their input, which was also called a challenge by Følstad et al. (2018a). This study also named straightforwardness as a perceived benefit, but we did not discover this in the present study. We could explain this discrepancy by the difference between chatbots' application in customer service, on which Følstad et al. focused, and the educational domain.

While the responses to the questions on expectations indicated students expected the chatbot would benefit their efficiency and stimulation, these scales were rated lowest in the subsequent evaluation. This finding suggests that a substantial gap exists between expectations for, and experience with, conversational user interfaces within the real-life context of our case study. This finding matches that of Luger and Sellen (2016), who found users failed to construct adequate mental models of the intelligence, capabilities, and goals of conversational agents in the general domain. In our case study, this seems to be focused on the capability of the chatbot to improve learning efficiency and stimulation, as a motivator to use a chatbot is to study faster, i.e., more efficiently.

4.2 Directions for Future Research

This study started to explore the design space of chatbots in real-life contexts by realising and evaluating an educational chatbot. While we believe our findings can, to an extent, be valuable to future design efforts, the validity of the results of this single case study with few respondents is limited. More case studies in the wild, such as this one, could contribute to a better understanding of context-dependent factors affecting the real use of chat interfaces.

For effective knowledge build-up across case studies, it is important to be very explicit about the scope of individual findings. Therefore, we propose using the CIMO logic (Denyer et al., 2008) to construct design propositions for future design and research efforts. The logic is as follows: "in this class of problematic contexts (C), use this intervention (I) type to invoke these generative mechanism(s) (M), to deliver these outcome(s) (O)" (Denyer et al., 2008).

We argue that using the CIMO logic in HCI holds several benefits. The components of the logic form a specification that can help others identify the applicability of the proposition to their case, making it easier to attribute different findings in different studies to a specific element of the CIMO specification. This is particularly valuable for knowledge build-up across multiple field studies where the precision of CIMO aids in diagnosing conflicting findings, leading to a refinement of the specificity, practical applicability, and robustness of the propositions; elements which are needed for all forms of prescriptive knowledge (van Turnhout et al., 2019).

New research and development challenges anchored in the contexts of use arise when novel technologies such as conversational interfaces reach maturity, sparking adoption in the real world. It is of key importance that we explore those situated challenges using field studies, like the one reported in this paper.

5. REFERENCES

- Braun, V. and Clarke, V. (2012) Thematic Analysis. In: Cooper, H. et al. (eds). *APA handbooks in psychology. APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological*. Washington: American Psychological Association.
- Candello, H. et al. (2019) Can Direct Address Affect Engagement with Chatbots Embodied in Physical Spaces? In: Cowan, B.R. and Clark, L. (eds). *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, August 2019*. New York: ACM. 1–3.
- Davis, F.D. (1989) Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13 (3), 319–340.
- Denyer, D., Tranfield, D. and Aken, J.E. van (2008) Developing Design Propositions through Research Synthesis. *Organisational studies*, 29 (3). 393–413.
- Drillster BV (2019) *The Science Behind Drillster*. Available from https://drillster.com/wp-content/uploads/2019/08/White-Paper-_The-science-behind-Drillster_.pdf (24 October 2020).
- Dubiel, M., Ceryone, A. and Riccardi, G. (2019) Inquisitive mind: a conversational news companion. In: Cowan, B.R. and Clark, L. (eds). *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, August 2019*. New York: ACM. 1–3.
- Ebbinghaus, H. (1885) *Über das Gedächtnis. Untersuchungen zur experimentellen Psychologie*. Leipzig: Decker & Humblot.
- Fischer, J.E. et al. (2019) Progressivity for voice interface design. *Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, August 2019*. New York: ACM. 1–8.
- Følstad, A., Nordheim, C.B. and Björkli, C.A. (2018a) What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In: Bodrunova, S. et al. (eds). *Internet Science. INSCI 2018. Lecture Notes in Computer Science, vol 11193*. Cham: Springer.
- Følstad, A., Skjuve, M. and Brandtzaeg, P.B. (2018b) Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design. In: Bodrunova, S. et al. (eds). *Internet Science. INSCI 2018. Lecture Notes in Computer Science, vol 11551*. Cham: Springer.
- Foster, M.E. (2019) Face-to-Face Conversation: Why Embodiment Matters for Conversational User Interfaces. In: Cowan, B.R. and Clark, L. (eds). *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, August 2019*. New York: ACM. 1–3.
- George, S. (2019) From sex and therapy bots to virtual assistants and tutors: how emotional should artificially intelligent agents be? In: Cowan, B.R. and Clark, L. (eds). *CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces, Dublin, August 2019*. New York: ACM. 1–3.
- Laugwitz, B., Held, T. and Schrepp, M. (2008) Construction and Evaluation of a User Experience Questionnaire. In: Holzinger, A. (eds). *HCI and Usability for Education and Work. USAB 2008. Lecture Notes in Computer Science, vol 5298*. Heidelberg: Springer. 63–76.
- Leitner, S. (1972) *So lernt man lernen. Der Weg zum Erfolg*. Freiburg im Breisgau: Herder.
- Luger, E. and Sellen, A. (2016) “Like Having a Really Bad PA”: The Gulf Between User Expectation and Experience of Conversational Agents. In: Kaye, J. and Druin, A. (eds). *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, May 2016*. New York: ACM. 5286–5297.
- Pimsleur, P. (1967) A Memory Schedule. *Modern Language Journal*, 51 (2). 73–75.
- Rheu, M. et al. (2020) Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human-Computer Interaction*, 37 (4). 1–16.
- Roda, C., Angehrn, A.A. and Nabeth, T. (2001) Conversational Agents for Advanced Learning: Applications and Research. *Proceedings of BotShow 2001 Conference, Fontainebleau, 2001*. Fontainebleau: INSEAD. 1–7.
- Schrepp, M. (2018) *UEQ_Plus_Data_Analysis_Tool.xlsx*. Available from http://ueqplus.ueq-research.org/Material/UEQ_Plus_Data_Analysis_Tool.xlsx (18 December 2020).
- Schrepp, M. and Thomaschewski, T. (2019) *Handbook for the modular extension of the User Experience Questionnaire*. Available from http://ueqplus.ueq-research.org/Material/UEQ+_Handbook_V1.pdf (18 December 2020).
- Sein, M.K. et al. (2011) Action Design Research. *MIS Quarterly*, 35 (1). 37–56.
- Turnhout, K. van et al. (2019) A Practical Take on Theory in HCI. White paper.
- Venkatesh, V. and Davis, F.D. (2000) A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, 46 (2). 186–204.

- Venkatesh, V. et al. (2003) User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27 (3). 425–478.
- Weber, P. and Ludwig, T. (2020) (Non-)Interacting with conversational agents: perceptions and motivations of using chatbots and voice assistants. *Proceedings of the Conference on Mensch und Computer, Magdeburg, September 2020*. New York: ACM. 321–331.
- Weizenbaum, J. (1966) ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9 (1). 36–45.