6-15-2020

Unleashing the Potential of Conversational Agents for Course Evaluations: Empirical Insights from a Comparison with Web Surveys

Thiemo Wambsganss
University of St. Gallen, thiemo.wambsganss@unisg.ch

Rainer Winkler
Institute of Information Management, rainer.winkler@unisg.ch

Pascale Schmid
University of St. Gallen, pascale.schmid@student.unisg.ch

Matthias Söllner
University of Kassel, matthias.soellner@unisg.ch

UNLEASHING THE POTENTIAL OF CONVERSATIONAL AGENTS FOR COURSE EVALUATIONS: EMPIRICAL INSIGHTS FROM A COMPARISON WITH WEB SURVEYS

Research paper

Thiemo Wambsganss, University of St.Gallen, St.Gallen, Switzerland, thiemo.wambsganss@unisg.ch

Rainer Winkler, University of St.Gallen, St.Gallen, Switzerland, rainer.winkler@unisg.ch

Pascale Schmid, University of St.Gallen, St.Gallen, Switzerland, pascale.schmid@student.unisg.ch

Matthias Söllner, University of Kassel, Kassel, Germany, soellner@uni-kassel.de & University of St.Gallen, St.Gallen, Switzerland, matthias.soellner@unisg.ch

Abstract

Recent advances in Natural Language Processing (NLP) bear the opportunity to design new forms of human-computer interaction with conversational interfaces. However, little is known about how these interfaces change the way users respond in online course evaluations. We aim to explore the effects of conversational agents (CAs) on the response quality of online course evaluations in education compared to the common standard of web surveys. Past research indicates that web surveys come with disadvantages, such as poor response quality caused by inattention, survey fatigue or satisficing behavior. We propose that a conversational interface will have a positive effect on the response quality through the different way of interaction. To test our hypotheses, we design an NLP-based CA and deploy it in a field experiment with 176 students in three different course formats and compare it with a web survey as a baseline. The results indicate that participants using the CA showed higher levels of response quality and social presence compared to the web survey. These findings along with technology acceptance measurements suggest that using CAs for evaluation are a promising approach to increase the effectiveness of surveys in general.

Keywords: Human-computer Interaction, Conversational Agents, Course Evaluation, Pedagogical Conversational Agents, Social Presence, Social Response Theory

1 Introduction

Nowadays, educational institutions face the challenge to adapt their curriculum ever faster for students to get them prepared for an increasingly volatile, uncertain, complex and ambiguous (VUCA) world (Fadel and Groff 2018). Technological innovation happens at a faster pace, job profiles in demand are adjusted continuously (vom Brocke et al. 2018) and therefore educational institutions need to adjust their courses accordingly. The rising number of courses and students confront lecturers with the difficulty to know students' needs and expectations and how they perceive the learning content (Blair and Valdez Noel 2014; Smithson et al. 2015; Spooren et al. 2013). In order to gain this information, course evaluation surveys are commonly used by educational institutions (Blair and Valdez Noel 2014). Here, online web surveys have developed as the standard format for course evaluations in most educational institutions, compared to paper-based surveys, since they are easy to distribute, simple to evaluate and less costly than paper-based approaches (Spooren et al. 2013). Despite the widespread use of this form of student feedback collection, previous research has shown that there are certain limitations of online surveys (Wambsganss, Winkler, Schmid, et al. 2020). Educational institutions are confronted with low acceptance and response rates, only time-related insights and low-quality answers in the open question sections that are hardly applicable for adapting courses to students' expectations (Erikson et al. 2016; Tucker et al. 2008). Explanations for these negative effects might be that student responses are affected by survey fatigue (Tucker et al. 2008) or respondents' satisficing behavior (Heerwegh and Loosveldt 2008; Kim et al. 2019). Using a method with a static interaction such as a web survey where it is difficult to control respondents likely leads to low-quality data (Kim et al. 2019; Wambsganss, Winkler, Söllner, et al. 2020), and this in return makes it difficult for educational institutions to adjust their courses to ever-changing environments. To address those issues, qualitative evaluation methods, such as individual interviews, are used to produce a higher quality of answers and to gain deeper insights (Steyn et al. 2019). However, these approaches are usually very resource-intensive since lecturers need to address every student individually, which is even more difficult in times of mass lectures such as massive open online courses (MOOCs) (Steyn et al. 2019).

One possible solution to benefit from the advantages of both – qualitative and quantitative – evaluation methods is using conversational agents (CAs). CAs are software programs which communicate with users through natural language interaction interfaces (Rubin et al. 2010; Shawar and Atwell 2005). Compared to traditional quantitative course evaluations, CAs are able to reach students on their everyday devices and build up a human-like interaction with them. CAs are able to adapt their answers to students' utterances and can therefore build up a meaningful dialog with the students, almost like a qualitative lecturer-student interview (Wambsganss, Winkler, Schmid, et al. 2020). Backing on social response theory (Moon 2000; Nass et al. 1994; Nass and Moon 2000), this form of human-computer interaction might encourage students to provide a higher quality of answers for lecturers to improve their courses. The popularity of CAs, such as Amazon's Alexa, Google's Assistant, Apple's Siri and other systems, has been steadily growing over the past few years (eMarketer 2017; Krassmann et al. 2018). The recent improvement in Natural Language Processing (NLP) and Machine Learning (ML) enables CA systems to ask and answer questions in natural conversation flows and use intelligent question answering to adapt to a certain task (Hobert and Wolff 2019). In education, CAs have been used for several purposes, such as to provide support for problem solving in mathematics (Aguiar et al. 2014), to mediate group learning processes during problem solving (Winkler et al. 2019), for collaborative language learning (Tegos et al. 2014) or for academic advising (Latorre-Navarro and Harris 2015). Existing research on CAs in education has mainly focused on providing learning support for students (Song et al. 2017). Research on CA support for lecturers is still scarce. Moreover, Winkler and Söllner (2018) emphasize that CAs might also have great potential as an evaluation tool for lecturers. Until now, only little empirical insight about how these CA interfaces influence the way how students evaluate the courses exist (Wambsganss, Winkler, Schmid, et al. 2020). A recent study by Kim et al. (2019) shows that a CA can perform part of a human interviewer's role by applying effective communication strategies and, therefore, encourages user engagement, which in return leads to high-quality quantitative data. Moreover, Wambsganss et al., 2020 firstly investigated the positive effect of a conversational interface on the response quality and the level of enjoyment in course evaluations. Addressing this gap, we aim to contribute to research and practice by investigating the effect of CAs on qualitative response data in course evaluations answering the following research question (RQ):

RQ: *How do conversational agents (CA) influence the response quality of online course evaluations compared to traditional web surveys?*

To answer our research question, we conducted a 2 x 2 field experiment design based on social response theory to test whether different interaction types (conversational vs. static) and different levels of anthropomorphism (low vs. high anthropomorphism) result in a higher response quality for course evaluations. We chose to conduct a field experiment to evaluate the different interaction types in different real-world scenarios to expand the validity of our results independent of lecture type and content. Drawing on social response theory (Moon 2000; Nass et al. 1994; Nass and Moon 2000), we designed an NLP-based CA and deployed a field experiment with 176 students in three different course formats in which participants were asked to provide feedback on the lecture content and the teaching style of the lecturer. We found that participants using a CA showed higher levels of social presence and the evlauation data showed higher levels of response quality. Moreover, the measured technology acceptance provides promising results to not only use our CA as a course evaluation tool in different learning settings but also rethink the use of CAs for other kinds of evaluations. The results suggest that CAs based on NLP may have a beneficial use for generating higher quality data for course evaluations and evaluations in general. Thereby, our study contributes to two different research areas in information systems: First, we contribute to the application of CAs in education, suggesting a successful use case to employ a CA with potential benefits for lecturers and educational institutions to better adjust their learning content based on high quality responses and potentially continuous student feedback. Second, we contribute to the human-computer interaction (HCI) and social response theory by measuring the influence of the interaction type and the level of anthropomorphism not only with self-reportings but also with more objective measures (e.g., sentiments and syntactical readability score). Our study suggests that highly anthropomorphistic design elements in the field of user-information-seeking scenarios such as course evaluation is not necessarily a benefit for response quality. The study also makes some practical contributions by exemplarily showing how educators can implement CAs as course evaluation tools in their learning environments, since they are widely available and easy to use from an educator's – non-programming expert's – point of view. Our research paper is organized as follows: First, we provide an overview of the theoretical background, the hypotheses development and our research question. Section 3 describes the experimental setup in more detail, including the procedure, the manipulations of the CA and the web survey, and the measurement and analysis of our constructs. Afterwards, we present the results for our research question. Then, we conclude with a discussion, limitations and future research. Finally, the paper closes with a conclusion.

2 Theoretical Background and Hypotheses Development

In this section, we provide an overview of CAs in education and discuss the current problems of course evaluations. Next, we will elaborate on social response theory as our theoretical lens on the way human-computer interaction is perceived. On that basis, we will develop our hypotheses which form our research model.

2.1 Conversational Agents in Education

CAs are software programs which are designed to communicate with users through natural language interaction interfaces (Rubin et al. 2010; Shawar and Atwell 2005). In today's world, conversational interfaces, such as Amazon's Alexa, Google's Assistant, Apple's Siri, are ubiquitous, with their popularity steadily growing over the past few years (eMarketer 2017; Krassmann et al. 2018). They are implemented in various areas, such as customer service (Hu et al. 2018; Xu et al. 2017), counselling (Cameron et al. 2017; Fitzpatrick et al. 2017), healthcare (Kowatsch et al. 2017; Laumer et al. 2019) or education (Kerly et al. 2007; Winkler and Söllner 2018). Hobert and Wolff (2019) define CAs used in education as a special form of learning application that interacts with learners individually. The development of CAs in education goes back to the 1970s research stream of Intelligent Tutoring Systems (ITS) (e.g., Atkinson and Shiffrin 1968; Suppes and Morningstar 1969). Similar to a human tutor, these systems can present instructions, ask questions and provide immediate feedback (Kulik and Fletcher 2016). ITS evolved from abstract entities with limited technological possibilities to systems that are able to interact with learners using multiple channels of communication, exhibit social skills and perform different roles, such as tutors (Payr 2003), motivators or learning companions (Kim and Baylor 2008). While existing research on CAs in education has mainly focused on providing learning support for students (Hobert and Wolff 2019; Song et al. 2017), Wambsganss, Winkler, Schmid, et al., 2020 pointed out that CAs might also have potential as an evaluation tool.

2.2 Course Evaluation in Education and Satisficing Behavior

Nowadays, course evaluations are a common feature used by higher educational institutions to deliver high-quality teaching and learning (Erikson et al. 2016; Steyn et al. 2019). Course evaluation surveys can provide valuable insights into teaching and course effectiveness, which allows institutions to react to changing student needs (Blair and Valdez Noel 2014; Steyn et al. 2019). They can be divided into quantitative and qualitative evaluation methods. The most frequent form of course evaluations are quantitative web surveys (Erikson et al. 2016). Respondents evaluate and respond to this evaluation form by themselves, which implies ease of measurement (Kim et al., 2019). However, this static interaction method has been broadly criticized. The reliability and validity of student feedback surveys is questioned (Spooren et al. 2013; Kim et al. 2019) due to problems such as survey fatigue (Tucker et al. 2008) or respondents' satisficing behavior (Krosnick 1991). Satisficing behavior is reflected in non-differentiation or straight lining, meaning an equal responsive behavior is used for an array of scaled questions (Kim et al. 2019). This behavior occurs because responding accurately and sincerely has a high level of cognitive demands (Krosnick 1999). Satisficing responses lead to response errors, thus producing data of lower quality (Heerwegh and Loosveldt 2008). A possibility to reduce satisficing behavior is qualitative course evaluations. According to Steyn et al. (2019), qualitative course evaluation has the potential to overcome the disadvantages of quantitative course evaluations. Qualitative course evaluation like class discussions promote conscientious responses and encourage participation through social presence (Kim et al. 2019). Having an interviewer can encourage students to participate in a survey, ask for clarification and check their answers to confirm their sincerity. These verbal and nonverbal interactions between students and lecturers promote accurate answers, which increases the feedback quality (Holbrook et al. 2003). However, lecturers may be reluctant to use qualitative course evaluation methods as they are not scalable, take longer to analyze and are resource-intensive (Steyn et al. 2019). CAs as a new survey method might be able to reduce the disadvantages of traditional quantitative and qualitative evaluations. Compared to qualitative evaluations, CAs are available at any time, can speed up response times and are less resource-intensive (Winkler and Söllner 2018). Furthermore, compared to quantitative evaluations, CAs can react to students' responses and therefore create interactivity. Lundqvist et al. (2013) show that the use of CAs provide deeper information about the users' opinions than normal Likert style surveys because of follow-up questions. Through reciprocal message exchange, conversational interactivity decreases respondents' satisficing behavior, thereby producing high-quality data (Kim et al. 2019). Kim et al. (2019) show in their study that participants in a CA survey were more likely to produce differentiated responses, which resulted in higher-quality data than responses from participants in a web survey. Moreover, Wambsganss et al., 2020 firstly investigated the positive effect of a conversational interface on the response quality and the level of enjoyment in course evaluations. Nevertheless, transferable insights and empirical proof about the influence of CAs and the design on data quality in the context of online course evaluations is still scarce in literature. Therefore, we argue that there is a need to better understand whether conversational interfaces result in a higher response quality for course evaluations. To evaluate the effect of different interaction types on online course evaluations, we formulate the following hypothesis:

H1a: *A CA will produce a higher response quality compared to a web survey.*

We measure response quality based on three constructs: self-reported response quality by the user, syntactic readability based on the Flesch-Readability score (Flesch 1943) and the intensity of sentiments in the answers (e.g., Pang and Lee, 2008). Self-reported response quality and the intensity of sentiments (measurement further explained in section 3) is based on a user self-disclosure. According to Kim et al. (2019), self-disclosure is the most commonly used standard for evaluating data quality. Literature shows contradictory findings about respondents' self-disclosure depending on the situation. Participants show an increased willingness to disclose sensitive information, e.g., in mental health contexts, when they believe that they are interacting with a computer and not with a human operator (Lucas et al. 2014). According to Lind et al. (2013), participants expose more sensitive information in the absence of a humanized interface as they perceive greater anonymity. On the other hand, a study by von der Pütten et al. (2010) revealed that a wordy agent facilitated self-disclosure.

2.3 Social Response Theory

Our research is motivated by social response theory. According to social response theory, humans tend to respond socially to agents that display characteristics similar to humans (e.g., to animals or technologies) (Moon 2000). Behavioral cues and social signals from computers, such as interacting with others, using natural language or playing social roles, subconsciously trigger responses from humans, no matter how rudimentary those cues or signals are (Nass et al. 1994; Nass and Moon 2000). Following the “Computers are Social Actors” (CASA) paradigm, existing research has examined different social cues and their influence on HCI. According to Tung and Deng (2006), students perceive a higher degree of social presence and social attraction in an active-interactivity environment than in a passive-interactivity environment. Also, Schuetzler et al. (2014) showed in their study that a dynamic CA compared to a static interview system is perceived as more engaging and more human-like and, thereby, increases the feeling of social presence. Thus, we hypothesize that:

H1b: *A CA will produce a higher perceived social presence compared to a web survey.*

Based on the “Computers are Social Actors” (CASA) paradigm from Nass et al. (1994), anthropomorphic design elements of personification are applied to human-like technologies (Gnewuch et al. 2018; Schuetzler et al. 2018). Anthropomorphic design elements include, e.g., a humanlike appearance of agents and socially oriented communication (Fink 2012). Anthropomorphism describes the tendency to apply humanlike characteristics, motivations, intentions or emotions to nonhuman agents (Epley et al. 2007). A study by Kim et al. (2019) shows that a CA can perform part of a human interviewer’s role by applying effective communication strategies. Anthropomorphic design elements might have the potential to overcome problems associated with online course evaluations, such as low acceptance rates. According to Fink (2012), adding humanlike design cues can have a positive impact on acceptance as they can elicit social responses. We propose that a higher acceptance rate induced by adding higher anthropomorphism to an agent leads to a higher response quality:

H2a: *Higher levels of anthropomorphism will produce a higher response quality.*

Adding anthropomorphic design elements such as a friendly communication style to an agent has a positive effect on the feeling of social presence (Gnewuch et al. 2018; Rietz et al. 2019; Verhagen et al. 2014). Gnewuch et al. (2018) use response delays to simulate the time it would take humans to respond to a message. In their study they demonstrate that CAs with dynamically delayed responses are perceived more humanlike by users and have a higher social presence (Gnewuch et al. 2018). Therefore, we derive the following hypothesis:

H2b*: Higher levels of anthropomorphism will produce a higher perceived social presence.*

According to Kim et al. (2019), social presence promotes conscientious responses, which has a positive effect on response quality. Social presence shows the degree to which participants in computer-mediated communication feel affectively connected to each other (Swan and Shih 2005). Verbal and nonverbal interactions draw the respondents’ attention and increase the social engagement with an agent, which leads to more appropriate answers (Holbrook et al. 2003). Therefore, we propose that a higher perceived social presence of an agent leads to a higher response quality:

H3*: Higher levels of perceived social presence, will produce a higher response quality.*

Our research model is illustrated in Figure 1.

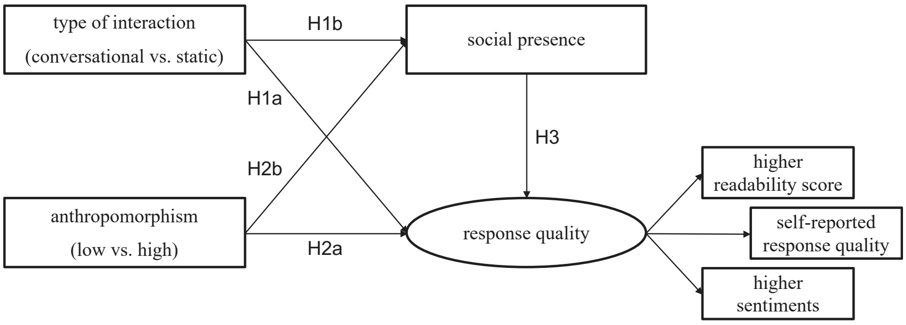


Figure 1. Overview of our research model

3 Research Methodology

To answer our research question, we chose to conduct a field experiment to test whether conversational compared to static interfaces result in higher quality response data for course evaluations. We chose this type of approach to evaluate the different interaction types in a real-world scenario and to ensure our results are independent of lecture type and content. Therefore, we designed a field experiment in which students of different lectures were asked to provide feedback on the lecture content and the teaching style of the lecturer. We used a fully randomized 2 (interaction type: conversational agent vs. static web survey) x 2 (anthropomorphism: low anthropomorphism vs. high anthropomorphism) between-subjects design resulting in one control group (CG) and three treatment groups (TG). The students were randomly assigned to one of the four conditions. The course evaluation questions were exactly the same for all groups, consisting of two quantitative and three qualitative questions following the standard content of the course evaluation of our university. In the following, we will explain the sample of participants, the artifact design, the experimental procedures and our measurements for the study.

3.1 Participants

Lecture	CG	TG1	TG2	TG3	All answers	Percentage
Small-sized	12	0	5	0	**17**	**9.6%**
Medium-sized	5	10	11	6	**32**	**18.2%**
Large-scale	36	35	32	24	**127**	**72.2%**
Participants	**53**	**45**	**48**	**30**	**176**	**100%**
Mean Age	24.71	24.43	24.56	24.57	**24.52**	
Gender	Female: 23 Male: 29 N/A: 1	Female: 22 Male: 22 N/A: 1	Female: 18 Male: 28 N/A: 2	Female: 9 Male: 17 N/A: 4	Female: 72 Male: 96 N/A: 8	Female: 41% Male: 54.5% N/A: 4.5%

Table 1. Overview of participant (CG = control group: low anthr. web survey, TG1: high anthr. web survey, TG2: low anthr. CA, TG3: high anthr. CA)

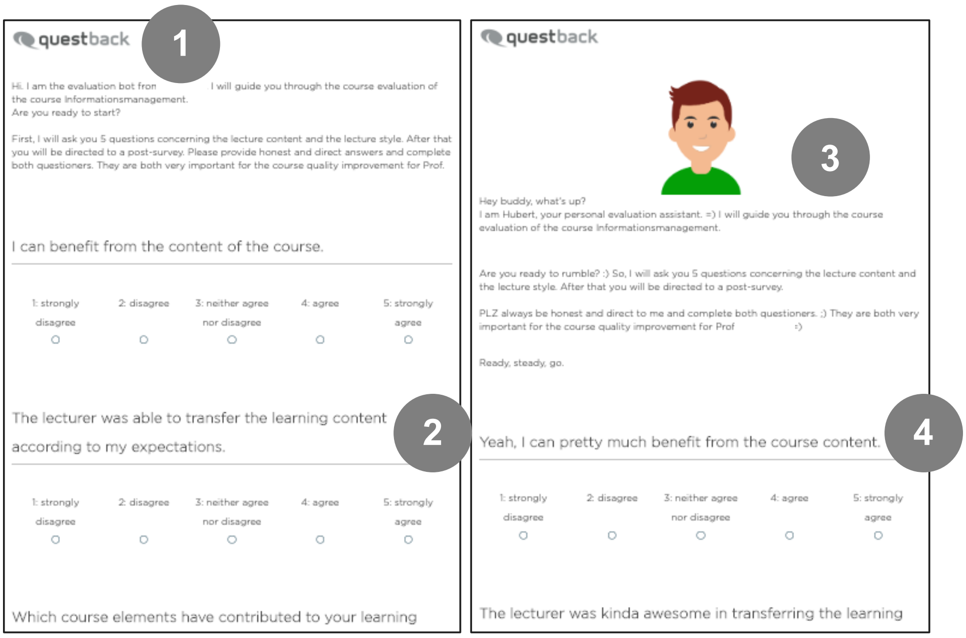
In order to validate our hypotheses in different learning environments, we conducted the experiment in three different lecture types: *small-sized lectures* (less than 25 students), *medium sized-lectures* (25 to 70 students) and *large-scale lectures* (more then 70 students). Since course evaluation feedback depend on the teacher-learner ratio, we aimed to ensure our results are independent of lecture type and content. The *small-sized lecture* was a course about *research methods*. In total, we obtained 17 valid answers from this type of lecture. The content of the *medium-sized lecture* was about *information management*. Here, we received 32 valid answers from students in this course. To capture *large-scale lectures*, we applied a course evaluation in a lecture about *digital business transformation* in which 127 valid answers were collected. All three lectures were mandatory master lectures at our university. In total, 176 students participated in our study with a mean age of 24.52 years, consisting of 72 males, 96 females and 8 students who chose not to reveal their gender. All participants were master students in economics or business in their first or third semester. Except for the *small-sized lecture*, we randomly assigned the participants to all four groups (see Table 1). In the *small-sized lecture*, we only test between low anthropomorphism of the web survey and low anthropomorphism of the CA to receive valid answers for the limited number of participants in this setting.

3.2 Design of Course Evaluation Artifacts

Every participant was asked to provide feedback on the lecture by answering five questions (see section 3.3). Every group received a different artifact leading the user through the course evaluation. We manipulated the user interaction (conversational agent vs. static web survey) in between the groups as well as the level of anthropomorphism (low anthropomorphism vs. high anthropomorphism) resulting in four artifacts. Our control group was the one with a low anthropomorphic web survey since this is the traditional format of course evaluation at our university.

Manipulation of User Interaction

For the interaction type of the course evaluation artifact, we used two different interfaces: a standard web survey and a CA. The participants of the two web survey groups (CG and TG1) conducted the course evaluation with a simple web survey tool called *unipark*. We chose this tool since it allowed us to design the survey similarly to the traditional web survey used at our university. The web survey could be completed by the students using either their personal notebook or a mobile device. The design of the web survey is presented in Figure 2. The quantitative questions were answered with a simple matrix format to ensure that the same-scaled options were used for multiple items to avoid repeating information. The qualitative items were answered with a simple plain text input field.



Hi. I am the evaluation bot fron I will guide you through the course evaluation of the course Informationsmanagement.
Are you ready to start?

First, I will ask you 5 questions concerning the lecture content and the lecture style. After that you will be directed to a post-survey. Please provide honest and direct answers and complete both questioners. They are both very important for the course quality improvement for Prof.

I can benefit from the content of the course.

1: strongly disagree	2: disagree	3: neither agree nor disagree	4: agree	5: strongly agree
○	○	○	○	○

The lecturer was able to transfer the learning content according to my expectations.

1: strongly disagree	2: disagree	3: neither agree nor disagree	4: agree	5: strongly agree
○	○	○	○	○

Which course elements have contributed to your learning

web survey - low anthropomorphism

Hey buddy, what's up?
I am Hubert, your personal evaluation assistant. =) I will guide you through the course evaluation of the course Informationsmanagement.

Are you ready to rumble? :) So, I will ask you 5 questions concerning the lecture content and the lecture style. After that you will be directed to a post-survey.

PLZ always be honest and direct to me and complete both questioners. ;) They are both very important for the course quality improvement for Prof :)

Ready, steady, go.

Yeah, I can pretty much benefit from the course content.

1: strongly disagree	2: disagree	3: neither agree nor disagree	4: agree	5: strongly agree
○	○	○	○	○

The lecturer was kinda awesome in transferring the learning

web survey - high anthropomorphism

Figure 2. Examples of manipulation of the level of anthropomorphism of the web survey (anonymized)

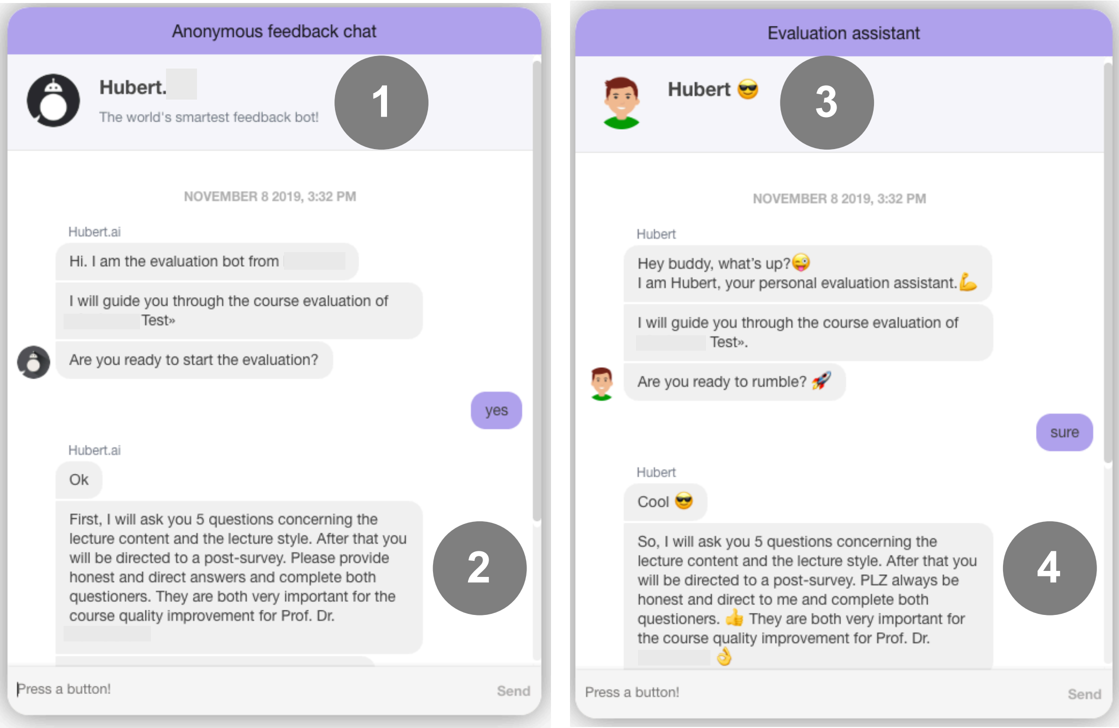
The participants of treatment group 2 and 3 were asked to use a CA to conduct the course evaluation. For the CA we cooperated with the company *Hubert.ai*[1] which is specialized on conducting chatbot-based surveys. The cooperation brought several benefits compared to developing our own solution: First, we could rely on proven design experience for questioning bots, which has been applied already in several practical scenarios including course evaluation. Second, the native designed chatbot of Hubert.ai

[1] We hereby thank *hubert.ai* for supporting our research endeavor by providing the conversational agent.

allowed us to control all design parameters, collect logs of interaction behavior and manipulate the interaction of the CA with the user. Like the web survey, the CA survey could also be conducted using a personal computer or a mobile device. The design of the CA is illustrated in Figure 3.

Manipulation of Anthropomorphism

Besides manipulating the interaction type of the course evaluation, we also differentiated between the level of anthropomorphism of the web survey and the CA based on social response theory. Therefore, we distinguished between two anthropomorphic design elements: 1) a humanlike appearance of the CA and personification elements in the web survey following Araujo (2018) and 2) socially oriented communication, meaning more casual and extensive communication behavior following Chattaraman et al. (2019) and Kim et al. (2019) (see Figure 2 and 3 for the anonymized versions). The humanlike appearance was differentiated in the web survey by creating an avatar represented by a picture and a name following Nowak and Biocca (2003), which guided the participant through the survey (high anthropomorphism) (see Figure 2: 1 vs. 3). The socially oriented communication consisted of two types: First, a more extensive and casual communication with a casual conversation style, informal question items and the use of emojis (such as those commonly used in text messaging) following Colley et al. (2004) and Stephens et al. (2009), representing high anthropomorphism. Second, a formal conversation tone with standardized form, proper grammar and punctuation, and formal question items was used for the low anthropomorphistic versions (see Figure 2 and 3: 2 vs. 4). For both, the CA and the web survey, we used exactly the same social elements, such as the same texts, images and questions, to ensure comparability of the effects. The humanlike appearance was differentiated by giving the chatbot a name, a certain character (named Hubert) and social elements such as a longer response time (high anthropomorphism) (see Figure 3: 1 and 3). A fast response would be perceived as unnatural because humans need time for reading and answering a message and, thus, are not able to respond immediately (Gnewuch et al. 2018). Therefore, we designed the high anthropomorphic version of the CA with a response time of one second and a typing indicator, whereas the low anthropomorphic one directly responded to the user without any indicator.



Hubert.[redacted]
The world's smartest feedback bot!

NOVEMBER 8 2019, 3:32 PM

Hubert.ai
Hi. I am the evaluation bot from [redacted]
I will guide you through the course evaluation of [redacted] Test»
Are you ready to start the evaluation?

yes

Hubert.ai
Ok
First, I will ask you 5 questions concerning the lecture content and the lecture style. After that you will be directed to a post-survey. Please provide honest and direct answers and complete both questioners. They are both very important for the course quality improvement for Prof. Dr. [redacted]

CA - low anthropomorphism
(0 second response time)

Hubert 😎

NOVEMBER 8 2019, 3:32 PM

Hubert
Hey buddy, what's up?😜
I am Hubert, your personal evaluation assistant.💪
I will guide you through the course evaluation of [redacted] Test».
Are you ready to rumble? 🚀

sure

Hubert
Cool 😎
So, I will ask you 5 questions concerning the lecture content and the lecture style. After that you will be directed to a post-survey. PLZ always be honest and direct to me and complete both questioners. 👍 They are both very important for the course quality improvement for Prof. Dr. [redacted] 👌

CA - high anthropomorphism
(1 second response time)

Figure 3. Examples of manipulation of the level of anthropomorphism of the CA (anonymized)

3.3 Experiment Procedure

The experiment consisted of three phases: 1) *randomization,* 2) *course evaluation* and 3) *posttest* (see Figure 4). The *randomization* and the *posttest* were consistent for all four groups. In the *course evaluation* phase, we asked all participants the same questions: two quantitative and three qualitative questions following the standard content of the course evaluation of our university. We only manipulated the human interaction of the artifacts between the groups by interaction (conversational agent vs. static web survey) and level of anthropomorphism (low anthropomorphism vs. high anthropomorphism).

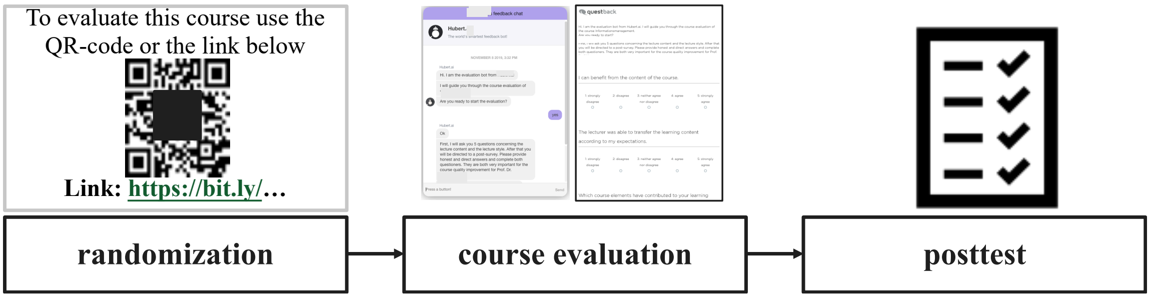


Figure 4. Overview of the experiment phases

Randomization: The field experiment started with the lecturer announcing the conduction of a midterm course evaluation of the lecture. The students were asked to either type a link into their notebook or scan a QR code with their mobile device. The link led to a web page, which fully randomly assigned the students to one of the four different groups (CG and TG1-TG3). As two students with differently assigned interfaces could be sitting next to each other, they were told that different user interfaces were being tested for improving the design of the course evaluation. All course evaluations were conducted in the middle of the lecture period (after about 50% of the content had been taught).

Course evaluation: In the *course evaluation* phase, the students were asked five different questions about the content of the course and the lecture style. Since we wanted to test the effects of our treatments with both quantitative and qualitative question styles, we asked two quantitative questions first, followed by three qualitative ones. The first two questions addressed the perceived benefit of the course *("I can benefit from the content of the course."*) and the expectation for the lecturing style *("The lecturer was able to transfer the learning content according to my expectations.")*. Both questions were measured with a 5-point Likert scale (1: strongly disagree to 5: strongly agree, with 3 being a neutral statement). Next, we asked the students three open qualitative questions ("*Which course elements have contributed to your learning success in a positive way?", "Which aspects of the course should be changed so that students benefit more from the course?"* and *"Are there any other points you would like to comment on?").*

Posttest: After the students conducted the course evaluation, they were led to a post-survey, in which we measured different constructs to validate our derived hypotheses. The constructs are described in the following section "Quantitative Data".

3.4 Measurement and Analysis

3.4.1 Quantitative Data

To measure social presence, we used items from Gefen and Straub (1997). In particular, we asked students about the following items: "*There is a sense of human contact in the course evaluation tool.", "There is a sense of personalness in the course evaluation tool.", "There is a sense of sociability in the course evaluation tool.", "There is a sense of human warmth in the course evaluation tool.",* and *"There is a sense of human sensitivity in the course evaluation tool."* For measuring the response quality*,* we used one subjective and two objective measurements: (1) self-reported response quality by the user, and

(2) syntactic readability based on the Flesch-readability score (Flesch 1943) as well as the intensity of sentiments in the answers (e.g., Joshi et al. 2014; Pang and Lee 2008). We measured the self-reported response quality by asking participants the following questions: *"The design of the evaluation tool made me think longer about my responses compared to traditional surveys."* and *"I would prefer using a chatbot as a survey tool."* Moreover, we captured the perceived usefulness, intention to use and ease of use following the technology acceptance model of Venkatesh et al. (2003), Venkatesh and Bala (2008). Exemplary items for the three constructs are: "*Imagine the evaluation tool was available in your next course, I am encouraged to use it.*", *"I would find the evaluation tool useful for giving course feedback.", "Using the evaluation tool would allow me to perform a more effective evaluation." or "The evaluation tool is easy to use."* All these items were measured with a 5-point Likert scale (1: strongly disagree to 5: strongly agree, with 3 being a neutral statement).

3.4.2 Qualitative Data

To measure the syntactic readability of texts, several measures have been used in research (Fromm et al. 2019; Khawaja et al. 2010). We selected the Flesh-Reading-Ease (FRE) (Flesch 1943) to capture the readability of received responses since this score combines language complexity measurements such as the average sentence lengths and the average syllables per word into one number (Flesch 1943). The score has been widely used before to determine the readability of a message in computer-mediated communication (Walther 2007; Wambsganss and Fromm 2019) or for the complexity of CA user responses (Gnewuch et al. 2018). Following Flesch (1943), we used the following formula since we received answers in English:

$$\boxed{\textit{Flesch Reading Ease} = 206.835 - (1.015 * \textit{asl}) - (84.6 * \textit{asw})}$$

asl: average sentence length of a response, asw: average syllable per word

The scores of our answers reach from *0* to *110*. The higher the FRE score, the better the readability of the responses. Moreover, we aimed to capture the sentiments of our received responses since a sentiment is a good indicator for an individual taking a position on a certain topic used, e.g., in opinion mining (Gruettner et al. 2020; Pang and Lee 2008). For example, if a student only answers "course content in learning unit 2" no action steps can be derived, since this message has no sentiment (positive or negative notion). Therefore, we used the Naïve-Bayes approach of TextBlob[2], using Python 3.7 to determine the sentiments of each response since it is an easy to use, openly available approach trained on review (evaluation) data. The scores are usually labeled between -1 and 1 according to a "positive", a "negative" and a "neutral" mood (-1 being negative, 0 neutral and 1 positive). However, we multiplied values smaller than 0 with **-1* since we did not distinguish between positive or negative sentiments. We believe "position talking" sentiments are valuable for the use case of course evaluation, similar to opinion mining (Pang and Lee 2008) or language complexity measurements (Joshi et al. 2014). Finally, we used a continuous normalized scale from 0 to 1 to measure the sentiments: 0 meaning no sentiment (neutral statement) and 1 meaning high sentiment (no matter if positive or negative). For measuring the FRE and the sentiments, the answers of all three qualitative questions from the course evaluation were combined to one string and analyzed using Python 3.6, utilizing the natural language toolkit (NLTK) (Bird et al. 2009). To construct one measurement for data quality, we normalized the construct's self-reported response quality, FRE and sentiments and weighted every measurement with one third to generate one final value to distinguish the responses.

In addition, we collected demographic information (age and gender) and asked participants if they had used a CA (e.g., Facebook Messenger Bot) before to control for technology usage between the groups. For data analysis, we used linear regression models and checked their assumptions visually with a test for normality and a test for homoscedasticity. All assumptions are met (see appendix).

[2] https://textblob.readthedocs.io/en/dev/index.html

4 Results

4.1 Descriptive results

To answer our research question, how a CA can improve the response quality of online course evaluations compared to a web survey, we calculated the means and standard derivations (SD) between the control and treatment groups (CG and TG1-TG3) depicted in Table 2. Moreover, we conducted multiple regressions for the corresponding variable as a control variable for all the hypotheses 1-3. The descriptive statistics are also illustrated in Table 2.

Condition	**n**	**Social presence**		**Response quality**		**Social presence**		**Response quality**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
CG: low anthr. web survey	53	2.307	0.864	0.415	0.112	2.372	0.927	0.431	0.116
TG1: high anthr. web survey	45*	2.431	0.988	0.449	0.1181				
TG2: low anthr. CA	48	3.070	0,980	0.506	0.108	3.211	0.930	0.507	0.108
TG3: high anthr. CA	30*	3.413	0.829	0.508	0.111				
Hypotheses						H1b: confirmed H2b: not confirmed		H1a: confirmed H2a: not confirmed	

*Table 2. Overview means and standard derivations (*not part of small-sized lecture experiment as explained above)*

To summarize our descriptive findings, we plotted the results on perceived social presence and response quality between the two interaction types web survey and CA in Figure 5.

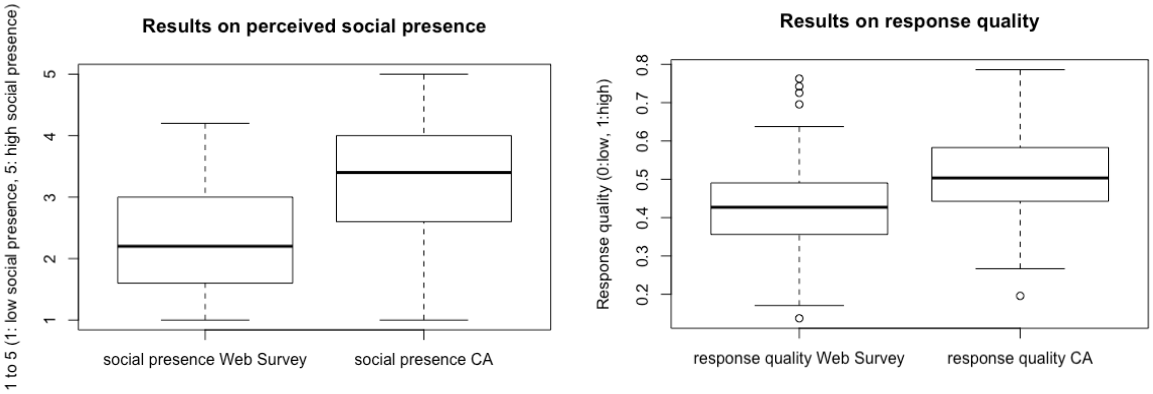


Figure 5. Differences in perceived social presence and response quality between Web Survey and CA

4.2 Analytical results

The results of our research model, the *r* values and the significances are illustrated in Figure 6. The statistical tests on our results support our hypotheses H1b and H1a, meaning that a conversational interaction type significantly influences the perceived social presence and response quality in online course evaluations. However, our hypothesis H2a and H2b were not supported, meaning that the level of anthropomorphism of a CA or a web survey has no significant influence on social presence and response quality in online course evaluations. Also, we could not confirm our hypothesis H3, meaning that social presence in online course evaluations does not automatically lead to a higher response quality.

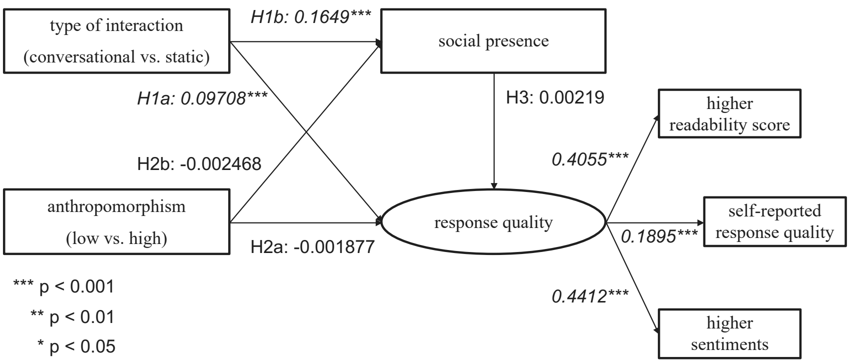


Figure 6. Overview of R-squared and significances of our hypotheses

In order to control for potential effects of interfering variables with our sample size and to ensure that the randomization was successful, we compared the difference in the mean of CA pre-usage, age and gender. We received p-values larger than 0.05 showing that there was no significant difference between the groups for all three variables. Around 70 percent of the participants had used a CA before across all treatments. Besides, we did not find any significances between the results of the qualitative course evaluations between the treatment groups. The course content was rated with mean = 3.60 (SD = 0.23) and the lecturer with mean = 3.72 (SD = 0.91) of participants using the CA. Participants conducting the course evaluation through the web survey judged the course content with mean = 3.75 (SD = 0.75) and the lecturer with mean = 3.66 (SD = 0.84).

5 Discussion

We conduct a 2 x 2 field experiment design to test if and how different interfaces (CA vs. static web survey) and different levels of anthropomorphism (low vs. high anthropomorphism) result in a higher response quality for online course evaluations. We chose to conduct a field experiment to evaluate the different interaction types in different real-world scenarios to ensure our results are independent of lecture type and content. Drawing on social response theory (Moon 2000; Nass et al. 1994; Nass and Moon 2000), we designed an NLP-based CA and deployed a field experiment with 176 students in three different course formats in which participants were asked to provide feedback on the lecture content and the teaching style of the lecturer. We found that participants using a CA showed higher levels of social presence and were more likely to share high quality feedback. Moreover, we measured the technology acceptance showing that the intention to use a CA for online course evaluations is higher (mean = 3.69, SD = 0.93) than the intention to use a web survey (mean = 3.47, SD = 0.83). These results are consistent with past studies that investigated beneficial effects of CAs over non-adaptive systems (such as surveys) (e.g., S. Kim et al., 2019; Wambsganss et al., 2020). One reason for the effect might be that conversational interfaces better direct the attention of the user to the question compared to a static web survey (Kim et al., 2019). We argue, that this might help to overcome the common challenges of surveys in general, such as survey fatigue (Heerwegh and Loosveldt 2008) or satisficing behavior (Tucker et al. 2008), and thus leads to better response quality. This can also be seen by the qualitative data of our post-survey, where multiple students made comments about CA such as about the perceived interaction: "*More natural and seems transparent*", "*It was more interactive than a questionnaire*", "*Feels more personal*"; the perceived usefulness: "*it is very easy to understand and to use*"; the perceived joy: "*It's different than other tools and more fun";* the response behaviour: "*I think your feedback with this tool is more honest*"; or in general: "*It's been time to implement these technologies".* Interestingly, our study was not able to show significant effects of anthropomorphism on the response quality. This is in contrast to the many studies that show the advantages of anthropomorphism in CAs (Fink 2012; Gnewuch et al. 2018; Rietz et al. 2019). However, our study verifies the results of Lind et al. (2013), who found that participants expose more sensitive information in the absence of a humanized interface as they perceive greater anonymity. We argue, that course evaluations are about revealing feedback, which can be, to an extent, sensitive information. Our study has several theoretical contributions and practical implications. First, we contribute to the application of CAs in education, suggesting a successful use case to employ a CA with potential benefits for lecturers and educational institutions to better adjust their learning content based on high quality responses and potentially continuous student feedback. Second, we contribute to the HCI and social response theory by investigating the influence of the interaction type and the level of anthropomorphism on an objective outcome (e.g., number of sentiments and readability). Our study suggests that high anthropomorphistic design elements in the field of user-information seeking scenarios such as course evaluation is not necessarily a benefit for response quality. The study also makes some practical contributions by exemplarily showing how educators can implement CAs as course evaluation tools in their learning environments, since they are widely available and easy to use from an educator's – non-programming expert's – point of view. Besides, our study faces some limitations. First, we only asked a representative subset of course evaluation question. Asking more questions with a longer interaction might lead to different response results. Second, it remains open if an ongoing usage of a CA as a course evaluation tool continuously leads to a higher response quality compared to a web survey or if this was only a short time effect. Even if 70 percent of the participants said they had used a CA before, novelty effects cannot be expelled. Therefore, we call for future work to test the effect of a CA as the course evaluation tool in a longitude study. Moreover, we did not investigate the different social cues about the CA design, since our objectives were rather to provide an empirical proof that a CA leads to a higher response quality. Also, we did not investigate potential effects of subsets (e.g., gender or lecture format) on the measured variables. Hence, we also call for future work to investigate the design of social cues for information-seeking bots such as for course evaluations or other types of feedback scenarios and the influence of lecture types on response quality.

6 Conclusion

We aimed to explore the effects of CAs on the response quality of online course evaluations in education compared to the common standard of web surveys. We proposed that a conversational interface will have a positive effect on the response quality through the change of interaction and perceived social presence. We designed an NLP-based CA and deployed it in a field experiment with 176 students in three different course formats and compared it with a web survey. The results indicate that participants using the CA showed higher levels of social presence and were more likely to share high quality feedback. These findings and the measured technology acceptance suggest using CAs not only as a standard tool for course evaluations but also for other forms of evaluation feedback or survey tools.

Appendix

Test for homogeneity of variance

Test for normality

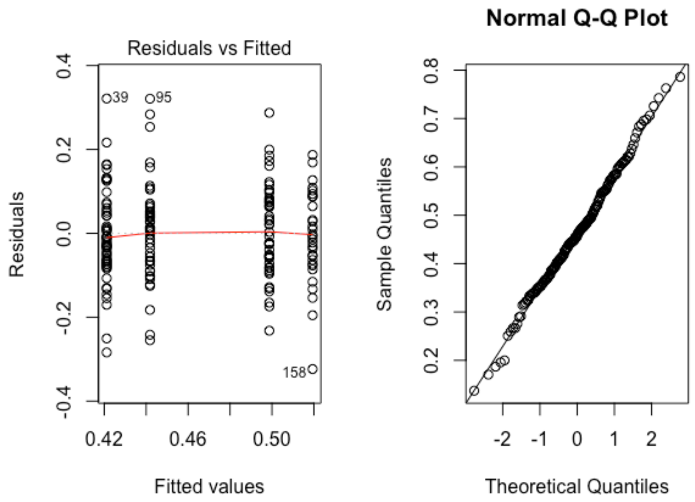


Figure 7. Check for assumptions of linear regression

References

Aguiar, E. V. B., Tarouco, L. M. R., and Reategui, E. 2014. "Supporting Problem-Solving in Mathematics with a Conversational Agent Capable of Representing Gifted Students' Knowledge," *Proceedings of the Annual Hawaii International Conference on System Sciences*, IEEE, pp. 130–137. (https://doi.org/10.1109/HICSS.2014.24).

Araujo, T. 2018. "Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions," *Computers in Human Behavior* (85), Elsevier Ltd, pp. 183–189. (https://doi.org/10.1016/j.chb.2018.03.051).

Atkinson, R. C., and Shiffrin, R. M. 1968. "Human Memory: A Proposed System and Its Control Processes," *Psychology of Learning and Motivation - Advances in Research and Theory* (2:C), pp. 89–195. (https://doi.org/10.1016/S0079-7421(08)60422-3).

Bird, S., Klein, E., and Loper, E. 2009. "Natural Language Processing with Python," *Text* (Vol. 43). (https://doi.org/10.1097/00004770-200204000-00018).

Blair, E., and Valdez Noel, K. 2014. "Improving Higher Education Practice through Student Evaluation Systems: Is the Student Voice Being Heard?," *Assessment and Evaluation in Higher Education* (39:7), Routledge, pp. 879–894. (https://doi.org/10.1080/02602938.2013.875984).

vom Brocke, J., Maaß, W., Buxmann, P., Maedche, A., Leimeister, J. M., and Pecht, G. 2018. "Future Work and Enterprise Systems," *Business and Information Systems Engineering* (60:4), pp. 357–366. (https://doi.org/10.1007/s12599-018-0544-2).

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., and McTear, M. 2017. "Towards a Chatbot for Digital Counselling," *HCI 2017: Digital Make Believe - Proceedings of the 31st International BCS Human Computer Interaction Conference, HCI 2017* (2017-July), pp. 1–7. (https://doi.org/10.14236/ewic/HCI2017.24).

Chattaraman, V., Kwon, W. S., Gilbert, J. E., and Ross, K. 2019. "Should AI-Based, Conversational Digital Assistants Employ Social- or Task-Oriented Interaction Style? A Task-Competency and Reciprocity Perspective for Older Adults," *Computers in Human Behavior* (90), Elsevier Ltd, pp. 315–330. (https://doi.org/10.1016/j.chb.2018.08.048).

Colley, A., Todd, Z., Bland, M., Holmes, M., Khanom, N., and Pike, H. 2004. "Style and Content in E-Mails and Letters to Male and Female Friends," *Journal of Language and Social Psychology* (23:3), pp. 369–378. (https://doi.org/10.1177/0261927X04266812).

eMarketer. 2017. "Alexa , Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year," *EMarketer*. (https://www.emarketer.com/Articles/Print.aspx?R=1015812).

Epley, N., Waytz, A., and Cacioppo, J. T. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism," *Psychological Review* (114:4), pp. 864–886. (https://doi.org/10.1037/0033-295X.114.4.864).

Erikson, M., Erikson, M. G., and Punzi, E. 2016. "Student Responses to a Reflexive Course Evaluation," *Reflective Practice* (17:6), Routledge, pp. 663–675. (https://doi.org/10.1080/14623943.2016.1206877).

Fadel, C., and Groff, J. S. 2018. "Four-Dimensional Education for Sustainable Societies," in *Sustainability, Human Well-Being, and the Future of Education*, Palgrave Macmillan, pp. 269–281. (https://doi.org/10.1007/978-3-319-78580-6_8).

Fink, J. 2012. "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (7621 LNAI), pp. 199–208. (https://doi.org/10.1007/978-3-642-34103-8_20).

Fitzpatrick, K. K., Darcy, A., and 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), p. e19. (https://doi.org/10.2196/mental.7785).

Flesch, R. 1943. "Marks of Readable Style; a Study in Adult Education.," *Teachers College Contributions to Education* (897).

Fromm, H., Wambsganss, T., and Söllner, M. 2019. "Towards a Taxonomy of Text Mining Features," in *European Conference of Information Systems (ECIS)*, pp. 1–12.

Gefen, D., and Straub, D. W. 1997. "Gender Differences in the Perception and Use of E-Mail: An Extension to the Technology Acceptance Model," *MIS Quarterly: Management Information Systems* (21:4), pp. 389–400. (https://doi.org/10.2307/249720).

Gnewuch, U., Morana, S., Adam, M. T. P., and Maedche, A. 2018. "Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction," in *26th European Conference on Information Systems (ECIS) 2018*. (https://aisel.aisnet.org/ecis2018_rp/113).

Gruettner, A., Vitisvorakarn, M., Wambsganss, T., Rietsche, R., and Back, A. 2020. "The New Window to Athletes' Soul-What Social Media Tells Us About Athletes' Performances," *Hawaii International Conference on System Sciences (HICSS).*

Heerwegh, D., and Loosveldt, G. 2008. "Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality," *Public Opinion Quarterly* (72:5), pp. 836–846. (https://doi.org/10.1093/poq/nfn045).

Hobert, S., and Wolff, R. M. Von. 2019. "Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents," *14th International Conference on Wirtschaftsinformatik, Siegen, Germany*.

Holbrook, A. L., Green, M. C., and Krosnick, J. A. 2003. "Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires," *Public Opinion Quarterly* (67:1), pp. 79–125. (https://doi.org/10.1086/346010).

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., and Akkiraju, R. 2018. "Touch Your Heart: A Tone-Aware Chatbot for Customer Care on Social Media," *Conference on Human Factors in Computing Systems - Proceedings* (2018-April). (https://doi.org/10.1145/3173574.3173989).

Joshi, A., Mishra, A., Senthamilselvan, N., and Bhattacharyya, P. 2014. "Measuring Sentiment Annotation Complexity of Text," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference* (Vol. 2), Association for Computational Linguistics, pp. 36–41. (https://doi.org/10.3115/v1/p14-2007).

Kerly, A., Hall, P., and Bull, S. 2007. "Bringing Chatbots into Education: Towards Natural Language Negotiation of Open Learner Models," *Knowledge-Based Systems* (20:2), pp. 177–185. (https://doi.org/10.1016/j.knosys.2006.11.014).

Khawaja, M. A., Chen, F., and Marcus, N. 2010. "Using Language Complexity to Measure Cognitive Load for Adaptive Interaction Design," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 333–336. (https://doi.org/10.1145/1719970.1720024).

Kim, C. M., and Baylor, A. L. 2008. "A Virtual Change Agent: Motivating Pre-Service Teachers to Integrate Technology in Their Future Classrooms," *Educational Technology and Society* (11:2), pp. 309–321.

Kim, S., Lee, J., and Gweon, G. 2019. "Comparing Data from Chatbot and Web Surveys Effects of Platform and Conversational Style on Survey Response Quality," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12. (https://doi.org/10.1145/3290605.3300316).

Kowatsch, T., Nißen, M., Shih, C. I., Rüegger, D., Filler, A., Künzler, F., Barata, F., Haug, S., Brogle, B., Heldt, K., Gindrat, P., Farpour-lambert, N., and Allemand, D. 2017. "Text-Based Healthcare Chatbots Supporting Patient and Health Professional Teams : Preliminary Results of a Randomized Controlled Trial on Childhood Obesity," *Persuasive Embodied Agents for Behavior Change (PEACH2017) Workshop* (1:Iva 2017), pp. 1–10.

Krassmann, A. L., Paz, F. J., Silveira, C., Tarouco, L. M. R., and Bercht, M. 2018. "Conversational Agents in Distance Education: Comparing Mood States with Students' Perception," *Creative Education* (09:11), pp. 1726–1742. (https://doi.org/10.4236/ce.2018.911126).

Krosnick, J. A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology* (5:3), pp. 213–236. (https://doi.org/10.1002/acp.2350050305).

Krosnick, J. A. 1999. "SURVEY RESEARCH," *Annual Review of Psychology* (50:1), pp. 537–567. (https://doi.org/10.1146/annurev.psych.50.1.537).

Kulik, J. A., and Fletcher, J. D. 2016. “Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review,” *Review of Educational Research* (86:1), pp. 42–78. (https://doi.org/10.3102/0034654315581420).

Latorre-Navarro, E., and Harris, J. 2015. “An Intelligent Natural Language Conversational System for Academic Advising,” *International Journal of Advanced Computer Science and Applications* (6:1). (https://doi.org/10.14569/IJACSA.2015.060116).

Laumer, S., Maier, C., and Gubler, F. T. 2019. “Chatbot Acceptance in Healthcare: Explaining User Adoption of Conversational Agents for Disease Diagnosis,” *Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden*, pp. 0–18. (https://aisel.aisnet.org/ecis2019_rp/88).

Lind, L. H., Schober, M. F., Conrad, F. G., and Reichert, H. 2013. “Why Do Survey Respondents Disclose More When Computers Ask the Questions?,” *Public Opinion Quarterly* (77:4), pp. 888–935. (https://doi.org/10.1093/poq/nft038).

Lucas, G. M., Gratch, J., King, A., and Morency, L. P. 2014. “It’s Only a Computer: Virtual Humans Increase Willingness to Disclose,” *Computers in Human Behavior* (37), Elsevier Ltd, pp. 94–100. (https://doi.org/10.1016/j.chb.2014.04.043).

Lundqvist, K. O., Pursey, G., and Williams, S. 2013. *Design and Implementation of Conversational Agents for Harvesting Feedback in ELearning Systems*, Springer, Berlin, Heidelberg, pp. 617–618. (https://doi.org/10.1007/978-3-642-40814-4_79).

Moon, Y. 2000. “Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers,” *Journal of Consumer Research* (26:4), Narnia, pp. 323–339. (https://doi.org/10.1086/209566).

Nass, C., and Moon, Y. 2000. “Machines and Mindlessness: Social Responses to Computers,” *Journal of Social Issues* (56:1), John Wiley & Sons, Ltd (10.1111), pp. 81–103. (https://doi.org/10.1111/0022-4537.00153).

Nass, C., Steuer, J., and Tauber, E. R. 1994. “Computers Are Social Actors,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI ’94*, New York, New York, USA: ACM Press, pp. 72–78. (https://doi.org/10.1145/191666.191703).

Nowak, K. L., and Biocca, F. 2003. “The Effect of the Agency and Anthropomorphism on Users’ Sense of Telepresence, Copresence, and Social Presence in Virtual Environments,” in *Presence: Teleoperators and Virtual Environments* (Vol. 12), , October, pp. 481–494. (https://doi.org/10.1162/105474603322761289).

Pang, B., and Lee, L. 2008. “Opinion Mining and Sentiment Analysis,” *Foundations and Trends in Information Retrieval* (2:12), pp. 1–135. (http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf).

Payr, S. 2003. “The Virtual University’s Faculty: An Overview of Educational Agents,” *Applied Artificial Intelligence* (17:1), Taylor & Francis Group , pp. 1–19. (https://doi.org/10.1080/713827053).

von der Pütten, A. M., Krämer, N. C., Gratch, J., and Kang, S.-H. 2010. “‘It Doesn’t Matter What You Are!’ Explaining Social Effects of Agents and Avatars,” *Computers in Human Behavior* (26:6), Elsevier Science Publishers B. V., pp. 1641–1650. (https://doi.org/10.1016/j.chb.2010.06.012).

Rietz, T., Benke, I., and Maedche, A. 2019. “The Impact of Anthropomorphic and Functional Chatbot Design Features in Enterprise Collaboration Systems on User Acceptance,” in *14th International Conference on Wirtschaftsinformatik*, pp. 1642–1656. (https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1306&context=wi2019).

Rubin, V. L., Chen, Y., and Thorimbert, L. M. 2010. “Artificially Intelligent Conversational Agents in Libraries,” *Library Hi Tech* (28:4), pp. 496–522. (https://doi.org/10.1108/07378831011096196).

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., and Nunamaker, J. F. 2018. “The Influence of Conversational Agent Embodiment and Conversational Relevance on Socially Desirable Responding,” *Decision Support Systems* (114:May), Elsevier, pp. 94–102. (https://doi.org/10.1016/j.dss.2018.08.011).

Schuetzler, R. M., Grimes, G. M., Giboney, J. S., and Buckman, J. 2014. “Facilitating Natural Conversational Agent Interactions: Lessons from a Deception Experiment,” *35th International Conference on Information Systems “Building a Better World Through Information Systems”,* *ICIS 2014*.

Shawar, B. A., and Atwell, E. S. 2005. “Using Corpora in Machine-Learning Chatbot Systems,” *International Journal of Corpus Linguistics* (10:4), pp. 489–516. (https://doi.org/10.1075/ijcl.10.4.06sha).

Smithson, J., Birks, M., Harrison, G., Sid Nair, C., and Hitchins, M. 2015. “Benchmarking for the Effective Use of Student Evaluation Data,” *Quality Assurance in Education* (23:1), pp. 20–29. (https://doi.org/10.1108/QAE-12-2013-0049).

Song, D., Oh, E. Y., and Rice, M. 2017. “Interacting with a Conversational Agent System for Educational Purposes in Online Courses,” *Proceedings - 2017 10th International Conference on Human System Interactions, HSI 2017*, pp. 78–82. (https://doi.org/10.1109/HSI.2017.8005002).

Spooren, P., Brockx, B., and Mortelmans, D. 2013. “On the Validity of Student Evaluation of Teaching,” *Review of Educational Research* (Vol. 83). (https://doi.org/10.3102/0034654313496870).

Stephens, K. K., Houser, M. L., and Cowan, R. L. 2009. “R U Able to Meat Me: The Impact of Students’ Overly Casual Email Messages to Instructors,” *Communication Education* (58:3), pp. 303–326. (https://doi.org/10.1080/03634520802582598).

Steyn, C., Davies, C., and Sambo, A. 2019. “Eliciting Student Feedback for Course Development: The Application of a Qualitative Course Evaluation Tool among Business Research Students,” *Assessment and Evaluation in Higher Education* (44:1), Routledge, pp. 11–24. (https://doi.org/10.1080/02602938.2018.1466266).

Suppes, P., and Morningstar, M. 1969. “Computer-Assisted Instruction,” *Science* (166:3903), pp. 343–350. (https://doi.org/10.1126/science.166.3903.343).

Swan, K., and Shih, L. F. 2005. “ON THE NATURE AND DEVELOPMENT OF SOCIAL PRESENCE IN ONLINE COURSE DISCUSSIONS,” *Online Learning* (9:3), The Online Learning Consortium. (https://doi.org/10.24059/olj.v9i3.1788).

Tegos, S., Demetriadis, S., and Tsiatsos, T. 2014. “A Configurable Conversational Agent to Trigger Students’ Productive Dialogue: A Pilot Study in the CALL Domain,” *International Journal of Artificial Intelligence in Education* (24:1), pp. 62–91. (https://doi.org/10.1007/s40593-013-0007-3).

Tucker, B., Jones, S., and Straker, L. 2008. “Online Student Evaluation Improves Course Experience Questionnaire Results in a Physiotherapy Program,” *Higher Education Research and Development* (27:3), pp. 281–296. (https://doi.org/10.1080/07294360802259067).

Tung, F. W., and Deng, Y. S. 2006. “Designing Social Presence in E-Learning Environments: Testing the Effect of Interactivity on Children,” *Interactive Learning Environments* (14:3), pp. 251–264. (https://doi.org/10.1080/10494820600924750).

Venkatesh, V., and Bala, H. 2008. “Technology Acceptance Model 3 and a Research Agenda on Interventions,” *Decision Sciences* (39:2), John Wiley & Sons, Ltd (10.1111), pp. 273–315. (https://doi.org/10.1111/j.1540-5915.2008.00192.x).

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. “User Acceptance of Information Technology: Toward a Unified View,” *MIS Quarterly* (27:3), pp. 425–478.

Verhagen, T., van Nes, J., Feldberg, F., and van Dolen, W. 2014. “Virtual Customer Service Agents: Using Social Presence and Personalization to Shape Online Service Encounters,” *Journal of Computer-Mediated Communication* (19:3), pp. 529–545. (https://doi.org/10.1111/jcc4.12066).

Walther, J. B. 2007. “Selective Self-Presentation in Computer-Mediated Communication: Hyperpersonal Dimensions of Technology, Language, and Cognition,” *Computers in Human Behavior* (23), pp. 2538–2557. (https://doi.org/10.1016/j.chb.2006.05.002).

Wambsganss, T., and Fromm, H. 2019. “Mining User-Generated Repair Instructions from Automotive Web Communities,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences* (Vol. 6), Hawaii, pp. 1184–1193. (https://doi.org/10.24251/hicss.2019.144).

Wambsganss, T., Winkler, R., Schmid, P., and Söllner, M. 2020. “Designing a Conversational Agent as a Formative Course Evaluation Tool,” in *15th International Conference on Wirtschaftsinformatik*, Potsdam, Germany.

Wambsganss, T., Winkler, R., Söllner, M., and Leimeister, J. M. 2020. “A Conversational Agent to Improve Response Quality in Course Evaluations,” *ACM CHI Conference on Human Factors in*

Computing Systems, pp. 1–9.

Winkler, R., and Söllner, M. 2018. "Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis . In : Academy of Management," *Meeting, Annual Chicago, A O M*. (https://www.alexandria.unisg.ch/254848/1/JML_699.pdf).

Winkler, R., Söllner, M., Neuweiler, M. L., Rossini, F. C., and Leimeister, J. M. 2019. *Alexa, Can You Help Us Solve This Problem? How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes*, SIGCHI. (https://www.alexandria.unisg.ch/256978/).

Xu, A., Liu, Z., Guo, Y., Sinha, V., and Akkiraju, R. 2017. "A New Chatbot for Customer Service on Social Media," *Conference on Human Factors in Computing Systems - Proceedings* (2017-May), pp. 3506–3510. (https://doi.org/10.1145/3025453.3025496).