Please quote as: Wambsganss, T.; Zierau, N.; Söllner, M.; Käser, T.; Koedinger, K. R.; Leimeister, J. M. (2022). Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations. Proceedings of the ACM on Human-Computer Interaction (PACM), 6 (CSCW2).

Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations

THIEMO WAMBSGANSS, EPFL, Switzerland
NAIM ZIERAU, University of St.Gallen, Switzerland
MATTHIAS SÖLLNER, University of Kassel, Germany
TANJA KÄSER, EPFL, Switzerland
KENNETH R. KOEDINGER, Carnegie Mellon University, United States
JAN MARCO LEIMEISTER, University of St.Gallen, Switzerland and University of Kassel, Germany

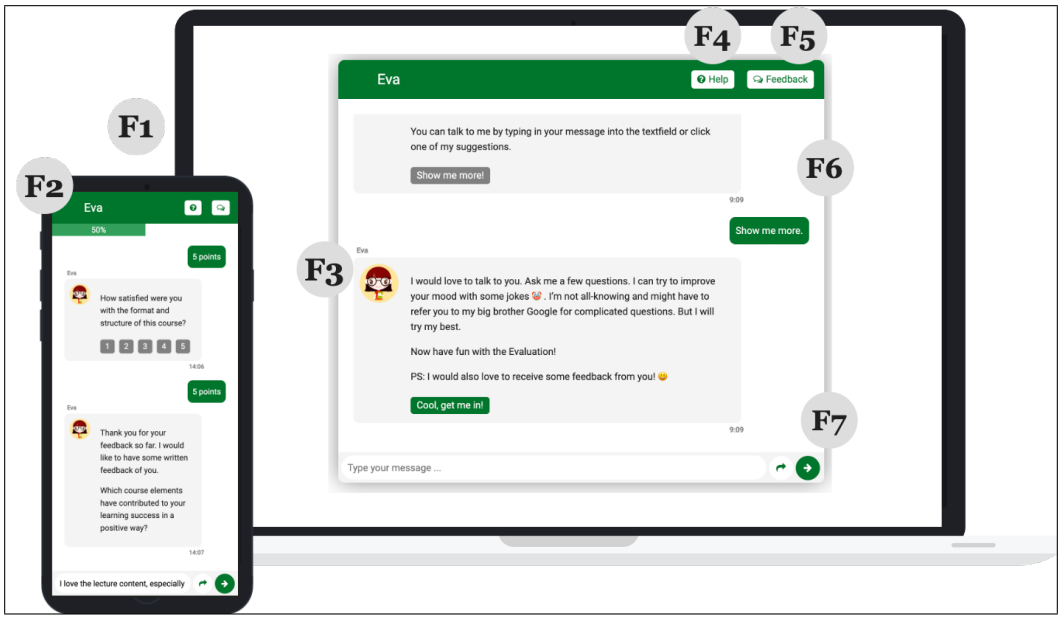

Fig. 1. Screenshot of the student-centered design of our conversational course evaluation tool: a student conducts a course evaluation in a natural conversation according to our design studies.

Authors' addresses: Thiemo Wambsganss, thiemo.wambsganss@epfl.ch, EPFL, Lausanne, Switzerland; Naim Zierau, naim.zierau@unisg.ch, University of St.Gallen, St.Gallen, Switzerland; Matthias Söllner, soellner@uni-kassel.de, University of Kassel, Kassel, Germany; Tanja Käser, tanja.kaeser@epfl.ch, EPFL, Lausanne, Switzerland; Kenneth R. Koedinger, kk1u@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, United States; Jan Marco Leimeister, janmarco.leimeister@unisg.ch, University of St.Gallen, St.Gallen, Switzerland, University of Kassel, Kassel, Germany.

2573-0142/2022/11-ART506 $15.00
https://doi.org/10.1145/3555619

Conversational agents (CAs) provide opportunities for improving the interaction in evaluation surveys. To investigate if and how a user-centered conversational evaluation tool impacts users' response quality and their experience, we build EVA - a novel conversational course evaluation tool for educational scenarios. In a field experiment with 128 students, we compared EVA against a static web survey. Our results confirm prior findings from literature about the positive effect of conversational evaluation tools in the domain of education. Second, we then investigate the differences between a voice-based and text-based conversational human-computer interaction of EVA in the same experimental set-up. Against our prior expectation, the students of the voice-based interaction answered with higher information quality but with lower quantity of information compared to the text-based modality. Our findings indicate that using a conversational CA (voice and text-based) results in a higher response quality and user experience compared to a static web survey interface.

CCS Concepts: • **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Laboratory experiments.*

Additional Key Words and Phrases: educational applications, course evaluations, conversational agents, voice interfaces

ACM Reference Format:
Thiemo Wambsganss, Naim Zierau, Matthias Söllner, Tanja Käser, Kenneth R. Koedinger, and Jan Marco Leimeister. 2022. Designing Conversational Evaluation Tools: A Comparison of Text and Voice Modalities to Improve Response Quality in Course Evaluations. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 506 (November 2022), 27 pages. https://doi.org/10.1145/3555619

1 INTRODUCTION

Student evaluations are an important tool for universities to continuously improve their teaching quality and to involve students in the further development of their curricula [2, 23]. However, many universities struggle to effectively capture students' needs and expectations due to financial and organizational constraints [4, 65, 67]. Large classroom sizes at high schools, mass lectures at universities with more than 100 students per lecturer, and the recent rise in massive open online courses (MOOCs) with more than 1000 participants provide few opportunities for individual and personal exchanges between students and educators [62]. Today, most educational institutions use online web surveys to assess teaching quality and collect student feedback [4]. These surveys have become the standard for course evaluations mainly because they improve the efficiency of the student evaluation management process [83]. For educational institutions, they offer a simple and scalable way to distribute and collect course evaluations. Moreover, they allow for (semi-)automated digital evaluations of quantitative assessment parts and partly also qualitative question parts [67]. For students, online web surveys represent a comfortable way to give feedback, however (e.g., various devices), wherever (e.g., at home), and whenever (e.g., 24x7).

Nevertheless, recent research indicates that the widespread adoption of online web surveys has created a number of problems that fundamentally question their effectiveness in course evaluation scenarios (e.g., [16, 73, 89]). First, the growing digitization of student-educator interactions, including the collection of student feedback, has resulted in reduced student satisfaction rates and an experienced loss of "human touch". As such, students feel that their feedback is not "heard" by educators [16, 73]. Second, most web surveys use static, old-fashioned designs that fall short of the expectations of often young students that are used to navigating the internet via interactive social media applications. Compared to these applications, web surveys fail to create an engaging user experience [25, 33, 89]. Third, web surveys provide few options to create adaptive and personalized forms of human-computer interaction that could foster increased participant engagement [33, 73, 89]. Finally, while online web surveys are efficient in gathering quantitative feedback, they fall short in motivating students to provide detailed accounts of their experiences when providing

qualitative feedback. However, long text answers are seen as increasingly important in higher education as they provide deeper insights into students' needs and wishes [68]. Together, these issues inform growing survey fatigue and satisfying behavior among students, ultimately leading to low acceptance and response rates as well as to low-quality responses in course evaluations [16, 52, 73, 89].

One solution to providing a human-like, adaptive, and engaging user experience and at the same time leverage efficiency-related advantages of digital technologies could be the development and deployment of conversational agents (CA) for conducting course evaluations. CAs offer a number of distinct features that might make them more effective when collecting student feedback than traditional web surveys. According to prior literature, the conversational interaction logic and the use of anthropomorphic features (e.g., the use of avatars) increase perceptions of social presence, in turn enhancing self-disclosure [33, 83, 89]. Moreover, dialogue-based interactions have shown to produce higher levels of enjoyment in survey contexts [17, 89]. Specifically, CAs are able to adapt their answers to students' utterances and can therefore build up a meaningful dialogue with the students, almost like a qualitative educator–student interview. Indeed, recent studies, e.g., by [33, 89], indicate that CAs are able to perform parts of a human interviewer's role by applying effective communication strategies and therefore encourage user enjoyment, which in turn leads to high-quality response data. We aim to build on these findings in two major ways. First, we investigate the potentially positive effects of conversational agents on response quality in the context of course evaluations using a rigorously designed field experiment. Second, we examine the effect of conversation modality (text vs. voice) on response quality and user experience. The recent proliferation of voice-based interfaces may provide new opportunities to further humanize human–computer interaction (e.g., [34, 43, 61, 70]). Consequently, comparing text and voice bots with regard to their effect on user experience and response quality in educational settings will likely lead to novel and relevant insights regarding the design of CAs for course evaluation.

In order to investigate the potential benefits of conversational course evaluation tools and to study the different modalities of voice and text-based interaction, we designed and built EVA (short for *Evaluation Agent*), an intelligent user-centered and theory-based conversational course evaluation tool that conducts conversational student evaluations with quantitative and qualitative questions. Specifically, we derived two hypotheses: (H1) *in comparison to a static web-based course evaluation, a user-centered conversational course evaluation tool increases response quality and user experience of students*; and (2) *in comparison to a text-based conversational course evaluation tool, a voice-based conversational course evaluation tool further increases responses quality and user experience of students.* Hence, we conducted a hierarchical evaluation set-up as a field experiment in a lecture with a total of 128 students. First, we compare our tool EVA against a static web survey with 93 of the 128 students (H1). In particular, we measured the impact of a conversational course evaluation in terms of quantity, quality, and clarity in information based on the Gricean Maxims [24] compared to a traditional web survey, the standard approach for course evaluations [4]. We observed that participants who used the text-based version of EVA to conduct the course evaluation answered with significantly higher response quality in terms of quantity and quality of information compared to a traditional web survey. Moreover, the users of EVA answered with more positive emotions in their course evaluations. Second, we then aimed to investigate the differences between a voice-based and text-based conversational human-computer interaction of EVA with 79 students (H2) from the same experiment. Against our hypothesis H2, we found that students provided answers with a lower information quantity but with higher quality when using a voice-based conversational interaction. Privacy concerns about voice interaction in the qualitative comments might be a first explanation for this observation.

In what follows, we first discuss and review related work on conversational agents in pedagogical scenarios and in survey contexts, as well as develop a set of hypotheses on how conversational agents alter response quality and user experience in the context of course evaluations. We then elaborate on how we developed EVA as an intelligent conversational course evaluation tool using a user-centered and theory-motivated design approach. Subsequently, we present the results of a field experiment designed to test our hypotheses. Finally, we discuss the theoretical and practical implications for the effective deployment and design of conversational agents for course evaluations.

2 RELATED WORK AND HYPOTHESES DEVELOPMENT

Our work and our hypotheses are inspired by previous studies on conversational interfaces in educational scenarios, by research about course evaluations, and by social response theory. We believe social response theory based on [45, 48] supports our two hypotheses that a user-centered design of an intelligent and conversational course evaluation tool leads to higher response quality and user experience.

2.1 Conversational Interfaces in Education

Conversational agents (CAs) are computer programs that are designed to communicate with users through natural language interaction interfaces [56, 63]. In today's world, CAs, such as Amazon's Alexa, Google's Assistant, and Apple's Siri are ubiquitous, with their popularity steadily growing [14, 37]. CAs are designed and implemented across various areas such as customer service [90], healthcare [36, 41], or education [30]. [27] define CAs in the education domain as distinct forms of learning applications that provide an individual and personalized interaction with students. There is a growing consensus that the interaction with conversational interfaces benefits students in their learning process. First of all, CAs offer a natural and intuitive way for students to express themselves in computer-mediated educational settings, which in turn drives learner outcomes (e.g., [55, 79, 85]). Second, CAs allow for an adaptive interaction path that goes beyond a static user interaction experience, thereby increasing student engagement [84, 87]. Third, the design of CAs often incorporates anthropomorphic features that further support educators in building trustworthy and engaging learning scenarios [84].

Inspired by the potential advantages of conversational over non-conversational user interfaces, the development of CAs in education goes back to the 1970s research stream of intelligent tutoring systems (ITS) (e.g., [3, 69]). Similar to a human tutor, these systems can present instructions, ask questions, and provide immediate feedback [40]. 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 [51, 79], motivators, or learning companions [32, 85]. While existing research on CAs in education has mainly focused on providing learning support for students [27, 66, 86], [83, 86] pointed out that CAs might also have potential as an evaluation tool in educational settings. However, literature that investigates the effects of CAs and their design on course evaluation response data is still scarce. Recent studies such as by [33, 89] indicate that CAs can perform part of a human interviewer's role, which in turn leads to high-quality response data. However, despite the rapidly developing knowledge base on different CA applications, current literature falls short in providing user-centered, theory-motivated, and empirically evaluated principles on how to design a conversational tool for course evaluations.

2.2 Course Evaluations in Education, Satisficing Behavior, and Response Quality

Today, course evaluations are a common feature used by higher education institutions to effectively gather student feedback [16, 68]. Course evaluation surveys can provide valuable insights into

teaching and course effectiveness, which allows educational institutions to react to changing student needs [4, 68]. They can be divided into quantitative and qualitative evaluation methods. The most frequent form of course evaluation is web surveys that usually include a battery of quantitative assessments and one or two qualitative-oriented questions at the end. Many institutions rely on web surveys as they improve the efficiency of the feedback collection and analysis process [33]. However, as outlined in the introduction section, web surveys are confronted with a number of issues, including a lack of interactivity and human warmth that lead to low acceptance rates and to a decrease in the quality of student feedback. [33, 67, 83, 89]. These issues cause problems such as survey fatigue [52, 73] and respondents' satisficing behavior [38]. Satisficing behavior is reflected in non-differentiation or straight lining, meaning an equal responsive behavior is used for an array of scaled questions [33, 38]. This behavior occurs because responding accurately and sincerely requires high levels of cognitive demands [39]. Satisficing responses lead to response errors, thus producing data of lower quality [38]. One way to reduce satisficing behavior is to use interactive and engaging survey formats that mirror conversations with a human educator [68]. In interpersonal contexts, it was shown that in-person surveys could promote more conscientious responses and encourage participation through personalized interaction. The presence of an interviewer can encourage students to participate in a survey, ask for clarification, and check their answers to confirm their sincerity [33]. These verbal and nonverbal interactions between students and lecturers promote accurate answers, which increases the feedback quality [28].

One way to provide students with a more interactive, adaptive, and human-like interaction experience in course evaluation settings is the design and deployment of CAs. As summarized in Table 1, three key factors help to understand the unique insights of related prior work with regard to (1) the nature or focus of computer-based assessments, (2) the type of assessments conducted, and (3) the kind of survey and user experience outcomes investigated. First, prior work has investigated the effects of conversational (as opposed to static) interactions showing that CAs are effective in decreasing respondents' satisficing behavior and increasing response quality (informativeness, relevance, specificity, and clarity) based on reciprocal message exchanges and conversational interactivity [33, 83, 89]. However, despite voice-based CAs becoming omnipresent across contexts, none of the previous studies examined how voice-based CAs affect response quality. Few studies examining the effect of interaction modality on response quality show that telephone-based assessment (as opposed to static web survey) can decrease satisficing behavior and increase data quality, hinting at the potential of voice-based CAs in survey contexts [19, 44, 54]. Second, most of the identified studies either focus on quantitative or qualitative assessment outcomes [89]. For a notable exception, refer to [83] which investigated differences between web surveys and chatbots on both quantitative and qualitative assessment outcomes. In the context of course evaluation, it is important to look at both types as educators are usually interested in getting both quantitative (i.e., overall course satisfaction) and qualitative-oriented insights (i.e., deep-level insights such as ideas on improvement). Third, there is very little research that both included survey and user experience outcomes to evaluate the effectiveness of CA-based survey tools across specific contexts such as course evaluations. As summarized in the review of Table 1, this paper examines the effectiveness of a fully conversational survey tool (as opposed to a static web survey) across two modalities (text vs. voice), incorporating both quantitative and qualitative assessments and examining both user experience and survey quality outcomes in the context of course evaluations.

2.3 Social Response Theory to Foster Students' Engagement in Course Evaluations

Our research is motivated by social response theory. According to social response theory, humans tend to respond socially to agents that display characteristics that are typically attributed to humans [45]. Behavioral cues and social signals from computers, such as using natural language or assuming

Source	Survey Context / Assessment Method	Modality / Device	User Experience	User Behaviour	Key Outcomes
[44]	Demographics / Quantitative	Static Web Survey vs. Static Telephone-based	n/a	Response Streamlining, Social Desirability	Static telephone-based assessments decrease the tendency to give social desirable answers in surveys.
[19]	Scenario-Based / Quantitative	Static Web Survey vs. Static Telephone-based	n/a	Response Streamlining, Satisficing Behavior (Dropout Rate, Non-responses, Response Time)	Static telephone-based assessments decrease dropout rates, non-responses, streamlining behavior. However, text-based assessment increased response times.
[54]	Market Research / Quantitative	Static Web Survey vs. Static Telephone-based	n/a	Response Behavior (Emotional Valence); Satisficing Behavior (Dropout Rate, Non-responses)	Web survey-based assessments lead to a lower response rate, more item omissions, and produced more negative and neutral evaluations.
[33]	Demographics / Quantitative + Qualitative	Statics Web-Survey vs. Text-Based Conversational Agent	Ease of Use, Enjoyment	Response Streamlining; Satisficing Behavior (Response Time; Dropout Rate)	Conversational agents increase response quality and decrease satisficing behavior. Moreover, perceptions of ease of use and enjoyment were enhance.
[89]	Market Research / Qualitative	Statics Web-Survey vs. Text-Based Conversational Agent	n/a	Response Quality (Informativeness, Relevance, Specificity, and Clarity); Participant Engagement (Engagement Duration; Response Length; Self Disclosure)	Conversational agents (as opposed to a static web-survey) increase both response quality and participant engagement.
[83]	Course Evaluation/ Qualitative	Statics Web-Survey vs. Text-Based Conversational Agent	Level of Enjoyment	Response Quality (Quality, Clarity)	Conversational agents (as opposed to a static web-survey) increase both response quality and level of enjoyment.
This Study	Course Evaluation / Quantitative + Qualitative	Static Web-Survey vs. Text-Based Conversational Agent vs. Voice-Based Conversational Agent	Interactional Enjoyability, Social Presence, Self-Disclosure and Qualitative Cluster Analysis	Response Quality in the form of Information Quality, Information Quantity, and Information Clarity; Response Time	Conversational surveys (as opposed to static surveys) enhance response quality and user experience. Voice based interfaces lead to higher information quality but lower information quantity compared to text-based modalities.

Table 1. Non-exhaustive overview on literature on conversational survey design.

a social role, subconsciously trigger social responses from humans, no matter how rudimentary

those cues or signals are [46, 49]. Following the "Computers are Social Actors" (CASA) paradigm, existing research has examined different social cues and their influence on human–computer interaction. Specifically, researchers and practitioners used anthropomorphic design elements to provide technologies with a human touch [22, 59]. Anthropomorphism describes the tendency to apply humanlike characteristics, motivations, intentions, or emotions to nonhuman agents [15]. Anthropomorphic design elements include, e.g., a humanlike appearance of agents, socially-oriented communication styles, or voice-based communication [17]. According to [17, 83], adding humanlike design cues can have a positive impact on technology acceptance, as they can elicit social responses.

Also, in the educational domain, anthropomorphic design elements were successfully leveraged to increase social presence in technology-mediated learning scenarios, which in turn enhances various learner outcomes [74]. Hence, it is reasonable to assume that using anthropomorphic dialogue systems such as a CA could bear the potential to overcome problems associated with online course evaluations, such as low acceptance rates and satisficing behavior. In fact, [60] show that a conversational interface compared to a static interviewing system is perceived as more engaging and more humanlike and thereby increases the feeling of social presence. Moreover, it was shown that a CA can effectively perform parts of a human interviewer's role, e.g., by applying effective communication strategies that increase social presence (e.g., [33, 83, 89]). Social presence shows the degree to which participants in computer-mediated communication feel affectively connected to each other [71]. Building on this research, we assume that the social presence experienced when interacting with a CA in a course evaluation setting may increase the likelihood of giving conscientious responses, thus increasing response quality [33]. Specifically, verbal and nonverbal interactions draw the respondents' attention and increase the social engagement with an agent, which should lead to more appropriate answers [28] - a mechanism we want to leverage in the design of a conversational course evaluation tool. Hence, based on prior literature on conversational course evaluations and social response theory, we derive the following hypothesis:

H1: *In comparison to a static web-based course evaluation, a user-centered conversational course evaluation tool increases the response quality and user experience of students.*

The recent proliferation of voice-enabled technologies raises the question of how communication modality (voice vs. text) affects response quality and user experience in course evaluations. From extant literature on media richness and multi-format communication, we know that voice-based (as opposed to text-based) communication affords a more synchronous (e.g., more immediate communication) and intuitive (e.g., less formal language, more flexible syntax) interaction experience [10, 91]. For instance, [91] show that voice-based interfaces lead to more immersive user experiences when filing an insurance claim, ultimately enhancing a wide range of perceptual and behavioral user outcomes. Furthermore, in the context of tasks that require information disclosure, a study by [44] has investigated that voice-based interactions compared to text-based interactions decrease the tendency to give socially desirable answers. Moreover, it has been found that talking to an interface makes people more willing to share sensitive information in a conversation setting [58]. In line with this work, we expect that the higher synchronicity and intuitiveness of voice-based exchanges enhance both response quality and user experiences in course evaluations [10]. We thus derive our second hypothesis:

H2: *In comparison to a text-based conversational course evaluation tool, a voice-based conversational course evaluation tool further increases responses quality and user experiences of students.*

3 DESIGN OF A CONVERSATIONAL COURSE EVALUATION TOOL

To investigate (1) the effects of a user-centered conversational course evaluation on students and (2) the differences between voice- and text-based evaluation modalities, we designed a CA called EVA. EVA is composed of two main components: an adaptive student-centered user interface and

an ML model to intelligently guide a student through a natural feedback conversation. The basic user interaction concept of EVA is illustrated in Figure 2. A user conducts a course evaluation in a natural conversation based on adaptive design features such as a progress bar, help functions, and a skip button.

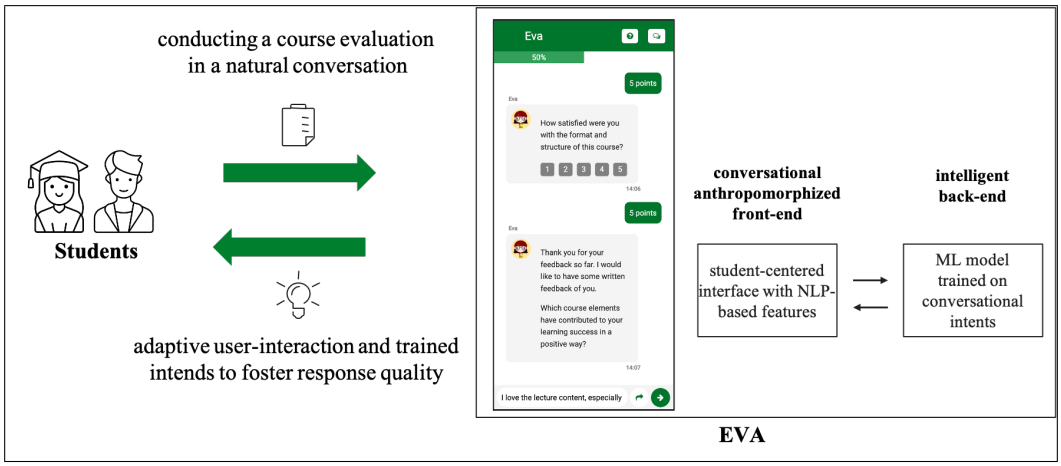

Fig. 2. Basic user interaction concept of EVA: a student conducts a course evaluation in a natural conversation based on an adaptive user interaction and trained intents to foster response quality.

3.1 User Interaction of EVA

3.1.1 Deriving Requirements from Literature and Users. In order to build a novel tool for course evaluations, we used two distinct but complementary approaches: a rigorous theory-driven approach and an agile user-centered approach following the build-measure-learn paradigm [53]. For the theory-driven approach, we followed [8] and [77] in conducting a systematic literature review with the aim of deriving theory-driven requirements for the design of an intelligent conversational course evaluation tool. To that end, our goal was to identify theory-driven design requirements on a conceptual level with a focus on an espousal of position and a representative coverage [8]. We identified two research areas to derive theory-driven design requirements: conversational design and human–computer interaction. To find relevant publications, we used Google Scholar to run a keyword search. We chose Google Scholar because it offers advanced full-text searching and a variety of filtering options for academic publications. We used the following keywords: „Course Evaluation", „Conversational Evaluation", „Conversational Survey", „Voice Survey", and „Speech Interaction Survey". To find the most relevant studies, we defined criteria for inclusion and exclusion and reviewed the titles and abstracts of our search results in the first step. We only included literature that deals with or contributes to a kind of conversational tool in the field of education. Several papers dealing with evaluation or voice analysis outside a conversational or educational domain were discarded. On this basis, we selected 51 papers for a more intensive analysis. We have summarized similar topics of these contributions, which served as theory requirements for the design of a conversational course evaluation tool. Requirements from literature included studies on the design of CAs in education [86], studies on improved response rates of online course evaluations [9, 67], or student perception and motivation for the assessment of course effectiveness [5].

Besides the rigorous theory-driven approach, we simultaneously followed an agile user-centered design approach. We conducted 20 semi-structured interviews with students to build an initial understanding of the needs and requirements of students with regard to course evaluations in general and for a conversational course evaluation tool in specific [21]. Our interview guideline consists of 33 questions and the interviews lasted on average (mean) 37.51 minutes. The interviewees were a subset of students at our university who are all potential users of conversational course evaluation tools. The interviewed students had a mean age of 23.8 years (SD = 2.23), and most students were enrolled in economics, law, or psychology; 13 were male, 7 were female. After transcribing the interviews, they were evaluated using qualitative content analysis. To that end, the interviews were coded and abstract categories were formed. The coding was performed using open coding to form a uniform coding system during evaluation [21]. We derived a total of 172 user stories and aggregated the most common ones following [6], resulting in eight common topics. The user stories treated requirements for the user interface design, the conversational flow, the confidential handling of results, and the efficient usage of a course evaluation tool. They gave us an overview of the needs and requirements of users for a conversational course evaluation tool.

Based on the user stories and the derived literature requirements, we designed several low-fidelity prototypes of EVA to pretest our design instantiations and to receive an initial understanding of the human–computer interaction with a conversational course evaluation tool. We pretested a number of iteratively developed design instantiations of EVA in different course settings in three different pre-studies with a total of 60 students (28 in pre-study 1, 12 in pre-study 2, and 20 in pre-study 3). The participants were all students from our university with a similar gender and age distribution to the ones from the interviews. We pretested different instantiations, e.g., to learn more about conversation strategies such as the length and the formulation of evaluation questions, the introduction and exit of the evaluation conversation, and the design of predefined buttons. For example, we experienced that students prefer to receive predefined answer buttons in the conversation introduction and in the quantitative question parts in order for them to go through an evaluation dialogue efficiently. To sum up our design findings, we derived five design principles on how to build an intelligent conversational course evaluation tool (see overview in Table 2). The design principles are instantiated as functionalities (e.g., F1 - F10) in the final version of EVA. An overview can be found in Table 2.

	Design Principle	**Instantiated Functionalities**
1)	"To design a student-centered conversational course evaluation tool", employ a conversational web-based agent, with a friendly, polite, and flexible interaction design in the educational institutions design to provide a student-centered and efficient evaluation experience in a natural dialogue.	e.g., F1, F2, F3, F4, F5, F6
2)	"...", employ intelligent conversational abilities for the agent to interact adaptively with the students and dig deeper into qualitative evaluation topics.	e.g., F6, F7, F11
3)	"...", design the agent with transparent explanation features and a "help" function, which constantly provides the student with information on how to use the agent.	e.g., F4
4)	"...", design the agent with student-centered trust-enhancing components to assure the user that their data is safe and kept private.	e.g., F9, F10
5)	"...", equip the agent with interaction feedback options, an overview of the user's responses, and predefined answers to enable a flexible interaction style for the user.	e.g., F5, F6, F8

Table 2. Overview of derived design principles on how to build an intelligent conversational course evaluation tool and instantiated functionalities addressing the principles.

3.1.2 User Interaction Design of EVA. Based on our design studies, we built EVA as a cloud-based web application that can be used on all kinds of devices (F1). A screenshot of EVA and its different functionalities (e.g., *F1 - F10*) can be seen in Figure 1 and Figure 3.

EVA guides students through a quantitative and qualitative student feedback conversation with the aim of imitating a human educator. The agent is designed in the corporate identity of our university (F2). For the sake of the anonymity of this submission, we blurred parts of this design. The CA proactively starts the evaluation dialogue and intelligently asks students predefined evaluation questions concerning a specific pedagogical scenario or a course. In general, the persona of EVA employs a friendly conversational style incorporating the use of emojis or colloquial language (F3). Moreover, we used an avatar to support students to empathize with the persona of EVA. The modeled interaction comprises a "declaration of consent" (see Figure 3, F10) at the start of the conversation. This assures the users that their data is only processed on university servers and only analyzed for the purpose of the course evaluation to make people feel more comfortable disclosing information. For the qualitative question parts, EVA is equipped with several NLP-based features, such as a sentiment analysis and an answer length analyzer (F11). If the user provides a feedback with less than five tokens or without any sort of sentiment (no matter if positive or negative), EVA asks a follow-up question to dig deeper into a certain topic. The NLP-based features, such as the word count and the sentiment analyzer, are implemented by utilizing the python-based library spacy[1]. For example, if a student answers a qualitative question with less than five tokens or no sentiment at all, EVA asks the user to elaborate a bit more on the topic (e.g., "Could you tell me more?").

Moreover, EVA provides students with the option to monitor the progress of the course evaluation through a progress bar with a percentage indicator (F2). This ensures the users' motivation to continue with the survey, as they can transparently monitor which part of the evaluation they are in at the moment and how long it might take in total. Next to the free text input field, we designed a skip button, which enables the user to decide against giving answers in the evaluation (F7). Interestingly, the skip button was rarely used in our pre-studies but showed positive effects on trust and engagement in the conversational tool. During the conversation, the users always have the choice to answer with predefined buttons or with freely written text (e.g., F6). The predefined answer buttons help the user to receive an overview of the evaluation process and ensure both flexibility and efficiency regarding user interactions with EVA. For example, EVA provides the users with predefined numerical buttons for the quantitative question parts to ensure a quick answer option (see Figure 1). Moreover, the user always has the option to receive an introduction or help on the user interaction of EVA (see Figure 1 and 3, F4). Also, EVA is provided with an always-present feedback button (see Figure 1 and 3, F6) for users to provide feedback to EVA whenever they want to. EVA is designed with a student-initiated casual chat mode. The user can ask EVA to tell jokes, fun facts, or talk about the weather to take a break from the evaluation (F3).

Finally, EVA closes every evaluation by giving the user the option to review their answers and change them if desired (F8). Moreover, if the users want, they can provide their email address to which the entire dialogue is sent (F9).[2]

3.2 Back end of EVA

For the adaptive back-end functionality of EVA, we utilized a combination of different NLP- and ML-based techniques. In general, EVA is built as a web app in HTML5 with CSS and JavaScript. The

[1] spacy.io

[2] We are only able to display selected design functionalities of EVA. For more insights, e.g., into other instantiated functionalities, such as the conversational introduction, the conversational closing, the adaptivity, or the casual chat mode, the interested reader might refer to the interaction videos in the appendix.

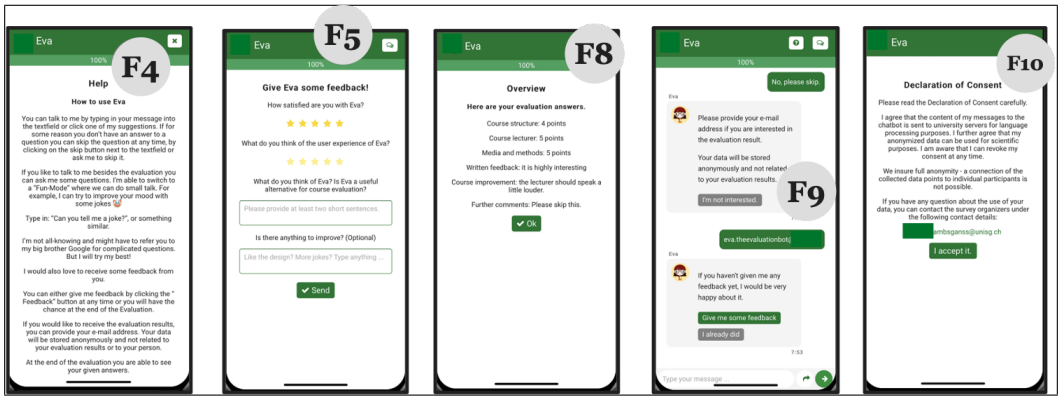

Fig. 3. Overview of different functionalities of EVA, including the feedback option, an overview function of the given responses, the declaration of consent, and explanations.

front end is connected to a python script that a) processes incoming user intents and b) provides predefined answers based on the incoming classifications.

For the conversational logic of a student-educator evaluation interview, we modeled 42 intents, including the introduction, the evaluation questions, and casual dialogue. The intents are then trained based on a "Naive Bayes classifier" in combination with semantic similarity matching as, e.g., also done in [55]. The conversational back end is implemented by utilizing the frameworks *chatterbot*[3] and *spacy*.

Moreover, as written above, we implemented several other NLP-based functions with *spacy*, such as a sentiment or a length analyzer. These functions increase the adaptivity of EVA, especially in the qualitative question parts, in order to re-ask or dig deeper if the answer contains too little information.

3.3 Voice-based and Alternative Course Evaluation Systems

In order to investigate the role of conversational modalities for course evaluations (H2), we designed a second version of EVA according to our design studies. The interaction of the text-based conversational version is displayed in Figure 1 and 3. The voice-based conversational version was built according to the same design findings based on the same back-end functionality, i.e., the same intents, conversational modes, the same interface design, and the same anthropomorphic features (e.g., the persona). We only manipulated the in- and output modality of the conversation between the two versions (see Figure 4). In the text-based version (version 1), students interacted with EVA through a chat interface. In the voice-based version (version 2), students interacted with the course evaluation tool using their voices. In order to do so, we added a speech-to-text function based on Google cloud services[4] and deleted the text input field. We did not manipulate any other interaction functionalities between the two versions. To evaluate both instantiations of our conversational course evaluation tool, we compared them to a standard web survey. The web survey was designed with the survey platform *unipark*. We chose this platform because it allowed us to design the survey similar to the traditional web-survey-based course evaluations used at our university. The design of the web survey is presented in Figure 4. The quantitative questions are answered through a

[3]https://chatterbot.readthedocs.io/en/stable/
[4]https://cloud.google.com/text-to-speech

simple matrix format to ensure that the same scaled options were used for multiple items to avoid repeating information. The qualitative items were answered through a simple plain text input field.

To keep the interaction of the two different versions of EVA and the web survey consistent with each other, there are many functions that are shared between all three tools. The entire course evaluation text, including the individual questions, is kept steady across the three course evaluation tools.

4 METHODS

To investigate our hypotheses, we conducted a field experiment where we asked students of a large-scale lecture to provide feedback on the lecture content and the teaching style of the lecturer (see Figure 4).

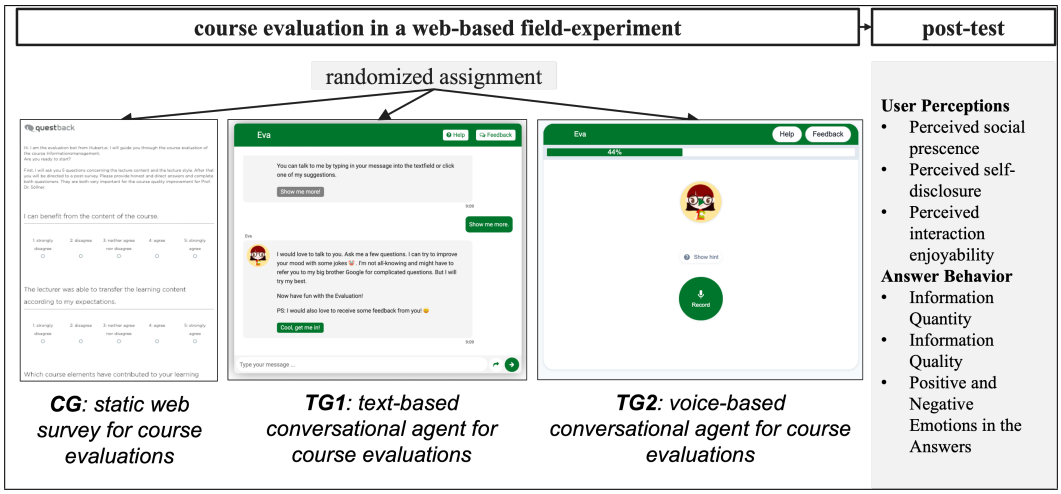

Fig. 4. Overview of our experimental design. Students were randomly assigned to either a control group (static web survey), treatment group 1 (text-based conversational agent), or treatment group 2 (voice-based conversational agent) to conduct the exact same course evaluation.

4.1 Participants

The students were invited to participate in the course evaluation through multiple channels. Students had a total of three days to complete the course evaluation. They could do this at a time and place of their own choosing. Students were randomly assigned to one of three tools. Participants in treatment group 1 (TG1) used the text-based instantiation of EVA according to our design findings to conduct the course evaluation (TG1 in Figure 4). Students in treatment group 2 (TG2) conducted the course evaluation based on the voice-based instantiation of EVA (TG2 in Figure 4). The participants of the control group (CG) used the standard web survey. The course evaluation questions were exactly the same for all groups, consisting of three quantitative and three qualitative questions following the standard procedure of the course evaluation of our university. In total, we received 128 valid answers from students participating in our field experiment. The experiment was conducted as a web experiment facilitated by the behavioral lab of our university and, thus, designed and reviewed according to the ethical guidelines of the lab and the university.

After randomization, we counted 44 valid results in TG1, 35 in TG2, and 49 in the control group[5]. Participants of TG1 had an average age of 24.86 (SD= 1.56); 36 were male, and 8 were female. In TG2, the average age was 24.2 (SD= 2.35); 25 were male, and 10 were female. In the CG, students had an average age of 25.02 (SD = 2.04); 33 were male, and 16 were female. In total, most participants took about 10 to 15 minutes to finish the experiment. The experiment was part of the curriculum of a graduate-level course at our university. The course evaluation was targeted at evaluating the student experience of this course.

4.2 Experimental Design

In the following, we will describe the experimental design. The results are then analyzed in a hierarchical evaluation, where we separably compare CG vs. TG1 for H1 and TG1 and TG2 for H2.

The experiment consisted of three main parts: 1) randomization, 2) course evaluation, and 3) a post-test. The post-test phases were consistent across all treatment groups. In the course evaluation phase, TG1 used the text-based version of EVA, TG2 the voice-based version of EVA, and the CG a standard web survey to conduct the same course evaluation.

1) 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 three different groups (TG1, TG2, and CG). The course evaluation was conducted in the middle of the lecture period.

2) Course evaluation: In the course evaluation phase, we asked all participants the same questions: three quantitative and three qualitative questions, following the standard procedure of course evaluation at our university. The first three questions addressed the perceived benefit of the course, the expectation for the lecturing style, and the satisfaction with the course format and structure. The questions were measured with a 5-point Likert scale (1 means strongly disagree, 5 means strongly agree, and 3 is a neutral statement). The exact items can be found in Table 8 in the appendix. 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?"*).

3) Post-test: In the post-survey, we measured the perceived social presence of the evaluation tool by asking five items based on [20], such as *"There is a sense of human contact in the course evaluation tool"*. Moreover, we assessed self-disclosure based on [7, 42] with five items, such as *"The course evaluation tool and I exchanged enough personal information."*. Furthermore, we captured the construct of interaction enjoyability [35, 42, 75] of all participants with five items, for instance *"It is fun and enjoyable to share an interaction with the evaluation tool."* All constructs were measured on a 1-to-5-point Likert scale (1 means totally disagree, 5 means totally agree). Furthermore, we asked three qualitative questions, which were "*What did you particularly like about the use of the course evaluation tool?*", "*What else could be improved?*", and "*Do you have any other ideas?*, and captured the demographics.

Also, to test whether the randomization was successful, we asked four items to assess the personal innovativeness in the domain of information technology of the participants following [1]. In total, we asked 28 questions, including two control questions (manipulation and attention check). [6]

[5]Initially, we received 47 answers for TG2. However, we had to delete twelve responses since they were incomplete in the course evaluation part and in the post-test. We only incorporated completed data points in our sample.
[6]All the items and variables of our study can be found in the appendix.

4.3 Measurement of Response Quality

Besides comparing theoretical constructs between the three groups, our main objective was to measure the response quality of the received answers. Hence, we used three objective measurements to capture the quality of students' responses. The objective variables are based on the lexical, syntactical, and semantic quality of the provided text answers of the students from the course evaluation.

For assessing the lexical, syntactical, and semantic quality of the provided answers, we followed [89] and applied the Gricean Maxims to determine response quality [24]. The Gricean Maxims provide a set of principles to measure the effectiveness of communication and have been applied as an information quality metric in survey responses (e.g., by [89]). Accordingly, the response quality of a student is measured in *1) quantity of information, 2) quality of information, and 3) clarity of information*. To provide an objective judgment, we followed [83] to measure these categories based on the lexical, syntactical, and semantic information of the provided answers. Hence, the answers to all three qualitative questions from the course evaluation were combined into one string and analyzed with state-of-the-art NLP-based techniques provided by the python-based libraries *TextBlob* and *spacy*.

1) Quantity of information: For the lexical information quantity, we preprocessed the answer string by tokenizing and lemmatizing the words. Afterward, we eliminated all stop words of the English language to ensure that no unnecessary information (e.g., the, a, and) are counted within this measurement. We took the length of the final tokens as a quantity of information measurement of the responses of a single student.

2) Quality of information: To measure the quality of the answers (e.g., the readability), several measures have been used in research [31]. Following [83], we selected the Flesh-Reading-Ease (FRE) [18] to capture the readability of received responses since the score combines language complexity measurements such as the average sentence lengths and the average syllables per word into one number [18]. The score has been widely used before to determine the readability of a message in computer-mediated communication [78] or for the complexity of CA user responses [22, 83]. Following Flesch (1943) [18], we used the following formula since we received answers in English:

$$\textbf{Flesch Reading Ease} = 206.835 - (1.015 \cdot asl) - (84.6 \cdot asw)$$

***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 language, and thus the higher the information quality (i.e., [18, 81]).

3) Clarity of information: We aimed to analyze the emotions of our received responses since emotional sentiments are a good indicator for what position an individual is taking on a certain topic [50]. For example, if a student only answers *"slide content in lecture unit 4"*, no clear action can be derived since this message contains no opinion. A clearly communicated emotion and thus a *"position taking"* is valuable for the use case of course evaluation, similar to opinion mining [50] or language complexity measurements [29]. Hence, we follow but refine the approach of [83] by not only using the polarity score of the sentiments but by capturing the core emotions of every student response [13]. We used the ML model of *text2emotions*[7], since the recent model is trained on text messages and provides a dictionary-based result based on the five core emotions following [13] (i.e., *happy, angry, sad, surprised, afraid*). For each student response string, we calculate the normalized scores of the positive (happy, surprise) and negative (angry, sad, fear) emotions of every student response. We decided to separately measure positive and negative emotions to avoid a leveling out. Positive and negative emotions represent conceptually different constructs and

[7] https://github.com/aman2656/text2emotion-library

provide different insights into human experiences [13]. The result is given as a normalized score for each emotion category from 0 to 1. The higher the score, the higher the clarity of the position taken in the student's response.

5 RESULTS

To better communicate our findings, we follow a hierarchical evaluation set-up, in which we first compare participants from the CG and TG1 to validate H1. Afterward, we compared TG1 and TG2 to shed light on the conversational interaction modalities of course evaluations to investigate H2. To ensure that the randomization resulted in randomized groups and to control for potential effects of interfering variables with our small sample size, we compared the differences of the construct of personal innovativeness. We received p values larger than 0.05 between the two treatment groups and the control group.

5.1 H1: Results of a Conversational Course Evaluation on Students' Response Quality and User Experience

To evaluate our first hypothesis, we compare the measured response quality of the provided answers (Table 3) and the self-reported constructs (Table 4) from participants of the CG and TG1. We applied a Welch's double-sided t-tests to control for statistical significance between the groups [11].

Group	n	**Information Quantity** (in number of tokens from 0 to 214)	**Information Quality** (in FRE from 0: low to 110: high)	**Positive Emotions** (0: low, 1:high)	**Negative Emotions** (0: low, 1:high)
TG1: text-based EVA	44	m= 59.45 (SD= 35.77)	m= 69.35 (SD= 12.67)	m= 0.44 (SD= 0.21)	m= 0.55 (SD= 0.21)
CG: web survey	49	m= 37.16 (SD= 28.70)	m= 59.45 (SD= 24.62)	m= 0.28 (SD= 0.27)	m= 0.49 (SD= 0.35)
p-value	93	0.0014**	0.0156*	0.0026**	0.2787

Table 3. Overview of the results on information quantity, information quality, positive and negative emotions across users of the text-based conversational version of EVA and the static web survey (*p < 0.05, **p < 0.01, ***p < 0.001).

Group	**Interactional Enjoyability**	**Perceived Social Presence**	**Self-Disclosure**
TG1: text-based EVA	m= 2.91 (SD= 1.13)	m= 2.72 (SD= 0.89)	m= 3.58 (SD= 0.54)
CG: web survey	m= 2.38 (SD= 0.87)	m= 2.06 (SD= 0.82))	m= 3.11 (SD= 0.76)
p-value	0.0056**	0.0006***	0.0012**

Table 4. Results on user experience across TG1 and CG measured on a 1-to-5-point Likert scale (1 means totally disagree, 5 means totally agree, *p < 0.05, **p < 0.01, ***p < 0.001).

Students using the text-based version of EVA (TG1) wrote their responses with a mean of 59.45 tokens (SD= 35.77). For participants in the control group, we counted a mean of 37.16 tokens per student (SD= 28.70) (see Table 3). The difference between the groups is statistically significant. The double-sided t-test confirmed that students using the text-based version of EVA wrote their text with a significantly higher quantity of information compared to the web survey (p= 0.0014). Also, we observed that students from TG1 wrote their answers with a mean readability score of 69.35

(SD= 12.67). Participants of the CG wrote their answers with mean syntactical readability of 59.45 (SD= 24.62). Hence, students in TG1 wrote their text with significantly higher readability compared to the control group (p= 0.0156).

Concerning the emotions, we captured the positive emotions (happy, surprise) and the negative emotions of the student responses (angry, sad, fear). We summed up the normalized scores of each emotion and calculated a total score for the positive and the negative emotions. Students in TG1 wrote their text with a significantly higher measured amount of positive emotions compared to the CG (mean TG1: 0.44, SD: 0.21, mean CG: 0.28, mean SD: 0.27, p= 0.0026). For students in TG1, we calculated a mean of 0.55 (SD= 0.21) for negative emotions, and in the CG, a mean of 0.49 (SD= 0.35). No significant difference was found in the double-sided t-test (p= 0.2787). To control that students are not unwillingly less critical in the way they communicate their feedback (i.e., controlling for unconstructive but positively framed responses), we manually compared qualitative answers in both conditions (CG vs. text-based version of EVA). Specifically, we identified a random subset of 10 comments per condition. One of the researchers identified all issues raised in each of these comments from an educator-based perspective. These results were discussed with the other researchers to check for potential misunderstandings and bias. We did not find any qualitative differences with regard to the extent of issues being raised. However, we found that responses given via the text-based version of EVA were more positive in tone (e.g., by writing improvement suggestions instead of harsh criticism; see attached an illustrative example of such a student feedback: "The overall course was useful as we students are about to tackle our master theses. Overview of research design and different research approaches dependent on which research problem one wants to solve.") Next, we calculated the average of the items for self-disclosure, interactional enjoyability, and perceived social presence to investigate the differences between a conversational course evaluation tool and a web survey on students' user experiences.

An overview of the results for user experience can be found in Table 4. First, we calculated the average of the five items of the construct of self-disclosure [7]. Students conducting the course evaluation using the text-based version of EVA (TG1) according to our design studies rated their self-disclosure with a mean of 3.58 (SD= 0.54). Students using the standard web survey rated their self-disclosure with a mean of 3.11 (SD: 0.76). The double-sided t-test revealed that the differences are statistically significant (p= 0.0012). The interactional enjoyability of the text-based version of EVA was rated with a mean value of 2.91 (SD= 1.13). Students using the traditional web survey judged the enjoyability with a mean value of 2.38 (SD= 0.87). The differences are statistically significant according to a double-sided t-test (p= 0.0056). Students of TG1 (text-based version of EVA) judged the social presence of the course evaluation tool with a mean of 2.72 (SD= 0.89), whereas participants in the CG judged the social presence of the web survey tool with a mean of 2.06 (SD= 0.82). The differences between the conversational course evaluation tool according to our design studies are statistically significant compared to the static web survey (CG) (p= 0.0006).

5.2 H2: Results on the Conversational Interaction Modalities of Course Evaluations on a Students' Response Quality and User Experience

To investigate the differences between a text-based conversational agent and a voice-based conversational agent on a students' response quality and user experience, we compared the results of the participants from TG1 and TG2 (see Table 5 and Table 6). Students who conducted the same course evaluation with the voice-based version of EVA (TG2) answered in mean 21.82 tokens (SD= 10.96) as compared to 59.45 tokens (SD= 35.77) from TG1. The double-sided t-test confirmed that students using the text-based version of EVA wrote their text with a significantly higher quantity of information compared to the voice-based version (TG2, p< 0.001). Moreover, students from TG2 had a mean readability score of 76.38 (SD= 15.68). The effect is statistically significant based on the

Group	n	**Information Quantity** (in number of tokens from 0 to 214)	**Information Quality** (in FRE from 0: low to 110: high)	**Positive Emotions** (0: low, 1:high)	**Negative Emotions** (0: low, 1:high)
TG1: text-based EVA	44	m= 59.45 (SD= 35.77)	m= 69.35 (SD= 12.67)	m= 0.44 (SD= 0.21)	m= 0.55 (SD= 0.21)
TG2: voice-based EVA	35	m= 21.82 (SD= 10.96)	m= 76.38 (SD= 15.68)	m= 0.37 SD= 0.32	m= 0.45 (SD= 0.34)
p-value	79	< 0.001***	0.035*	0.3016	0.1252

Table 5. Overview of the results on information quantity, information quality, positive and negative emotions across users of the text-based and the voice-based conversational interaction of EVA (*p < 0.05, **p < 0.01, ***p < 0.001).

FRE (p= 0.035). No significant difference was found for the positive emotions between the voice and text modality (mean TG2: 0.37, SD: mean TG1: 0.44, SD: 0.21; p= 0.3016). Also, no significance was found for the negative emotions between TG1 and TG2 (TG1: mean 0.55, SD= 0.21; TG2: mean 0.45, SD= 0.34; p= 0.1252).

Group	**Interactional Enjoyability**	**Perceived Social Presence**	**Self-Disclosure**
TG1: text-based EVA	m= 2.91 (SD= 1.13)	m= 2.72 (SD= 0.89)	m= 3.58 (SD= 0.54)
TG2: voice-based EVA	m= 2.90 (SD= 0.74)	m= 2.63 (0.76)	m= 3.46 (SD= 0.48)
p-value	0.974	0.8865	0.6603

Table 6. Results on user experience across the groups measured on a 1-to-5-point Likert scale (1 means totally disagree, 5 means totally agree, *p < 0.05, **p < 0.01, ***p < 0.001).

Concerning the user experience, participants using the voice-based version (TG2) judged their self-disclosure with a mean of 3.46 (SD= 0.48). No significant differences were found between TG1 and TG2 (p= 0.6603). The interactional enjoyability of the text-based version of EVA was rated with a mean value of 2.91 (SD= 1.13), and the average for the voice-based version was 2.90 (SD= 0.74). No difference was found between both groups (p= 0.97401). In our experiment, students of TG1 (text-based version of EVA) judged the social presence of the course evaluation tool with a mean of 2.72 (SD= 0.89). Students in TG2 (voice-based version of EVA) rated the social presence with a mean of 2.63 (0.76). The differences are not significant according to a double-sided t-test (p= 0.8865).

With regard to the quantitative assessment, we did not find any significant differences with regard to how the course was evaluated. Also, we did see any significant differences in straight-lining. However, we did find differences with regard to the overall response time. Students in TG1 spend 18.81 min in mean to conduct the course evaluation (SD= 7.88 min). Students in CG needed 14.96 (mean) min (SD= 9.55) and TG2 had a mean usage time of 18.01 min (SD= 7.11). When comparing both conversational settings (TG1 and TG2) together, the difference to the static web survey in the usage time is statistically significant (p= 0.0497).

5.3 Qualitative User Feedback

We also asked open questions in our survey to receive the participants' opinions about their user experience with the respective tool they used and to shed led light on possible psychological mechanisms explaining our findings. Based on the participants' answers, we created a thematic map by identifying and clustering topics embedded in the data. In sum, four major themes emerged (please see Table 7 for representative quotes for each thematic cluster):

(1) Attitude towards conversational agents: The first topic that emerged was related to the deployment of CAs for course evaluations. Generally, the participants' welcomed the introduction

of innovative technologies in educational settings. Specifically, they were positive about the use of CAs due to prior experiences. This also had positive downstream consequences for the institution itself, as students felt that the university was making an effort to improve teaching quality. Moreover, the institution itself was perceived as being innovative (as compared to its peers). However, few participants also rejected CAs as a technology by principle. This could indicate that students should be offered to choose different interfaces according to their preferences.

(2) Interactivity creates experiences even when doing mundane tasks: The results show that participants using EVA felt more positive about conducting course evaluations, a task that is often perceived by students as a burden rather than a value-adding activity. This is in line with previous findings from [33]. Specifically, the participants highlighted that providing feedback to the CA felt like talking to somebody, which motivated them to stay committed throughout the interaction. Moreover, it was mentioned that a feeling of being "heard" emerged, ultimately making students feel more positive about giving feedback. However, to some, the high interactivity of EVA also increased task complexity, particularly when using the voice-based version of EVA. One participant commented that due to the high immediacy of voice-based interactions, it would be difficult to provide thoughtful responses. This is in line with media synchronicity theory, according to which higher synchronicity of a communication channel can negatively impact user outcomes when the need for behavioral coordination is high [12]. One solution to this could be to allow students to edit transcripts of their recordings after the interaction.

(3) Leveraging the potential of the conversational logic: Generally, the conversational interaction paradigm was evaluated in positive terms as participants did not only have more fun interacting with the interface but also cited efficiency gains as a benefit. Specifically, it was mentioned that using the voice-based version of EVA to answer qualitative assessments was more convenient and time-saving. However, the interviews also indicate that a key requirement for positive user experiences is to provide guidance throughout the conversational interaction, particularly for the voice-based interaction [64]. In this regard, the progress bar, in particular, was highlighted as a favorable feature. Finally, with regard to the voice-based interaction, some users were highlighting privacy concerns when providing feedback via voice, which may have quieted the positive effects of voice-based modalities on response quality and user outcomes. This is in line with findings of [47] showing that voice-based modalities decrease the willingness to disclose sensitive information when completing an online survey on sensitive topics. Specifically, speech includes a wider range of social cues that afford for the identification of individual users [26]. Hence, one key requirement for the success of voice-based evaluations may be to ensure users of effective anonymity measures [76].

(4) Providing the right amount of anthropomorphic features: The last topic was related to the degree of anthropomorphism of EVA and has been raised again and again in the process of EVA's user-centered development. Thus, it was important to the participants that the bot is presented as not too human-like. For example, one participant highlighted the importance of using an avatar of a fictitious personality. On the other hand, the follow-up questions from EVA and the opportunity to have a joke told were positively highlighted. In line with [88], our results indicate that the use of anthropomorphic features must be carefully tailored to the user group to avoid feelings of creepiness.

6 DISCUSSION

In our study, we aimed to build on the recent literature on conversational evaluations and surveys (e.g.,[33, 81, 89]) to (1) enrich current insights on potential positive effects of a user-centered course evaluation tool for educational settings in a field experiment (H1) and (2) dig deeper into the modalities and design options of conversational evaluation tools by investigating the differences

Cluster	Feature
Attitude towards conversational agents	*"I like the new technologies and I find chatbots to be an innovative way of communication for course evaluations."*
Interactivity creates experiences even when doing mundane tasks	*"It gives you more of a feeling of talking to somebody. It is more interacting and therefore more exciting than a standard survey."*
Leveraging the potential of conversations for surveys	*"I only focused on the question that I was asked and could not skim all the questions at a glance. This actually made me think more about each question."*
Providing the right amount of anthropomorphic features	*"I liked that you named it after a human and that it's not a photo but a comic profile picture and that you asked for pardon when not understanding something."*

Table 7. Representative examples of qualitative user responses for EVA.

between a voice-based human-computer interaction and a text-based interaction for conversational course evaluations (H2).

Our results confirm H1 that in comparison to a static web-based course evaluation, a user-centered conversational course evaluation tool increases response quality and user experience of students. The response quality was significantly higher for the qualitative assessments with regard to the quantity and the quality of information using a text-based conversational interface for the exact same evaluation task compared to a static survey. Our results indicate that the higher interactivity afforded by conversational interactions enhances response quality for qualitative question types. Accordingly, a dialogue-based evaluation episode increases the engagement of the student compared to a static evaluation episode (i.e., a traditional web survey). Therefore, our results show that the theory-based and user-centered design of a conversational interface (i.e., EVA) combined with intelligent algorithms fosters student participation in course evaluations and increases response quality. Moreover, we saw that students seem to be positively biased in the emotions of their answers when using a conversational evaluation tool. Participants using EVA wrote their answers with significantly more positive emotions compared to the web survey. The positive notion of users' answer behavior is also reflected in the perception measures. Self-disclosure, perceived social presence, and interactional enjoyment were significantly more positively evaluated with EVA as compared to the traditional web survey. Specifically, a high level of enjoyment during the evaluation episode is important for the long-term adoption of such information retrieval tools, since this is likely to counter satisficing behavior and survey fatigue in course evaluations. These results are consistent with past studies that investigated the beneficial effects of CAs over non-adaptive systems (such as surveys) (e.g., [33, 81, 89]). One reason for these effects might be that conversational interfaces, compared to static web surveys, better direct the attention of the user to the question at hand [33]. Moreover, we believe social response theory could explain these effects as students might feel more inclined to answer a question more thoroughly when they have the feeling that another entity is present [44, 49]. Also, our results suggest that the increased user experience when using conversational interfaces might help to overcome the common challenges of surveys in general, such as survey fatigue [73] or satisficing behavior [25], and thus lead to better response quality.

However, we have to reject H2 that in comparison to a text-based conversational course evaluation tool, a voice-based conversational course evaluation tool further increases responses quality and user experiences of students. The results implied that a voice-based interaction does not automatically improve response quality or user experiences. Although the measured syntactical complexity of the

answers (FRE, information quality) was significantly higher of the voice-based group compared to the text-based interaction, the information quantity was lower. Against our expectation, this might imply that students answer less but with higher quality when conducting a course evaluation with a voice-based interaction modality. The emotions and the perceived user experience (enjoyability, social presence, and self-disclosure) seem to not significantly differ between text and voice-based interaction. Both were equally high when compared to a static web survey. Interestingly, the interaction time between both conversational tools (voice and text) was significantly higher together compared to the static web survey. This indicates that students are willing to spend more time when using a conversational evaluation. However, these effects regarding time differences may also be explained by differences in the way how users operate the different interface modalities. Hence, future research should take a more fine-granular perspective on response behavior when using a voice-based as opposed to a text-based interface.

Therefore, our study makes several contributions to research on the effect of interface modality in the context of course evaluations. First, although the voice-based version of EVA was on par with the text-based version with respect to perceived user experience, according to our qualitative data, two major issues arose that may inhibit response quality: A) users did not trust the ability of the speech-to-text engine to correctly transcribe their input. B) users were worried about providing incorrect or ambiguous information. In particular, users missed the option to rehearse and iteratively build on arguments like it is possible in text-based interface modalities. At the same time, this lack of behavioral control when interacting with a voice-based interface may provide a heightened potential to elicit more truthful responses - a potential we aim to investigate in future studies. Second, our qualitative data suggests that some of the functional issues can be attributed to the, in fact, still poor quality of the speech-to-text engine since long answers were not always recorded, and thus, users, in some instances, had to go through several iterations of the same interaction turn. Hence, in a future version of EVA, we want to further improve the functionality of the voice bot. Moreover, we aim to provide a multimodal option of EVA to fully capture and augment the potential of different modalities (i.e., voice to quickly formulate first thoughts and text to refine those initial thoughts or just to give users the option to choose their preferred modality). We believe, based on our results and as the functionality of speech-based technologies further matures and people become more familiar with using them, that there is enormous potential in voice-based conversational evaluation tools in education and for collecting survey data in general.

In sum, our work makes several contributions to current research. This study is one of the first to present empirical insights into how to design a dialogue-based information tool for course evaluation to increase response quality and user experience based on adaptive and intelligent interaction. Moreover, our study is the first to contrast the effect of text- and voice-based modalities in digital course evaluations. Thus, our research provides a foundation for researchers who also aim to develop information retrieval tools in different settings to compare their solutions with ours. Educational institutions and educators can use our design principles to build their own adaptive and dialogue-based course evaluation tools in their large-scale and distance-learning scenarios.

Based on the user feedback, we see two main improvements. First, we aim to improve the user interface of EVA such that it gives more control and overview to students when providing feedback – a key theme that emerged in the text-based and in the voice-based version. In the text-based version, we aim to design the text input field in such a way that it allows to comfortably write longer texts but at the same time preserves the conversational logic. Moreover, we will provide a multimodal version of EVA that allows choosing the preferred way of proving data (text or voice-based). Second, we want to train EVA on improved corpora so that the agent can follow up more elaborately on the statements of users.

7 LIMITATIONS AND FUTURE WORK

This research has several limitations that need to be acknowledged. First, we used a student sample from a university in a Western European country. While we believe that it is reasonable to assume that our results can be replicated in similar contexts, we believe that intercultural differences may play an important role. Hence, we invite other researchers to conduct similar studies, especially in contexts where different results are to be expected. For instance, in some countries, the use of voice-based modalities is much more common, so students might be more inclined to give feedback via voice. Second, although 74 percent of students indicated that they had interacted with a CA before, novelty effects cannot be ruled out. Therefore, it would be interesting to study whether these results hold across different evaluation episodes when students get used to this new form of course evaluations. Thus, we would urge future research to employ longitudinal designs in a real-world setting to study the effect of conversational interfaces on response quality over a number of evaluation episodes. Third, our used metrics and measurements of response quality come with natural limitations. Although we followed different literature on measuring response quality (e.g., [33, 81, 89]) based on the lexical, syntactical, and semantic information of the provided answers to enhance the objectivity of our results, the measurements we used might be biased and only show a limited perspective of information quality. For example, to measure the clarity of information, we not only used traditional sentiment analysis approaches to measure the valence of a string. We base our approach on a state-of-the-art ML model, which was trained on text messages and provides a dictionary-based result based on the five core Ekman emotions, which are rooted in the psychological literature ([13]; i.e., happy, angry, sad, surprised, afraid). Nevertheless, the used ML model and the conceptual background of [13] is also only representing a limited and potentially biased perceptive on information clarity. Hence, further studies are needed to confirm our results with more precise measurements. With ML algorithms advancing and also measuring more sophisticated constructs of human communication such as empathy (e.g., [82]) or argumentation (e.g., [80]), we believe more precise methods to measure the Gricean Maxims might evolve in future work that can be used to assess response quality more accurately [24]. In addition, future research could investigate more advanced metrics on response quality. For example, it is conceivable that deploying conversational interfaces in surveys influences feedback specificity.

Our results also suggest fruitful avenues for future research that could further support researchers and practitioners to exploit the potential of conversational technologies for all sorts of surveys. First, future research could manipulate different dimensions of human speech to improve response quality and user experience. Using voice-based modalities afford the use of an additional layer of social cues manifested in human speech that increase "human touch" in conversational technologies [57, 61]. Specifically, researchers could morph voice pitch, which has been shown to affect a wide range of user perceptions. Specifically, the literature suggests that voice pitch can influence trust attributions towards the interface [57, 72]. Still, it is unknown whether pitch in synthesized voices can affect trust-related behaviors such as information disclosure. Hence, future researchers could study the effect of voice pitch on disclosure behaviors of participants in survey contexts or other educational scenarios, such as writing reflective journals. Second, future work could explore when conversational interfaces increase response quality in survey contexts and when they become a nuisance or even detrimental. Especially, sensitive contexts such as providing potentially "shameful" information may be negatively impacted by interfaces that facilitate high social presence (i.e., conversational agents). Hence, future researchers could systematically explore contexts when conversational survey tools may foster response quality and when not. Third, future work might also address how voice analytics can be used to enrich survey data. Voice features reveal a host of behavioral markers that allow drawing inferences on the emotional status of a participant [26].

These insights could be linked to enriching participants' actual responses in course evaluations and other survey contexts (e.g., scientific surveys). For instance, emotional inferences could be used to rate the importance of answers or to evaluate their veracity.

8 CONCLUSION

In our research project, we designed, built, and evaluated EVA - an adaptive conversational course evaluation tool. EVA individually conducts evaluations with students by interactively asking both quantitative and qualitative questions, which follows up with the student when there is unclear input and engages students by using established design approaches (i.e., telling jokes). We compared two versions of EVA – a text-based and a voice-based interface – to a traditional web survey format in an experiment with 128 participants. We found that students using EVA to conduct a course evaluation gave more comprehensive feedback judged by the quantity and the quality of the information given. Moreover, the perceived social presence and enjoyment offer promising results for using EVA as an evaluation tool in different educational scenarios. In sum, our research offers specific design knowledge to further improve dialogue-based information retrieval systems in educational scenarios based on techniques from NLP and ML. With further advances of these technologies, we hope our work will attract researchers to design more intelligent information retrieval systems for different educational scenarios or other areas (i.e., scientific surveys) and thus contribute to the improvement of response quality in surveys, ultimately leading to higher data quality.

REFERENCES

[1] Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly* 24, 4 (12 2000), 665. https://doi.org/10.2307/3250951

[2] Mamta Agrawal. 2004. Curricular reform in schools: The importance of evaluation. *Journal of Curriculum Studies* 36, 3 (2004), 361–379. https://doi.org/10.1080/0022027032000152987

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

[4] Erik Blair and Keisha Valdez Noel. 2014. Improving higher education practice through student evaluation systems: Is the student voice being heard? *Assessment and Evaluation in Higher Education* 39, 7 (2014), 879–894. https://doi.org/10.1080/02602938.2013.875984

[5] Yining Chen and Leon B. Hoshower. 2003. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education* 28, 1 (2003), 71–88. https://doi.org/10.1080/02602930301683

[6] Mike Cohn. 2004. *User Stories Applied For Agile Software Development.* Technical Report.

[7] Nancy L. Collins and Lynn Carol Miller. 1994. Self-Disclosure and Liking: A Meta-Analytic Review. *Psychological Bulletin* 116, 3 (1994), 457–475. https://doi.org/10.1037/0033-2909.116.3.457

[8] Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society* 1, 1 (1988), 104–126. https://doi.org/10.1007/BF03177550

[9] Tena B. Crews and Dylan F. Curtis. 2011. Online Course Evaluations: Faculty Perspective and Strategies for Improved Response Rates. *Assessment and Evaluation in Higher Education* 36, 7 (2011), 865–878. https://doi.org/10.1080/02602938.2010.493970

[10] Richard L. Daft and Robert H. Lengel. 1986. Organizational Information Requirements, Media Richness and Structural Design. *Management Science* 32, 5 (1986), 554–571. http://www.jstor.org/stable/2631846

[11] Marie Delacre, Daniël Lakens, and Christophe Leys. 2017. Why psychologists should by default use welch's t-Test instead of student's t-Test. *International Review of Social Psychology* 30, 1 (2017), 92–101. https://doi.org/10.5334/irsp.82

[12] Alan R. Dennis, Robert M. Fuller, and Joseph S. Valacich. 2008. Media, Tasks, and Communication Processes: A Theory of Media Synchronicity. *MIS Quarterly* 32, 3 (2008), 575–600. http://www.jstor.org/stable/25148857

[13] Paul Ekman. 1992. An Argument for Basic Emotions. *COGNITION AND EMOTION* 6, 3/4 (1992), 169–200.

[14] eMarketer. 2017. *Alexa , Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year.* Technical Report. https://www.emarketer.com/Articles/Print.aspx?R=1015812

[15] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review* 114, 4 (2007), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
[16] Malgorzata Erikson, Martin G. Erikson, and Elisabeth Punzi. 2016. Student responses to a reflexive course evaluation. *Reflective Practice* 17, 6 (2016), 663–675. https://doi.org/10.1080/14623943.2016.1206877
[17] Julia Fink. 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 (2012), 199–208. https://doi.org/10.1007/978-3-642-34103-8{_}20
[18] R Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education* 897 (1943).
[19] Scott Fricker, Mirta Galesic, Roger Tourangeau, and Ting Yan. 2005. An experimental comparison of web and telephone surveys. In *Public Opinion Quarterly*, Vol. 69. Oxford Academic, 370–392. https://doi.org/10.1093/poq/nfi027
[20] David Gefen and Detmar W. Straub. 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 (1997), 389–400. https://doi.org/10.2307/249720
[21] Jochen. Glaser and Grit. Laudel. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen*. VS Verlag fur Sozialwiss. http://www.springer.com/de/book/9783531172385
[22] Ulrich Gnewuch, Stefan Morana, Marc T P Adam, and Alexander Maedche. 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
[23] Anthony G. Greenwald and Gerald M. Gillmore. 1997. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology* 89, 4 (1997), 743–751. https://doi.org/10.1037/0022-0663.89.4.743
[24] H.P. Grice. 1975. Logic and Conversation. In *Speech Acts*. BRILL, 41–58. https://doi.org/10.1163/9789004368811{_}003
[25] Dirk Heerwegh and Geert Loosveldt. 2008. Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly* 72, 5 (2008), 836–846. https://doi.org/10.1093/poq/nfn045
[26] Christian Hildebrand and Anouk Bergner. 2020. Conversational robo advisors as surrogates of trust: onboarding experience, firm perception, and consumer financial decision making. *Journal of the Academy of Marketing Science* (2020). https://doi.org/10.1007/s11747-020-00753-z
[27] Sebastian Hobert and Raphael Meyer Von Wolff. 2019. Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. *14th International Conference on Wirtschaftsinformatik, Siegen, Germany* (2019).
[28] Allyson L. Holbrook, Melanie C. Green, and Jon A. Krosnick. 2003. Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires. *Public Opinion Quarterly* 67, 1 (2003), 79–125. https://doi.org/10.1086/346010
[29] Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 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, 36–41. https://doi.org/10.3115/v1/p14-2007
[30] Alice Kerly, Phil Hall, and Susan Bull. 2007. Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems* 20, 2 (2007), 177–185. https://doi.org/10.1016/j.knosys.2006.11.014
[31] M. Asif Khawaja, Fang Chen, and Nadine Marcus. 2010. Using language complexity to measure cognitive load for adaptive interaction design. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. 333–336. https://doi.org/10.1145/1719970.1720024
[32] Chan Min Kim and Amy L Baylor. 2008. A virtual change agent: Motivating pre-service teachers to integrate technology in their future classrooms. *Educational Technology and Society* 11, 2 (2008), 309–321.
[33] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 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* (2019), 1–12. https://doi.org/10.1145/3290605.3300316
[34] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: A chat and voice-based conversational agent. *DIS 2018 - Proceedings of the 2018 Designing Interactive Systems Conference* (6 2018), 881–894. https://doi.org/10.1145/3196709.3196784
[35] Marios Koufaris. 2002. Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior on JSTOR. https://www.jstor.org/stable/23011056
[36] Tobias Kowatsch, Marcia Nißen, Chen-hsuan Iris Shih, Dominik Rüegger, Andreas Filler, Florian Künzler, Filipe Barata, Severin Haug, Björn Brogle, Katrin Heldt, Pauline Gindrat, Nathalie Farpour-lambert, and Dagmar Allemand. 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 (2017), 1–10.

[37] Aliane Loureiro Krassmann, Fábio Josende Paz, Clóvis Silveira, Liane Margarida Rockenbach Tarouco, and Magda Bercht. 2018. Conversational Agents in Distance Education: Comparing Mood States with Students' Perception. *Creative Education* 09, 11 (2018), 1726–1742. https://doi.org/10.4236/ce.2018.911126

[38] Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5, 3 (5 1991), 213–236. https://doi.org/10.1002/acp.2350050305

[39] Jon A. Krosnick. 1999. SURVEY RESEARCH. *Annual Review of Psychology* 50, 1 (2 1999), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

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

[41] Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 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* (2019), 0–18. https://aisel.aisnet.org/ecis2019_rp/88

[42] Seo Young Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (7 2017), 95–105. https://doi.org/10.1016/J.IJHCS.2017.02.005

[43] Raju Maharjan, Darius Adam Rohani, Per Baekgaard, Jakob E Bardram, Kevin Doherty, and Kevin 2021 Doherty. [n.d.]. Can we talk? Design Implications for theQuestionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent; Can we talk? Design Implications for theQuestionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. *CUI 2021 - 3rd Conference on Conversational User Interfaces* 11 ([n. d.]). https://doi.org/10.1145/3469595

[44] Youngme Moon. 1998. Impression management in computer-based interviews: The effects of input modality, output modality, and distance. *Public Opinion Quarterly* 62, 4 (1998), 610–622. https://doi.org/10.1086/297862

[45] Youngme Moon. 2000. Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. *Journal of Consumer Research* 26, 4 (2000), 323–339. https://doi.org/10.1086/209566

[46] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[47] Clifford Nass, Erica Robles, Charles Heenan, Hilary Bienstock, and Marissa Treinen. 2003. Speech-Based Disclosure Systems: Effects of Modality, Gender of Prompt, and Gender of User. *International Journal of Speech Technology 2003 6:2* 6, 2 (4 2003), 113–121. https://doi.org/10.1023/A:1022378312670

[48] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. ACM Press, New York, New York, USA, 72–78. https://doi.org/10.1145/191666.191703

[49] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. ACM Press, New York, New York, USA, 72–78. https://doi.org/10.1145/191666.191703

[50] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 12 (2008), 1–135. http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[51] Sabine Payr. 2003. The virtual university's faculty: An overview of educational agents. *Applied Artificial Intelligence* 17, 1 (1 2003), 1–19. https://doi.org/10.1080/713827053

[52] Stephen R. Porter. 2004. Raising response rates: What works? *New Directions for Institutional Research* 2004, 121 (1 2004), 5–21. https://doi.org/10.1002/IR.97

[53] Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.

[54] Catherine A. Roster, Robert D. Rogers, Gerald Albaum, and Darin Klein. 2004. A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research* 46, 3 (1 2004). https://doi.org/10.1177/147078530404600301

[55] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System System for Factual Knowledge. Chi (2019), 1–13. https://doi.org/10.1145/3290605.3300587

[56] Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert. 2010. Artificially intelligent conversational agents in libraries. *Library Hi Tech* 28, 4 (2010), 496–522. https://doi.org/10.1108/07378831011096196

[57] Anuschka Schmitt, Naim Zierau, Andreas Janson, and Jan Marco Leimeister. 2021. Voice as a Contemporary Frontier of Interaction Design. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*. Virtual, 1–17.

[58] Juliana Schroeder and Matthew Schroeder. 2018. Trusting in machines: How mode of interaction affects willingness to share personal information with machines. *Proceedings of the Annual Hawaii International Conference on System Sciences* 2018-Janua (2018), 472–480. https://doi.org/10.24251/HICSS.2018.061

[59] Ryan M. Schuetzler, Justin Scott Giboney, G. Mark Grimes, and Jay F. Nunamaker. 2018. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. *Decision Support Systems* 114, August (2018), 94–102. https://doi.org/10.1016/j.dss.2018.08.011

[60] Ryan M Schuetzler, G Mark Grimes, Justin Scott Giboney, and Joseph Buckman. 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* (2014).

[61] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human-agent interaction: A survey. https://doi.org/10.1145/3386867

[62] Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. *Higher Education Reports - Babson Survey Research Group*. Technical Report. http://www.onlinelearningsurvey.com/highered.htmlhttps://www.onlinelearningsurvey.com/highered.html

[63] Bayan Abu Shawar and Eric Steven Atwell. 2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics* 10, 4 (2005), 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

[64] J. Sherwani, Dong Yu, Tim Paek, Mary Czerwinski, Y. C. Ju, and Alex Acero. 2007. Voicepedia: Towards speech-based access to unstructured information. In *International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007*, Vol. 3. 2245–2248. https://doi.org/10.21437/interspeech.2007-60

[65] John Smithson, Melanie Birks, Glenn Harrison, Chenicheri Sid Nair, and Marnie Hitchins. 2015. Benchmarking for the effective use of student evaluation data. *Quality Assurance in Education* 23, 1 (2015), 20–29. https://doi.org/10.1108/QAE-12-2013-0049

[66] Donggil Song, Eun Young Oh, and Marilyn Rice. 2017. Interacting with a conversational agent system for educational purposes in online courses. *Proceedings - 2017 10th International Conference on Human System Interactions, HSI 2017* (2017), 78–82. https://doi.org/10.1109/HSI.2017.8005002

[67] Pieter Spooren, Bert Brockx, and Dimitri Mortelmans. 2013. *On the Validity of Student Evaluation of Teaching: The State of the Art*. Vol. 83. 598–642 pages. https://doi.org/10.3102/0034654313496870

[68] Carly Steyn, Clint Davies, and Adeel Sambo. 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 (2019), 11–24. https://doi.org/10.1080/02602938.2018.1466266

[69] Patrick Suppes and Mona Morningstar. 1969. Computer-assisted instruction. *Science* 166, 3903 (1969), 343–350. https://doi.org/10.1126/science.166.3903.343

[70] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings* (5 2019). https://doi.org/10.1145/3290605.3300833

[71] Karen Swan and Li Fang Shih. 2005. ON THE NATURE AND DEVELOPMENT OF SOCIAL PRESENCE IN ONLINE COURSE DISCUSSIONS. *Online Learning* 9, 3 (2005). https://doi.org/10.24059/olj.v9i3.1788

[72] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 455, 7 pages. https://doi.org/10.1145/3411763.3451623

[73] Beatrice Tucker, Sue Jones, and Leon Straker. 2008. Online student evaluation improves Course Experience Questionnaire results in a physiotherapy program. *Higher Education Research and Development* 27, 3 (2008), 281–296. https://doi.org/10.1080/07294360802259067

[74] Fang Wu Tung and Yi Shin Deng. 2006. Designing social presence in e-learning environments: Testing the effect of interactivity on children. *Interactive Learning Environments* 14, 3 (2006), 251–264. https://doi.org/10.1080/10494820600924750

[75] Hans Van Der Heijden. 2004. User acceptance of hedonic information systems. *MIS Quarterly: Management Information Systems* 28, 4 (2004), 695–704. https://doi.org/10.2307/25148660

[76] M. Vimalkumar, Sujeet Kumar Sharma, Jang Bahadur Singh, and Yogesh K. Dwivedi. 2021. 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior* 120 (7 2021), 106763. https://doi.org/10.1016/J.CHB.2021.106763

[77] Jan vom Brocke, Alexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems* 37, 1 (8 2015), 205–224. https://doi.org/10.17705/1cais.03709

[78] Joseph B Walther. 2007. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior* 23 (2007), 2538–2557. https://doi.org/10.1016/j.chb.2006.05.002

[79] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors*

in Computing Systems. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445781

[80] Thiemo Wambsganss and Christina Niklaus. 2022. Modeling Persuasive Discourse to Adaptively Support Students' Argumentative Writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8748–8760. https://doi.org/10.18653/v1/2022.acl-long.599

[81] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*. Barcelona, Spain. https://doi.org/10.18653/v1/2020.coling-main.74

[82] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting Cognitive and Emotional Empathic Writing of Students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4063–4077. https://doi.org/10.18653/v1/2021.acl-long.314

[83] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3334480.3382805

[84] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Lernaing : A Taxonomy of Conversational Agents in Education. In *Forty-Second International Conference on Information Systems*. Austin, Texas, 1–17.

[85] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376781

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

[87] Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2021. Enhancing problem-solving skills with smart personal assistant technology. *Computers and Education* 165 (2021). https://doi.org/10.1016/j.compedu.2021.104148

[88] Pawel W. Wozniak, Jakob Karolus, Florian Lang, Caroline Eckerth, Johannes Schoning, Yvonne Rogers, and Jasmin Niess. 2021. Creepy technology: What is it and how do you measure it? *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). https://doi.org/10.1145/3411764.3445299

[89] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2019. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. 27, 3 (2019). https://doi.org/10.1145/3381804

[90] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 3506–3510. https://doi.org/10.1145/3025453.3025496

[91] Naim Zierau, Christian Hildebrand, Anouk Bergner, Francesc Busquet, Anuschka Schmitt, and Jan Marco Leimeister. 2022. Voice bots on the frontline: Voice-based interfaces enhance flow-like consumer experiences & boost service outcomes. (2022). https://doi.org/10.1007/s11747-022-00868-5

Received January 2022; revised April 2022; accepted August 2022

A APPENDIX A

Test	Variable	Items	Scale
Pre-Test	personal innovativeness in the domain of information technology [1]	*"I like to experiment with new information technologies."* *"If I heard about a new information technology, I would look for ways to experiment with it."* *"In general, I am hesitant to try out new information technologies."* *"Among my peers, I am usually the first to try out new information technologies"*	1 - 5 (5: highest)
Course Evaluation	Quantitative	"*I can benefit from the content of the course.*" "*The lecturer was able to transfer the learning content according to my expectations.*" *"How satisfied were you with the format and structure of this course?.*	1 - 5 (5: highest)
Course Evaluation	Qualitative	"*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?"* *"Are there any other points you would like to comment on?"*	
Post-test	manipulation check	*With what kind of course evaluation tool did you just provide course feedback?*	Condition 1, 2, or 3
Post-test	perceived social presence [20]	*"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 human sensitivity 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."*	1 - 5 (5: highest)
Post-test	self-disclosure [7, 42]	*"The course evaluation tool and I exchanged enough personal information."* *"I expressed my thoughts when the evaluation tool asked me."* *"I can speak to the evaluation tool frankly and candidly."* *"My thoughts and feelings were conveyed frankly to the evaluation tool."* *"My answers contained information and facts."*	1 - 5 (5: highest)
Post-test	interaction enjoyability [35, 42, 75]	*"It is fun and enjoyable to share a interaction with the evaluation tool."* *"I am so absorbed in the interaction with the evaluation tool."* *"I enjoyed answering question more with this evaluation tool compared to a newer tool.",* *"The interaction with the evaluation tool is exciting.".* *"Services provided by the evaluation tool are more entertaining and attractive than without this tool."*	1 - 5 (5: highest)
Post-test	attention check	*Please select "strongly agree".*	1 - 5
Post-test	qualitative impression	*"What did you particularly like about the use of the course evaluation tool?"* *"What else could be improved?"* *"Do you have any other ideas?"*	open
Post-test	pre-experience	*"Have you used a chatbot before (e.g., a Facebook Messenger Bot)?"*	Yes/No
Post-test	demographics	1. Age 2. Gender	open

Table 8. Overview of measured items in the pre-and post-test as well as in the qualitative and quantitative course evaluation.