Please quote as: Wambsganss, T., Winkler, R., Schmid, P., and Söllner, M. (2020) Designing a Conversational Agent as a Formative Course Evaluation Tool. In 15th International Conference on Wirtschaftsinformatik (WI), pp 1234-1249.

Designing a Conversational Agent as a Formative Course Evaluation Tool

Thiemo Wambsganss[1], Rainer Winkler[1], Pascale Schmid[1], Matthias Söllner[1,2]

[1] University of St.Gallen (HSG), Institute of Information Management, St.Gallen, Switzerland
{thiemo.wambsganss, rainer.winkler, matthias.soellner}@unisg.ch,
pascale.schmid@student.unisg.ch
[2] University of Kassel, Information Systems and Systems Engineering, Kassel, Germany,
soellner@uni-kassel.de

Abstract. Today's graduating students face ever-changing environments when they enter their job life. Educational institutions must therefore continuously develop their course structure and content in order to prepare their students to be future employees. A very important means for developing the courses is the students' course evaluations. Due to financial and organizational restrictions, these course evaluations are usually carried out quantitatively and at the end of the semester. However, past research has shown that this kind of evaluation faces certain constraints such as low acceptance rates, only time-related insights and low-quality answers that do not really help the lecturer to improve the course. Drawing on social response theory, we propose that conversational agents as a formative course evaluation tool are able to address the mentioned problems by interactively engaging with students. Therefore, we propose a set of design principles and evaluate them with our prototype Eva.

Keywords: Conversational agents, formative course evaluation, design science research, human-computer interaction

1 Introduction

Students graduating today face increasingly volatile, uncertain, complex and ambiguous (VUCA) business environments. Job profiles in demand change constantly [1]. Educational institutions, such as universities, are therefore required to continuously adapt their courses in order to prepare the students for the job life afterwards. However, lecturers often struggle to anticipate what students expect from a course, how they perceive the learning content and how their needs can be effectively addressed [2–4]. To tackle this challenge, educational institutions commonly use quantitative, standardized, online or paper-based surveys, in which students are encouraged to share their insights in a given form. Most of the time, these surveys reach the students at the end of a course or, at best, once in the middle of the semester [5].

However, past research has shown that this kind of feedback faces certain constraints, such as low acceptance rates, only time-related insights and low-quality answers that are hardly usable for adapting the course to students' expectations [6]. One

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

explanation for these negative results might be that student responses are affected by survey fatigue resulting from repeated requests at the end of the semester [6]. Students feel frustrated because lecturers often miss to react appropriately to evaluations [5] and lecturers often adapt courses only for the next cohort of students [7]. To address those issues, qualitative evaluation methods, such as interviews, are used to produce a higher quality of answers and to gain deeper insights. However, these approaches are usually very resource-intensive since lecturers need to address every student individually.

One possible solution to benefit from the advantages of both – qualitative and quantitative – evaluation methods is using conversational agents (CAs). CAs are software programs which communicate with users through natural language interaction interfaces [8, 9]. Compared to traditional quantitative course evaluations, CAs are able to reach students on their everyday devices and build up a human-like interaction with them. CAs are able to adapt their answers to students' utterances and can therefore build up a meaningful dialogue with the students almost like a qualitative lecturer-student interview. Backing on social response theory [10–12] this form of human-computer interaction might encourage students to provide a higher quality of answers for lecturers to improve their courses.

The popularity of CAs, such as Amazon's Alexa, Google's Assistant, Apple's Siri and other systems, has been steadily growing over the past few years [13, 14]. The recent improvement in Natural Language Processing (NLP), Natural Language Understanding (NLU) and Machine Learning (ML) enables CA systems to ask and answer questions in natural conversation flows and use intelligent question answering to adapt to a certain task [15]. In education, CAs have been used for several purposes, such as to provide support for problem solving in mathematics [16], to mediate group learning processes during problem solving [17], for collaborative language learning [18] or for academic advising [19]. Existing research on CAs in education has mainly focused on providing learning support for students [20]. Research on CA support for lecturers is still scarce. Moreover, Winkler and Söllner [21] emphasize that CAs might also have great potential as an evaluation tool for lecturers. In a recent study Kim et al. [22] compared a CA against online surveys in an experiment and found that participants using a chatbot were providing more differentiated responses and, thus, the CA survey resulted in higher-quality data compared to the online survey. However, transferable insights and design knowledge on how to build CAs as a formative evaluation tool for lecturers is still missing. Hence, in this paper we seek to answer the following research question:

***RQ**: How to design a conversational agent as a formative course evaluation tool?*

To answer our research question, we follow the design science research approach (DSR) of Hevner et al. [23]. Drawing on social response theory, we propose that CAs might be better able to be used as evaluation tool compared to existing solutions by changing the way how computer systems are interacting with students. We argue that a CA might be able to address the current challenges of quantitative and qualitative course evaluation methods by combining the advantages of both. With a formative course evaluation tool, we implicate a tool which enables students and lecturers, to provide and receive continuous and on-going feedback during a course.

First, we define the problem and gather requirements from practice and literature. Second, we propose design principles and instantiate and evaluate our CA *Eva* (*Evaluation Agent*). Based on the insights of our first evaluation, we create our second version of *Eva* and evaluate it in a real learning environment.

The remainder of our paper is structured as follows: In section 2, we describe different types of CAs in education and the current state of literature on course evaluations in education. In section 3, we explain how we proceed to develop our CA *Eva*. In section 4, we design and evaluate *Eva*. Finally, we close with a discussion, limitations and the contributions of our study.

2 Theoretical Background

2.1 Conversational Agents in Education

CAs are software programs which communicate with users through natural language interaction interfaces [8, 9]. In modern society, CAs have become ubiquitous and have been implemented in various areas, such as e-commerce [24], entertainment [24] or the health sector [25]. CAs used in education can be defined as a special form of learning application that interacts with learners individually [15]. The development of CAs in education can be traced back to the 1970s research stream of Intelligent Tutoring Systems (ITS). An ITS exhibits characteristics similar to a human tutor such that it may be able to answer student questions, detect misconceptions and provide feedback. While the original ITS were abstract entities with limited technological possibilities, the next three decades saw advances in agent representation (i.e., visual embodiment) and interactive capabilities. Over the years, ITS are able to interact with learners using multiple channels of communication (e.g., text and speech) and are able to exhibit social skills and intelligence by communicating with users on a broad range of issues and expand its scope in terms of roles. Such roles include tutors, coaches and actors [26]; experts, motivators and mentors; learning companions [27, 28], change agents [29]; and lifelong learning partners [30].

Until now, CA research mainly focused on how to improve students' learning outcomes by offering them individual support [15]. However, there is little research on how lecturers can use CAs as a formative course evaluation tool to improve the learning content. The question remains how CAs can be successfully used and designed as a formative course evaluation tool.

2.2 Course Evaluation in Education

Course evaluation can be differentiated into quantitative and qualitative evaluation methods. The most frequent form of course evaluation is paper-based or online questionnaires used for quantitative analysis [31]. However, quantitative course evaluations have been broadly criticized. The reliability and validity of student feedback surveys is questioned [4]. Another concern involves the use of evaluations because they have become a mere ritual to be followed despite poor survey results and have a limited capability to contribute to course improvements [32]. The motivation of

students to participate in course evaluations is often low because they see no benefit [31]. According to Chen and Hoshower [33], students can be motivated to participate in course evaluations by informing them about the intension of the evaluation and showing the implementation of the feedback provided. The alternative approach, i.e., qualitative course evaluation, is rather scarce in literature [34]. A recent study by Steyn et al. [34] shows that qualitative course evaluation has the potential to overcome the disadvantages of quantitative course evaluations. Qualitative course evaluations offer students the possibility to provide individual, open feedback, which increases the feedback quality [34]. Moreover, depending on the type of qualitative course evaluation (e.g., class discussion), lecturers have the possibility to ask questions and, thus, receive better and more extensive feedback. However, qualitative course evaluations are not scalable, take longer to analyze and are resource-intensive, resulting in a deterrent effect [34].

CAs could have the potential to overcome the disadvantages of traditional quantitative and qualitative evaluations. Compared to quantitative evaluations, CAs can adapt their questions to the answers of the students and, thus, allows to gather richer feedback similar to qualitative course evaluation methods. Compared to qualitative evaluations, CAs are available 24/7 and save personal resources (e.g., lecturers do not have to use the lecture time or additional time for qualitative feedback). Moreover, data gathered by CAs can be stored in a database and can be analyzed further by information extraction algorithms (e.g., through the classification of answers to categories, topic modelling or sentiment analysis).

2.3 Kernel Theory: Social Response Theory

We built our research endeavor on social response theory. We believe that this theory supports our underlying hypothesis that CAs can improve the quality and acceptance of course evaluations in education. Moon [11] found that humans tend to respond socially to agents that display characteristics similar to humans (e.g., to animals or technologies). When people experience any human-like characteristics in any form of communication, their evolutionary behavior subconsciously applies social rules to their interaction [10]. Nass et al. [12] argue that the underlying reasons for this human behavior are that social cues from computers trigger subconscious responses from humans, no matter in which rudimentary form they occur. Seeing computers as social actors, researchers have investigated how different social cues impact human-computer interaction, e.g., language style [12] or response time [35].

We argue that social response theory can explain why CAs might be able to better imitate an individual lecturer-student interaction compared to standardized, quantitative surveys, which might result in a higher quality of students' evaluation input [4].

3 Research Methodology

Our study follows the three cycle view of Hevner [36]. This approach allows us to design an artifact (design cycle) that solves a set of practical problems that researchers

and practitioners experience in their own practice (relevance cycle) and to contribute to the existing body of knowledge (rigor cycle). Figure 1 shows the steps we carry out.

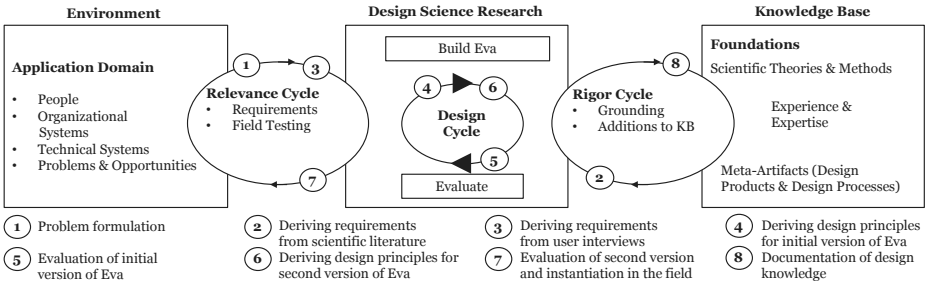

Figure 1. Three cycle design science process according to Hevner [36]

The first step of the DSR cycle includes the formulation of the problem. The relevance of the practical problem was therefore described in the introduction of this work. In the second step, we start the rigor cycle by deriving a set of meta-requirements (MRs) from scientific literature for the design of a CA for formative course evaluation. In the third step, we conducted eleven semi-structured interviews with students and five interviews with lecturers using the expert interview method by Gläser and Laudel [37]. Based on the interviews, we gathered user stories (USs) and defined user requirements (URs) for the design of a CA as a formative course evaluation tool following the method of Cohn [38]. In the fourth step, we derived preliminary design principles (DPs) addressing the MRs and URs from the prior steps and designed an initial version as a first instantiation of these DPs. In the fifth step, we conducted a proof-of-concept evaluation based on evaluation criteria proposed by Venable et al. [39]. Based on the design principles, we created a mock-up prototype called *Eva* (short for evaluation agent), where lecturers and students were able to have a simple interaction with a CA. Subsequently, we interviewed them to capture their perception of *Eva*. The goal of this evaluation was to see how lecturers and students perceive the value of our instantiated design principles, to note change requests and to gather additional design principles. Then in step 6, we refined our design principles based on the findings from this evaluation before designing a second version of *Eva*, which then was tested in a real learning environment with one lecturer and twelve students (step 7). The goal of this evaluation was a proof-of-usefulness proposed by Sonnenberg and vom Brocke [40]. Here our aim was to see whether the design principles and the resulting CA are useful in a real-life setting. In step 8, we close with a short discussion thereby documenting the design knowledge.

4 Design and Evaluation of *Eva – an Evaluation Agent*

In this section, we describe and discuss how we gathered the requirements and derived the DPs. The problem formulation (step 1) described in the introduction serves as the foundation for the derivation of the requirements from literature and users. An overview

of the practical and theoretical requirements as well as the derived design principles is illustrated in Figure 2.

4.1 Step 2: Rigor - Deriving Requirements from Scientific Literature

We initiate the rigor cycle by gathering requirements from theory. We conducted a systematic literature review following established methodical approaches from Cooper [41] and vom Brocke et al. [42]. Based on that, we (1) defined the review scope, (2) conceptualized the topic, (3) searched the literature, and (4) analyzed the findings regarding requirements. Regarding step 1 (define the review scope), we primarily focused our literature review on research outcomes that show successfully implemented CAs. Furthermore, our goal is to identify requirements on a conceptual level with a focus on an espousal of position and a representative coverage [41]. Regarding step 2 (conceptualization of the topic), we identified two broad areas for deriving requirements: *Technology* and *Education*. We focused on these two areas because evaluating a course is a complex phenomenon being investigated through different lenses by psychologists, educationists and computer scientists [43]. Regarding step 3 (literature search), we conducted a keyword search on Google Scholar to identify relevant publications. We used Google Scholar because this web search engine enables advanced full-text search and several filter options for academic literature. We decided to use the following search string: ("*pedagogical conversational agent*" OR "*conversational agent*" OR "*chatbot*") AND ("*student evaluation*" OR "*course evaluation*" OR "*online course evaluation*" OR "*assessment*" OR "*student feedback*"). In total, we obtained 6'090 articles. We defined criteria for inclusion and exclusion and reviewed titles and abstracts of our search results in a first step. We only included papers that address some kind of evaluation in higher education and the use of conversational agents in order to define successfully derived requirements. Several papers were excluded because they address different research areas of CAs, such as health care or customer service. Furthermore, we excluded papers that mainly focus on other educational fields such as using CAs as learning tutors. Based on that, we selected 43 papers. Regarding step 4 (literature analysis), we clustered similar issues together (LIs), resulting in four clusters. The clusters including exemplary papers and meta-requirements are depicted in Figure 2.

The first meta-requirement (MR1) derived from theory raises the importance for including an assessment of the individual teaching quality to increase the acceptance rates of students (e.g.,[4]). As literature shows, students show improved response rates when course evaluations are conducted online and contain open questions [7]. Therefore, the second meta-requirement (MR2) expresses the need for an online assessment format with open questions. To enhance student perception and motivation for the assessment of the effectiveness of a course, the third meta-requirement (MR3) focuses on information on the intension and implementation of the course evaluation (e.g., [33]). The fourth meta-requirement (MR4) deals with the need for an individualized CA to be easy accessible, available twenty-four hours a day, and speeds up response times to enhance user satisfaction (e.g.,[21]).

4.2 Step 3: Relevance - Deriving Requirements from Expert Interviews

To gain a wider picture of the requirements of a formative course evaluation system, we interviewed different user groups, including students as lead users and lecturers as users of the results systems. Based on the derived LIs and MRs, we conducted eleven semi-structured interviews with students and five interviews with lecturers according to Gläser and Laudel [37]. We chose these two user groups in order to get a holistic picture about the needs of CAs as a formative evaluation tool. The interview guideline consisted of questions regarding the following topics: perception of actual online and paper-based course evaluation (e.g., advantages, disadvantages), motivation for evaluation participation (e.g., benefits, influence), requirements concerning an evaluation tool (e.g., interface, functionalities), use of a CA for course evaluations (e.g., application scenarios). Each interview lasted around 15 to 30 minutes. The eleven student interviewees were chosen out of a random subset of the population of students at our university. Nine students were male and two were female, all aged between 21 and 27. The interviewee group consisted of nine master and two bachelor students, all majoring in Business Studies. Additionally, we conducted five lecturer interviews. The interviewees were professors teaching on a regular basis at our university on bachelor and master levels, all aged between 27 and 50. Based on the interview results, we gathered user stories from students (USSs) and user stories from lecturers (USLs) and herewith identified user requirements (URs) following Cohn [38]. The first user story of students (USS1) highlights the perceived lack of student influence on a university course during the semester. The students want to provide feedback during the semester to improve their education experience while still enrolled in the course (UR1). The second user story of students (USS2) describes the request for a responsive and convenient user interface of the evaluation tool. Hence, the second user requirement (UR2) demands a responsive and lean user interface which can be adapted for the specific course content. The interviews showed that students complained about missing or delayed survey results and about their perception of not being taken seriously by the lecturer (USS3). Therefore, the third user requirement (UR3) indicates that evaluation results and course adjustments should be shared with the students as soon as possible after the evaluation. The fourth user requirement (UR4) represents the need for a time specification for course evaluations as students want to see the progress of their evaluation (USS4). The first user story of lecturers (USL1) addresses the fact that the lecturers would like to get a clear analysis of the conducted course evaluation. This allows them to benefit from valuable insights to improve their courses. Hence, from a user perspective the evaluation tool needs an intelligent analysis function which can filter and display the important insights (UR5). To give and evaluate feedback efficiently and effectively, lecturers demand a user-friendly and adaptive evaluation tool. Therefore, the second user requirement (UR2) demands a responsive and lean user interface which can be adapted for the specific course content, as mentioned above. The last user story (USL3) deals with the fact that educational institutions have to comply with data protection regulations. Also, user anonymity has to be guaranteed. Thus, the resulting user requirement (UR6) states that the evaluation tool has to be compliant with

data protection regulations and has to allow anonymous evaluation.

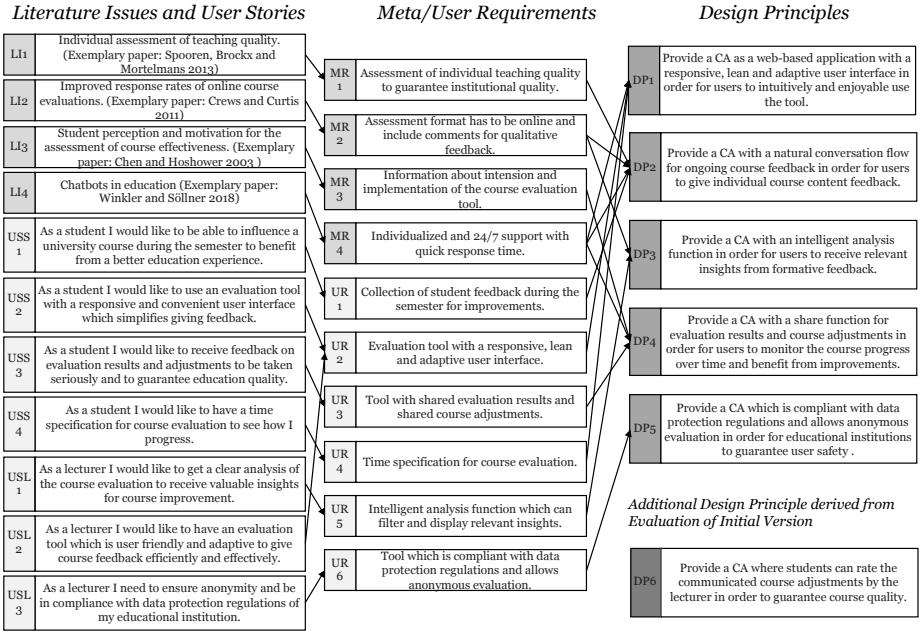

Figure 2. Overview of the derived design principles according to Chandra et al. [44]

4.3 Step 4: Deriving Design Principles for an Initial Version of Eva

Based on our identified meta- and user-requirements, we established an initial set of design principles as shown in Figure 2. The first design principle (DP1) specifies that the CA is a web-based application tool with a responsive, lean and adaptive user interface. The web-based application tool has to be accessible for devices that students commonly use. With the frequent use of web-based devices today, students have access to the CA at any time and at any place. The second design principle (DP2) contains a CA for ongoing course feedback. As described above, the increasing number of students and the resource constraints do not allow a personal dialog between students and lecturers. Course insights therefore have to be gathered differently. According to social response theory, CAs have the capabilities to provide valuable formative feedback when interacting in a natural conversation flow. The format of a CA enables higher educational institutions to collect comments for qualitative feedback. Conversation flows between users and CAs can take place anytime and individually. In addition, assessments can be conducted with a CA during the semester and allow a quick response time. Lecturers are able to adapt their courses according to the given feedback and therefore students can benefit from improvements while still enrolled in the course. The use of a CA makes it possible to address this important issue of the students. The third design principle (DP3) describes the need for an intelligent analysis function in order to deliver relevant insights from formative feedback. Because a CA

can gather qualitative feedback in large amounts, an intelligent analysis function must process this information. The design principle four (DP4) demands a share function for evaluation results and course adjustments. As mentioned above, sharing information of the course evaluation motivates students to participate and increases the probability of them providing meaningful feedback in further surveys [33]. The fifth design principle (DP5) includes the data protection regulations and the requirement of anonymous course evaluations. Data protection regulations are determined externally and must be applied accordingly. The tool must guarantee educational institutions their data sovereignty and must ensure immediate adaptations to regulatory changes.

To instantiate and evaluate the design principles above, we created a mock-up-based prototype called *Eva* with design features derived from our design principles (see Table 1). The prototype *Eva* allows lecturers to send evaluation questions to the students where they are able to insert responses.

Table 1. Instantiation of design principles with design features

Design Features of the Initial Version of Eva		**Implemented Design Principles**				
		DP1	**DP2**	**DP3**	**DP4**	**DP5**
DF1	Link or QR code to access course evaluation on a web-enabled device	X				
DF2	User authentication via active directory using single sign-on	X				
DF3	Simple overview of functions with appealing design (e.g., pictures, color, icons)	X				
DF4	Onboarding message	X	X			
DF5	Course specific questions	X	X			
DF6	Human-like language		X			
DF7	Analysis overview (e.g., sentiment analysis, percentages, pie charts)			X		
DF8	Filter function for the evaluation results			X		
DF9	Function “Share with students”				X	
DF10	Function “Defined Actions”				X	
DF11	Data storage on an internal university server					X
DF12	Notification of anonymous data handling					X

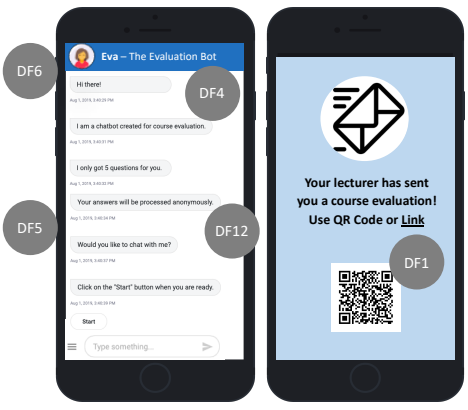

Figure 3. CA mock-up of *Eva* on a mobile phone with certain design features

The initial version of *Eva* was created with the platform *Snatchbot*[1]. Figure 3 and 4 show the designed mock-up prototype for different devices (smartphone, tablet and notebook). *Eva* can be accessed via a link or a QR code (DF1). It delivers an onboarding message (DF4) and uses human-like language (DF6) to guarantee a natural conversation flow. The mock-up prototype contains course-specific questions (DF5) to receive relevant insights from formative feedback. In order to inform students of anonymous data handling, the CA sends a notification message (DF12).

Figure 4. CA mock-up of *Eva* on different devices with certain design features

Eva allows the lecturers to set up their individual questions (DF5) for the course evaluation with the CA. Figure 3 and 4 show the designed user interface of the evaluation tool. To make *Eva* convenient, the mock-up contains a user authentication via active directory using a single sign-on functionality (DF2, not shown in Figure 3 or 4). With this functionality, students as well as lecturers can access *Eva* using their university login. The user interface was designed as an easy-to-use and intuitive mock-up. Furthermore, a scalable user interface allows to consider specific course contents, which guarantees efficient and effective evaluation. An appealing design with pictures, color and icons as well as a simple overview of the offered functions (DF3) supports an intuitive and enjoyable use of the tool. An analysis of the evaluation results gathered from *Eva* is presented in percentages, pie charts, diagrams, sentiment analysis or text form (DF7). Additionally, the mock-up contains a filter function for the evaluation results (DF8) and the function "*Share with students*" (DF9). The function "*Defined Actions*" (DF10) allows the lecturer to inform the students about actions taken based on the insights from the course evaluation. To be compliant with data protection regulations, data has to be stored on an internal university server (DF11).

4.4 Step 5: Proof-of-Concept Evaluation of Initial Version of Eva

In this section, we describe the proof-of-concept evaluation of the initial version of *Eva*. The evaluation serves to verify if the design principles are of value to the lecturers and

[1] https://snatchbot.me/

students and to identify change requests and additional design principles. For this purpose, an online course evaluation invitation was sent to 266 students via a personal messenger. The link was forwarded to bachelor and master students studying Business Management, Business Innovation, Accounting and Finance. As the evaluation was conducted anonymously, no further participant characteristics are available. When the students clicked on the link, the CA *Eva* appeared on their device similar to Figure 3, left picture. After having a short interaction with *Eva*, the students were put into the scenario of a course evaluation. Then, the students were asked six questions regarding the use of the designed CA they were interacting with, its features and possible change requests. Out of the 266 students contacted, we got responses from 28 students from our university. Table 2 shows the consolidated results of the questions.

Table 2. Evaluation results of questions from proof-of-concept user survey

I would like to use a CA for lecture evaluation.	**21** Yes	**1** No	**6** N/A	
I would appreciate to influence a course with my personal feedback.	**23** Yes	**2** I don't care	**0** No	**3** N/A
I would appreciate to give feedback anytime during the semester.	**19** Yes	**2** I don't care	**4** No	**3** N/A
I would appreciate getting instant evaluation results.	**22** Yes	**2** I don't care	**1** No	**3** N/A
My preferred question types are…	**10** Predefined	**1** Open	**11** Combination	**6** N/A
For me the ideal evaluation duration would be…	**16** 2-5 min	**4** 5-10 min	**2** 10-15 min	**6** N/A
I would like to evaluate courses in group chats.	**5** Yes	**18** No	**5** N/A	
How would you rank a CA versus a traditional online evaluation?	**12** Positive	**7** Neutral	**7** Critical	**2** N/A

The results of the student questions revealed the usefulness of our design principles DP1, DP2 and DP4. 21 out of 28 students liked using *Eva* for evaluation. 23 out of 28 students would appreciate being able to influence a university course with their feedback. 19 out of 28 students would appreciate giving feedback anytime during the semester. These results support the design principles DP1 and DP2. The open question regarding their preference to use *Eva* instead of the traditional online evaluation led to 12 positive comments supporting DP2. 22 students stated that they would appreciate getting instant evaluation results, which is described in DP4. In addition, the preferred question types and ideal evaluation duration were questioned in order to design the artifact. The idea of adding a group chat feature was rejected based on the survey results. In order to also have the design principles assessed by lecturers, we contacted different university professors via mail. The mail contained a link which directed the professors to the CA *Eva*. After having a short interaction with *Eva*, we gathered qualitative feedback via the CA and asked questions regarding its features and possible change requests. Seven professors from our university responded anonymously with detailed information. The lecturer feedback showed that they liked the idea of using CAs as a formative course evaluation tool because of the opportunity to ask individualized and course-specific questions similar to qualitative evaluation methods. Moreover, the use of CAs allows a more comprehensive analysis of the evaluation results compared to qualitative evaluation methods, such as interviews or class

discussions. However, the lecturers stated that they want to use only one evaluation tool at the educational institution. Thus, the CA should not be implemented as an additional tool to the existing course evaluation tool. Finally, we compared student and lecturer responses with the existing design principles. The findings revealed that both students and lecturers miss the possibility to rate the course adjustments based on the preceding course evaluation and to communicate whether lecturers were able to implement student requests correctly. In the following, the design principle derived from this requirement and the additionally developed design features will be discussed further.

4.5 Step 6: Validating and Deriving Design Principles for Second Version of *Eva*

In this section we refined our design principles based on the findings from the evaluation of the initial version. Based on the results of step 5, the design principles DP1-DP5 were validated and a new design principle was derived. The new design principle (DP6) specifies that the evaluation tool should allow students to rate the communicated course adjustments by the lecturer in order to guarantee course quality and close the feedback loop. DP6 is shown in Figure 2 as an additional design principle. We developed a design feature to instantiate and evaluate DP6 by creating the function "*Students Feedback*" (DF13) for the defined actions. For the second version of *Eva*, we used the existing design features presented in Table 1 and added the feature "*Students Feedback*". This feature allows students to provide feedback on the course adjustments of the lecturer. The second version of *Eva* was also created on the platform *Snatchbot* by adjusting the initial version.

4.6 Step 7: Proof-of-Usefulness Evaluation with Second Version of *Eva*

In a next step, we tested the second version of *Eva* in a real learning environment with one lecturer and twelve students. According to Sonnenberg and vom Brocke [40], it is important to evaluate the proof-of-usefulness of an artifact. Our aim was to show whether the design principles and the resulting CA are of value in a real-life setting. The evaluation of the second version of *Eva* took place in a university course, a didactic course for prospective business educators. The participants consisted of twelve business education students, five male and seven female participants, who were between the ages of 23 and 28. Out of twelve business education students, ten were enrolled in a master course and two were doing postgraduate studies. A small course was chosen as this allowed us to observe the participants while using the CA and hence gather deeper insights. The lecturer sent a link via mail to his students to open *Eva*. During a lecture, students were then asked to complete the CA course evaluation and to answer questions about the used CA. The evaluation took between 10 and 15 minutes.

Table 3. Survey results of the proof-of-usefulness evaluation in a university course

Would more likely fill out a course evaluation if the survey format was a CA	8/12
Liked the CA more than the used standard online course evaluation format (EvaSys)	10/12
Liked the reaction intensity of the CA reflected in the individual follow-up questions	8/12

The findings of the evaluation with *Eva* showed that 8 out of 12 students would more likely fill out a survey if the tool was a CA. 10 out of 12 students preferred *Eva* more than the standard survey tool used at their university. 8 out of 12 students liked the reaction intensity of *Eva* reflected in the individual follow-up questions (see Table 3). After the course evaluation, the lecturer was interviewed about the presentation of the results in the CA mock-up (see Figure 4). The lecturer appreciated the analysis overview of *Eva* with, e.g., percentages and pie charts, which reduces his effort and provides important information in a short time. Furthermore, the lecturer liked sharing the evaluation results with his students and that he gets informed about how his course adaptations are perceived by the students.

5 Discussion and Conclusion

In this paper, we report on the development of design principles for CAs as a formative course evaluation tool. Our work makes several contributions to research. First, we show how CAs can be used to build up a dialogue with students in order to increase evaluation quality. This kind of dialogue was previously only possible with qualitative evaluation methods conducted by humans. Thus, our work contributes to a better understanding of how computer systems can imitate lecturers in the area of course evaluation. Second, to the best of our knowledge, this study is the first one that creates design knowledge on how to use CAs as formative course evaluation tools. Our DPs were formulated based on social response theory. We argue that a course evaluation that instantiates our DPs might be able to better imitate an individual lecturer-student interaction compared to standardized, quantitative surveys. This might result in a higher quality of students' evaluation input. Finally, we provide a stronger basis for researchers to report on alternative course evaluation designs to compare and contrast them with our solution. Our work also has several implications for practice. First, we argue that CAs represent a better course evaluation method compared to existing quantitative and qualitative evaluation methods and we were able to prove its usefulness. The extent to which this is indeed the case should be the aim of future research. Second, lecturers and educational institutions can now use these design principles to create their own CAs. A number of limitations have to be considered with respect to our study. First, we gathered requirements from a certain theoretical perspective (social response theory) and a specific user group. It might be possible that other areas of literature and user groups might have led to different results. Moreover, we were not yet able to evaluate our CA *Eva* in a large-scale lecture during a whole semester. This would have given us further insights about the long-term usage of CAs as formative evaluation tools and would help us to evaluate if lecturers are triggered to adjust their course within a semester. Additionally, we did not address the design of social cues and the level of anthropomorphism of our CA in the evaluation. Thus, we not only call for future research to evaluate CAs in large-scale lectures, but also on more research on how to design the level of anthropomorphism of a CA as a course evaluation tool.

References

1. vom Brocke, J., Maaß, W., Buxmann, P., Maedche, A., Leimeister, J.M., Pecht, G.: Future Work and Enterprise Systems. Bus. Inf. Syst. Eng. 60, 357–366 (2018).
2. Smithson, J., Birks, M., Harrison, G., Sid Nair, C., Hitchins, M.: Benchmarking for the effective use of student evaluation data. Qual. Assur. Educ. 23, 20–29 (2015).
3. Blair, E., Valdez Noel, K.: Improving higher education practice through student evaluation systems: Is the student voice being heard? Assess. Eval. High. Educ. 39, 879–894 (2014).
4. Spooren, P., Brockx, B., Mortelmans, D.: On the Validity of Student Evaluation of Teaching. (2013).
5. Shah, M., Cheng, M., Fitzgerald, R.: Closing the loop on student feedback: the case of Australian and Scottish universities. High. Educ. 74, 115–129 (2017).
6. Tucker, B., Jones, S., Straker, L.: Online student evaluation improves Course Experience Questionnaire results in a physiotherapy program. High. Edu. Res. Dev. 27, 281–296 (2008).
7. Crews, T.B., Curtis, D.F.: Online Course Evaluations: Faculty Perspective and Strategies for Improved Response Rates. Assess. Eval. High. Educ. 36, 865–878 (2011).
8. Shawar, B.A., Atwell, E.S.: Using corpora in machine-learning chatbot systems. Int. J. Corpus Linguist. 10, 489–516 (2005).
9. Rubin, V.L., Chen, Y., Thorimbert, L.M.: Artificially intelligent conversational agents in libraries. Libr. Hi Tech. 28, 496–522 (2010).
10. Nass, C., Moon, Y.: Machines and Mindlessness: Social Responses to Computers. (2000).
11. Moon, Y.: Intimate Exchanges: Using Computers to Elicit Self-Disclosure From Consumers. J. Consum. Res. 26, 323–339 (2000).
12. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94. pp. 72–78. ACM Press, New York, New York, USA (1994).
13. Krassmann, A.L., Paz, F.J., Silveira, C., Tarouco, L.M.R., Bercht, M.: Conversational Agents in Distance Education: Comparing Mood States with Students' Perception. Creat. Educ. 09, 1726–1742 (2018).
14. eMarketer: Alexa, Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130%. (2017).
15. Hobert, S., Wolff, R.M. Von: Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. (2019).
16. Aguiar, E.V.B., Tarouco, L.M.R., Reategui, E.: Supporting problem-solving in Mathematics with a conversational agent capable of representing gifted students' knowledge. Proc. Annu. Hawaii Int. Conf. Syst. Sci. 130–137 (2014).
17. Winkler, R., Söllner, M., Neuweiler, M.L., Rossini, F.C., Leimeister, J.M.: Alexa, Can You Help Us Solve This Problem? How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes. (2019).
18. Tegos, S., Demetriadis, S., Tsiatsos, T.: A configurable conversational agent to trigger students' productive dialogue: A pilot study in the CALL domain. (2014).
19. Latorre-Navarro, E., Harris, J.: An Intelligent Natural Language Conversational System for Academic Advising. Int. J. Adv. Comput. Sci. Appl. 6, (2015).
20. Song, D., Oh, E.Y., Rice, M.: Interacting with a conversational agent system for educational purposes in online courses. Proc. - 2017 10th Int. Conf. Hum. Syst. Interact. 78–82 (2017).
21. Winkler, R., Söllner, M.: Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis . In : Academy of Management. Meet. Annu. Chicago, A O M. (2018).
22. Kim, S., Lee, J., Gweon, G.: Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. Conf. Hum. Factors Comput. Syst. - Proc. 1–12 (2019).

23. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. Des. Sci. IS Res. MIS Q. 28, 75 (2004).
24. Kerly, A., Hall, P., Bull, S.: Bringing chatbots into education: Towards natural language negotiation of open learner models. Knowledge-Based Syst. super20, 177–185 (2007).
25. Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N.R., Rajeswar, S., de Brebisson, A., Sotelo, J.M.R., Suhubdy, D., Michalski, V., Nguyen, A., Pineau, J., Bengio, Y.: A Deep Reinforcement Learning Chatbot. 1–40 (2018).
26. Payr, S.: The virtual university's faculty: An overview of educational agents. Appl. Artif. Intell. 17, 1–19 (2003).
27. Kim, Y., Baylor, A.L., Shen, E.: Pedagogical agents as learning companions: The impact of agent emotion and gender. J. Comput. Assist. Learn. 23, 220–234 (2007).
28. Kim, Y., Baylor, A.L.: Research-Based Design of Pedagogical Agent Roles: a Review, Progress, and Recommendations. Int. J. Artif. Intell. Educ. 26, 160–169 (2016).
29. Kim, C.M., Baylor, A.L.: A virtual change agent: Motivating pre-service teachers to integrate technology in their future classrooms. Educ. Technol. Soc. 11, 309–321 (2008).
30. Chou, C.-Y., Huang, B.-H., Lin, C.-J.: Complementary machine intelligence and human intelligence in virtual teaching assistant for tutoring program tracing. Comput. Educ. 57, 2303–2312 (2011).
31. Erikson, M., Erikson, M.G., Punzi, E.: Student responses to a reflexive course evaluation. Reflective Pract. 17, 663–675 (2016).
32. Freeman, R., Dobbins, K.: Are we serious about enhancing courses? Using the principles of assessment for learning to enhance course evaluation. As. Eval. Hi. Edu. 38,142–151 (2013).
33. Chen, Y., Hoshower, L.B.: Student evaluation of teaching effectiveness: An assessment of student perception and motivation. Assess. Eval. High. Educ. 28, 71–88 (2003).
34. Steyn, C., Davies, C., Sambo, A.: Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students. Assess. Eval. High. Educ. 44, 11–24 (2019).
35. Gnewuch, U., Morana, S., Adam, M., Maedche, A.: Faster is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. Res. Pap. (2018).
36. Hevner, A.R.: A three cycle view of design science research. Scand. J. Inf. Syst. 1–6 (2007).
37. Gläser, J., Laudel, G.: Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen. VS Verlag für Sozialwiss (2010).
38. Cohn, M.: User Stories Applied For Agile Software Development. (2004).
39. Venable, J., Pries-Heje, J., Baskerville, R.: FEDS: A Framework for Evaluation in Design Science Research. Eur. J. Inf. Syst. 25, 77–89 (2016).
40. Sonnenberg, C., vom Brocke, J.: Evaluation Patterns for Design Science Research Artefacts. Presented at the October 14 (2012).
41. Cooper, H.M.: Organizing knowledge syntheses: A taxonomy of literature reviews. (1988).
42. vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., Cleven, A.: Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. Commun. Assoc. Inf. Syst. 37, (2015).
43. Lajoie, S.P., Azevedo, R.: Teaching and Learning in Technology-Rich Environments. In: Handbook of educational psychology. pp. 803–821. (2006).
44. Chandra, L., Seidel, S., Gregor, S.: Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. HICSS (2015).