How Are You, Chatbot? Evaluating Chatbots in Educational Settings – Results of a Literature Review

Sebastian Hobert[1]

Abstract: Evaluation studies are essential for determining the utilization of technology-enhanced learning systems. Prior research often focuses on evaluating specific factors like the technology adoption or usability aspects. However, it needs to be questioned if evaluating only specific factors is appropriate in each case. The aim of this research paper is to outline which methods are suited for evaluating technology-enhanced learning systems in interdisciplinary research domains. Specifically, we focus our analysis on pedagogical conversational agents – i.e. learning systems that interact with learners using natural language. For instance, in addition to technology acceptance, further factors like learning success are more important in this case. Based on this assumption, we analyze the current state-of-the-art literature of pedagogical conversational agents to identify evaluation objectives, procedures and measuring instruments. Afterward, we use the results to propose a guideline for evaluations of pedagogical conversational agents.

Keywords: chatbots, pedagogical conversational agents, technology-enhanced learning, evaluation

1 Introduction

The rising of intelligent systems based on methods known from machine learning and artificial intelligence affects information systems in many areas. This is especially true for the domain of knowledge management and technology-enhanced learning. On the one hand, those technologies enable the analysis of massive amounts of data as it is used, for instance, for learning analytics [Ch12, Jü18]. Thus, artificial intelligence and machine learning algorithms improve the data analysis in this domain. On the other hand, the technologies enable the rise of new or massively improved systems for educational purposes. One topic that is especially promising and receives increasing interest are chatbot-based systems that use natural language processing algorithms [TP12]. In the domain of education, such systems are known as pedagogical conversational agents, i.e. chatbots used for educational purposes. Even though chatbots have been researched for many years, the application of machine learning and artificial intelligence algorithms enables new possibilities regarding natural language processing. One example in which major progress has been made due to machine learning is the so-called intend classification. The intent classification is an important part of the development of conversational agents as it is used to analyze natural language input with the aim of classifying it [Br17]. Due to these advances in natural language processing, the availability of easy-to-use libraries for chatbots increased rapidly. For this reason, researchers as well

[1] Universität Göttingen, Professur für Anwendungssysteme und E-Business, Platz der Göttinger Sieben 5, 37073 Göttingen, shobert@uni-goettingen.de

doi:10.18420/delfi2019_289

as practitioners were encouraged to develop conversational agents.

By analyzing the literature targeting pedagogical conversational agents, it is noticeable that researchers from different fields of research like computer science, information systems, education, and psychology are participating in this research stream. This interdisciplinarity results on the one hand in interesting complementary research results. On the other hand, there is no uniform procedure for evaluating the pedagogical conversational agent prototypes. The different research disciplines apply their own evaluation approaches to pedagogical conversational agents. Due to this, pedagogical conversational agents are often only evaluated regarding selected aspects. For instance, researchers in the information systems discipline often apply widely adopted models for technology acceptance or adoption like TAM [DBW89]. For analyzing usability in other fields, UEQ [LHS08] or meCUE [MR13] are common standardized questionnaires. In some cases, software prototypes are evaluated from a technical perspective whereas others focus on learning success. However, as the field of pedagogical conversational agents is interdisciplinary, it needs to be questioned if evaluating further aspects from other closely related fields might be beneficial for this interdisciplinary research domain. Thus, the aim of this research paper is to outline which commonly used evaluation models exist and which methods can be used for evaluating educational chatbots comprehensively. Based on this, we discuss possible implications for future research studies that aim at evaluating pedagogical conversational agents comprehensively targeting evaluation objectives from multiple perspectives.

As a starting point, we analyze the current state-of-the-art of evaluating pedagogical conversational agents based on a systematic literature review. Our focus is on identifying cases that have conducted evaluation studies. In particular, we are interested in the cases' objectives, procedures, and corresponding measuring instruments as well as discussing implications for future research:

RQ1: Which state-of-the-art approaches to evaluate pedagogical conversational agents are common?

RQ2: Which implications can be derived for future evaluation studies of pedagogical conversational agents?

To answer these two research questions, the remainder of this article is structured as follows: First, we briefly describe the theoretical background by defining the term pedagogical conversational agent and outlining related research. Based on this, we present the procedure of our systematic literature review that we applied to identify state-of-the-art evaluation approaches of the pedagogical conversational agent in the literature base. Following this, we discuss implications for future evaluation studies. Finally, we summarize our work in the conclusion.

2 Theoretical Background

Conversational agents – also known as chatbots, chatterbots, bots or interactive agents – that focus on educational purposes are known under different terms in this interdisciplinary research field. For instance, they are named pedagogical conversational agents [TP16]. These agents are a special form of chat applications that interact with learners automatically using a natural language interface [KHB07].

In recent research studies, especially messenger-like pedagogical conversational agents are popular [HM19]. This type of agents uses user interfaces known from instant messenger apps (like WhatsApp, Facebook or Telegram). In some cases, the pedagogical conversational agents are using APIs from those messenger platforms to integrate themselves into those platforms. In other cases, messenger-like agents are implemented as standalone apps and implement chat interfaces on their own. In contrast to messenger-like agents that are designed like messenger apps, pedagogical conversational agents can also be visualized using artificial bodies. This type of pedagogical conversational agents is called 'embodied conversational agent'. They often also interact with learners using text-based inputs. In some cases, also voice-enabled systems exist. The main difference to messenger-like agents is that the visualization does not only include a chat-interface but also shows an artificial body of the system. A common approach is to show a human-like avatar that is able to show facial expressions and gestures. According to [Pe16], messenger-like systems are currently preferred because "interaction takes place through messaging applications to which students are already very keen on" [Pe16].

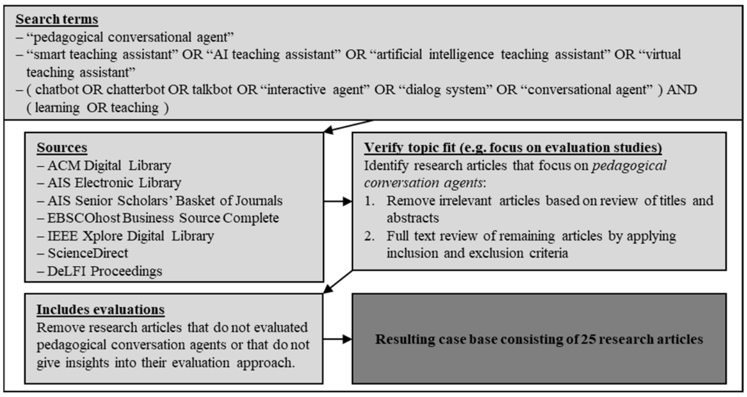
From a technical perspective, pedagogical conversational agents are often similar to other implementations of chatbot-based systems. Usually, the user's natural language input is processed in the natural language understanding step. In this step, the user's input is often analyzed regarding known patterns in order to identify predefined intents. Based on the identified intents, information that is needed to respond to the user's input is processed before an answer is generated in the natural language generation step. Common approaches for handling natural language inputs are machine learning approaches, e.g., based on *spacy* or pattern matching approaches based on the *Artificial Intelligence Markup Language* (see e.g. [Mi09]).

In addition to conversational agents in other application domains, more advanced pedagogical conversational agents often store learning objects that encompass learning materials that can be requested by learners (see e.g. Co18). In some cases, even complex learning paths are stored that describe which learning objects should be learned next. To use this information, a pedagogical conversational agent needs to retain the learner's current state of knowledge to be able to deliver the next appropriate learning object. Due to this increasing complexity, these pedagogical conversational agents can also be classified as intelligent tutoring systems [CLW17].

3 Research Design

To answer our research questions, we conducted a literature search to identify research studies that evaluated pedagogical conversational agents. The resulting list of relevant studies creates the case base for the following analyses.

To create our case base, we conducted a structured literature search within major IS journals and IS-related databases and the digital library of the DeLFI proceedings. In particular, we searched the digital libraries of ACM, AIS, EBSCOhost, IEEE, and ScienceDirect as well as the AIS Senior Scholars' Basket of Journals [As18]. To search for literature, we used several different search terms to identify relevant papers using as outlined in Figure 1. Due to the focus of our study, we only included results targeting educational settings, which is reflected in the search terms. After conducting the literature search, we first checked titles and abstracts of search results and removed irrelevant articles. Then, we reviewed the full texts of the remaining articles for content fit using inclusion and exclusion criteria in accordance with [vo15]. Finally, we included all articles in our case base that describe evaluation studies.



Search terms
– "pedagogical conversational agent"
– "smart teaching assistant" OR "AI teaching assistant" OR "artificial intelligence teaching assistant" OR "virtual teaching assistant"
– (chatbot OR chatterbot OR talkbot OR "interactive agent" OR "dialog system" OR "conversational agent") AND (learning OR teaching)

Sources
– ACM Digital Library
– AIS Electronic Library
– AIS Senior Scholars' Basket of Journals
– EBSCOhost Business Source Complete
– IEEE Xplore Digital Library
– ScienceDirect
– DeLFI Proceedings

Verify topic fit (e.g. focus on evaluation studies)
Identify research articles that focus on *pedagogical conversation agents*:
1. Remove irrelevant articles based on review of titles and abstracts
2. Full text review of remaining articles by applying inclusion and exclusion criteria

Includes evaluations
Remove research articles that do not evaluated pedagogical conversation agents or that do not give insights into their evaluation approach.

Resulting case base consisting of 25 research articles

Fig. 1: Literature-based approach to create our case base

As a result, we got a case base consisting of 25 research papers that provides insights into the evaluation of pedagogical conversational agents.

4 State-of-the-Art Evaluation Approaches

We present the results of our analysis that outline how researchers evaluate pedagogical conversational agents. Before we describe the three dimensions "evaluation objectives", "procedures" and "measuring instruments" in detail, we give an overview of our classification matrix in the following section 4.1. A complete list of all considered papers including further information is available at https://publikationen.as.wiwi.uni-goettingen.de/getfile?DateiID=743.

4.1 Classification matrix

By applying the steps described above, we inductively identified seven characteristics of evaluation objectives, four characteristics of procedures and five characteristics of measuring instruments in our case base. Table 2 shows the result of our classification process. We categorized all 25 studies of our case base and highlight dominant intersections between procedures and evaluation objectives as well as measuring instruments and evaluation objectives in the matrix.

		Evaluation objective						
		Acceptance & adoption	**Learning success**	**Increased motivation**	**Further beneficial effects**	**Usability**	**Technical correctness**	**Further psychological factors**
Procedures	Wizard-of-Oz experiment with a non-functional simulated pedagogical conversational agent	x				x		x
	Laboratory study with a functional pedagogical conversational agent under controlled conditions	x	x	x	x	x		x
	Field study with a functional pedagogical conversational agent in the field	x	x	x	x	x		x
	Technical validation of algorithms or software components without an experimental setup						x	
Measurement instruments	Quantitative survey after a laboratory or field study with users of a pedagogical conversational agent (*Post*)	x	x	x	x	x		x
	Quantitative survey before and after a laboratory or field study with users of a pedagogical conversational agent (*Pre/Post*)	x	x	x	x	x		x
	Qualitative interview study with users of a pedagogical conversational agent		x			x		x
	Discourse analysis of transcripts of the interaction of learners with a pedagogical conversational agent	x					x	
	Analysis of technical log files produced during the usage of a pedagogical conversational agent						x	
Legend: x indicates dominant intersections identified in the literature base								

Tab. 2: Classification matrix of our literature-based analysis that includes the dimensions procedures, measurement instruments and evaluation objective

4.2 Evaluation objectives

The objectives of evaluation studies influence the procedures to be applied and corresponding measuring instruments. In our analysis, we identified that many papers do not only focus on one specific evolution goal but often target multiple. In some cases, the authors of studies do not even state which objective they pursued. Due to these difficulties, we chose to allow the classification of one study in multiple characteristics. Thus, the total sum of evaluation objectives may be larger than the number of studies in our case base.

When the authors did not state the evaluation objective directly, we derived it manually from the result sections.

One of the most common evaluation objectives in our case analysis is to analyze the *technology acceptance and adoption*. Approx. 30 % of all analyzed studies focused on this main evaluation goal. In many cases, the analyses also include the evaluation of educational goals like *learning success*. Although analyzing the overall learning success seems to be the most important aspect of a technology-enhanced learning solution, only approx. 33 % of the studies focused on this evaluation goal. However, closely related evaluation objectives were analyzed in the case base: For instance, the effect of an *increased motivation* triggered by using a pedagogical conversational agent was analyzed. Researchers tested whether the interactive natural language-based conversation with a pedagogical conversational agent can result in an increased motivation [e.g. TP12]. Even though this is closely related to an increased learning success, we decided to emphasize this effect separately, because research studies do not necessarily try to measure whether the increased motivation also results in a higher learning success – also this assumption suggests itself. Further studies analyzed individual education-related goals that cannot be aggregated to a specific term. We opted to aggregate them under the generic term *further beneficial effects on learning processes* because the aggregated studies target individual but closely related positive effects of pedagogical conversational agents that are not explicitly attributable to an increased learning success or an increased motivation. An exemplary beneficial impact that has been analyzed is for instance that pedagogical conversational agents support learners to understand a given topic [e.g. KHB07].

In addition to these educational objectives, the *usability* of pedagogical conversational agents has been targeted in approx. 25 % of the studies. Although we would have expected that correlations between user-friendliness, acceptance and learning success exist, this was only analyzed rarely. In particular, interrelations between the technology acceptance, the actual system use, and the learning success would provide additional insights into the effectivity of providing pedagogical conversational agents to students. In this way, the results of TAM-based questionnaires could be verified with actual usage data of pedagogical conversational agents.

Another technical evaluation goal is to verify the *algorithmic or technical correctness* of software implementations of pedagogical conversational agents. Research items in this category validate for instance if the algorithmic implementation responds with a correct answer (see e.g. [PPR08]).

Finally, we identified evaluation studies that analyzed further *psychological factors* – in addition to motivation mentioned above – like the enjoyment (see [e.g. Fr17]) or the implication of the design of pedagogical conversational agents on the social interaction of participants with the agent.

Summarizing the evaluation objectives, the interdisciplinarity of the research stream becomes apparent in a large variety of objectives. Our classification of the evaluation objectives outlines that the goals targeted in research studies vary a lot from technical validations of algorithms via design-oriented evaluations of user acceptance to psychological effects. Although these goals are very heterogeneous, they all have their justification and are therefore useful for a comprehensive examination of pedagogical conversational agents.

4.3 Procedures

While the total number of evaluation goals was quite large, we could aggregate the applied procedures in the case studies to only four different approaches:

First, we can distinguish evaluation studies using functional software implementations in comparison to early-stage non-functional software artefacts. A common approach for evaluating the concepts of pedagogical conversational agents at an early stage of development is to conduct so-called *Wizard-of-Oz experiments* [MHH12]. In these experimental setups, participants get access to a semi-functional prototypical implementation of a pedagogical conversational agent. The software often only consists of a mock-up-like user interface that can accept inputs of learners that are automatically redirected to a remote operator. In contrast to fully functional pedagogical conversational agents, complex functional implementations including natural language processing are usually still missing. Thus, the remote operator takes over all control tasks and responds to the users' input. A critical success factor of this evaluation approach is that the study participants are assuming that they are interacting with a fully operational pedagogical conversational agent. Using this evaluation approach, for instance, the learners' technology acceptance or the usability of a pedagogical conversational agent can be evaluated in an early stage of development (see [e.g. KHB07]).

In our case base, we identified two more types of studies that are common in other fields of research as well: *laboratory studies* and *field studies*. In contrast to Wizard-of-Oz experiments, researchers in our case base usually conduct both, laboratory and field studies, using fully-functional software implementations. However, as it was expected, in most cases the software implementations used in field studies are often more advanced than in laboratory studies. Thus, we conclude that the conduction of laboratory studies is suited to test prototypes in an early state of the development process. In contrast to that, pedagogical conversational agents can only be evaluated in a field study after the software is fully implemented and production-ready.

Besides these three procedures in which human participants interact with a software artefact, we identified *technical validations* of algorithms or software components that did not conduct any experiment with human participants. Evaluations in those research studies are often focused on verifying the correctness of algorithms. To do this, researchers, for instance, provide artificial natural language input to the system and examine whether it responds with a valid output.

Summarizing the identified evaluation procedures, we conclude that besides laboratory and field studies that are common in other domains as well, especially Wizard-of-Oz experiments and technical validations are suited in research on pedagogical conversational agents. Particularly Wizard-of-Oz experiments seem important to assess the suitability of natural-language-based systems in an early stage of development. Algorithmic evaluations seem important as well, as the handling of natural language user inputs are currently a critical success factor for pedagogical conversational agents.

4.4 Measuring instruments

During analyzing the measuring instruments used in the cases of the pedagogical conversational agent literature, we could identify five different instruments.

The most prominent types of instruments are *quantitative surveys* before or/and after a field or laboratory study. In our classification matrix, we distinguish between surveys that ask the participants to complete a questionnaire after the study has finished (post-survey) and those who ask the participants to complete a questionnaire before and after it (pre- and post-survey). Whereas the post-surveys are often used to assess the technology acceptance or usability, pre- and post-surveys have the ability to evaluate the learning success by comparing the state of knowledge before the intervention with the state afterward.

As a counterpart to these quantitative instruments, researchers in two cases conducted *qualitative interviews* with participants of a study to get detailed feedback on the impact of using pedagogical conversational agents on the learning success and on resulting effects.

Finally, we identified two technical measuring instruments used in multiple cases: *Transcripts of dialogs* and *technical log files*. With a focus on evaluating the interaction among learners and pedagogical conversational agents, researchers in multiple studies stored all dialogues during evaluation studies. Afterward, the transcripts could be analyzed using discourse analyses to get insights into the technical correctness of the logic behind the pedagogical conversational agents as well as into the learners' behavior. In addition to that, researchers often collect technical log data in field experiments.

Summarizing the results from analyzing the measuring instruments, we conclude that most often quantitative data, as well as technical log data, are collected. Both can be analyzed statistically. In contrast to that, the analysis of transcripts using discourse analysis seems especially time-consuming as most researchers in our case base did it manually (e.g. by coding the answers of the conversational agent for correctness).

5 Discussion and Implications

The first research goal of our study (RQ1) was to analyze the literature base of pedagogical conversational agents to identify evaluation studies and to examine them. In our analysis, we aggregated the results to depict the evaluation objectives in seven categories, the procedures in four categories and the measuring instruments in five categories. To derive implications for future research studies (RQ2), we carefully analyzed these categories and propose four evaluation steps based on the evaluation methods described above. By applying all four evaluation procedures, it is possible to cover all identified evaluation objectives. Thus, a comprehensive evaluation of pedagogical can be achieved.

(1) Wizard-of-Oz experiments are suited to gather early feedback of learners about a pedagogical conversational agent. The main advantage of this method is that it can be applied even before a functional software prototype has been developed – only a remotely controllable user interface is required. By conducting a Wizard-of-Oz experiment, the evaluation objectives *acceptance & adoption* and *usability* can be targeted. (2) Subsequently, an important step of developing pedagogical conversational agents is the technical implementation of the natural language understanding and subsequent answer generation. As this step is critical to the success of a pedagogical conversational agent, we follow the researchers in our case base who validated the technical algorithms in their studies. Consequently, we propose to conduct technical validations after the natural language processing component has been implemented. As a precondition, researchers must create a dataset consisting of valid input data (e.g. questions) and expected output data (e.g. in-tended answers). Using such a dataset, the algorithms can be tested for validity. To verify its performance, the transcripts of the dialogues (input data and resulting output data) as well as technical log files of the algorithms should be collected and analyzed. (3) Following the technical validation and the finalization of a first functional prototype of a pedagogical conversational agent, we propose to conduct a pre-test in a laboratory setting with the aim to test the agent with potential users. This gives researchers the possibility to get feedback concerning the learning processes, the related learning content as well as the suitability of the conversational dialog form from potential users before an agent is used productively on a large scale. Furthermore, laboratory experiments can be conducted even when the prototype is not yet fully functional and can only operate in a limited setting. In addition to the analyses of transcripts of dialogues and technical log files, post-experiment questionnaires should be handed out to the participants. In this way, users can express their opinions, for instance, targeting technology acceptance and adoption. We state that learning success should not be evaluated in this phase of the development of a pedagogical conversational agent. Analyzing the long-term learning success is usually a complex task that requires elaborated measuring instruments. (4) After a successfully completed laboratory experiment and the finalization of the software implementation, a field experiment gives researchers the opportunity to test the conversational agent in a real setting (like in [PP13]). The overall goal of this last evaluation step is to test the implications of using a pedagogical conversational agent towards the learning success. Due to this, it seems particularly important to us that the pedagogical conversational agent including the natural language processing is fully operational before conducting the field experiments. If a pedagogical conversational agent is not able to understand the intents of learners, they may refuse to use it directly. If, however, the agent is used in a large-scale field study, researchers can test whether the expected effects of pedagogical conversational agents (e.g. learning success, increased motivation or further psychological effects) occur. In

particular, the analysis of long-term effects – like learning success – must be measured in longitude studies with pre- and post-experiment surveys. Additionally, interviews with key users of the pedagogical conversational agent may give researchers the possibility to get more detailed feedback about the effects on the learning process. Finally, in accordance with the other evaluation steps, we also propose to save transcripts of dialogues and technical log files for an in-depth analysis of usage patterns and user engagement. The analysis of the transcripts of dialogues also gives researchers the possibility to assess the most important learning contents (i.e. the contents that are mentioned most often in the chat dialogues). Thus, it is possible to identify key learning concepts as well as contents that are formulated in a misleading way.

		Evaluation objective						
		Acceptance & Adoption	Usability	Learning Success	Increased Motivation	Further beneficial effects	Psychological Factors	Technical Correctness
Procedure	Wizard-of-Oz experiment	x	x					
	Technical validation							x
	Laboratory experiment	x	x					x
	Field experiment			x	x	x	x	

Tab. 3: Comparison of the evaluation methods with the evaluation objectives

By applying these four steps as part of evaluations of pedagogical conversational agents, the identified objectives can be addressed (see Tab. 3). Thereby, researchers might be able to contribute not only to one field of research but to multiple subfields in this interdisciplinary research domain of pedagogical conversational agents. However, targeting multiple objectives from different fields of research at the same time might not be easy: This might be particularly true if multiple models (like TAM [DBW89]) and standardized questionnaires (like UEQ [LHS08] or meCUE [MR13]) should be considered at the same time. For instance, [HST18] indicated that using and comparing different standardized questionnaires might be difficult. Thus, in future evaluation studies, it might be interesting to analyze this in more detail.

6 Conclusion

The aim of our research study was to analyze the current state-of-the-art of evaluating pedagogical conversational agents in research studies in order to derive implications for future evaluation studies. Our case-based analysis of the literature depicts that there is a large number of different evaluation approaches common in research on pedagogical conversational agents. Nevertheless, our analysis shows that often only selected aspects are being analyzed. We attribute this to the interdisciplinary nature of the research area.

This leads to the fact that researchers only analyze particular aspects, which are in the focus of their discipline. Due to this, comprehensive evaluations that analyze pedagogical conversational agents from different perspectives are usually missing. To provide a guideline for future research, we proposed four evaluation steps that might be used as a guideline for evaluation studies. However, future research is needed to test if evaluating multiple objectives in one research study is practicable and provides adequate contributions.

Limitations of our research approach are based on the research design of our literature review. After conducting our structured literature review, we selected our cases as described above. As with any literature review, the assessment and classification of the relevant articles may be subjective in some cases. However, we tried to formalize it as much as possible. Nevertheless, our results contribute to research and practice as they provide a guideline that can be applied in future research studies as well as in projects of practitioners. Both target groups can benefit from the analysis as it provides a guideline of how to analyze pedagogical conversational agents during development processes.

Bibliography

[As18] Association for Information Sytems: Senior Scholars' Basket of Journals, https://aisnet.org/page/SeniorScholarBasket, accessed: 12/07/2018.

[Br17] Braun, D. et al.: Evaluating Natural Language Understanding Services for Conversational Question Answering Systems. In (Jokinen, K. et al. ed.): Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. Association for Computational Linguistics, Stroudsburg, PA, pp. 174–185, 2017.

[Ch12] Chatti, M. A. et al.: A reference model for learning analytics. International Journal of Technology Enhanced Learning 5/6 /12, pp. 318–331, 2012.

[CLW17] Crockett, K.; Latham, A.; Whitton, N.: On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees. International Journal of Human-Computer Studies 97/17, pp. 98–115, 2017.

[Co18] Coronado, M. et al.: A cognitive assistant for learning java featuring social dialogue. International Journal of Human-Computer Studies 117/18, pp. 55–67, 2018.

[DBW89] Davis, F. D.; Bagozzi, R. P.; Warshaw, P. R.: User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science 8/89, pp. 982–1003, 1989.

[Fr17] Fryer, L. K. et al.: Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. Computers in Human Behavior 75/17, pp. 461–468, 2017.

[HM19] Hobert, S.; Meyer von Wolff, R.: Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. In (Ludwig, T.; Pipek, V. ed.): 14. Internationale Tagung Wirtschaftsinformatik (WI 2019) Tagungsband, pp. 301–314, 2019.

[Jü18] Jülicher, T.: Education 2.0: Learning Analytics, Educational Data Mining and Co. In (Hoeren, T.; Kolany-Raiser, B. ed.): Big Data in Context. Legal, Social and Technological Insights. Springer International Publishing, Cham, pp. 47–53, 2018.

[KHB07] Kerly, A.; Hall, P.; Bull, S.: Bringing chatbots into education: Towards natural language negotiation of open learner models. Knowledge-Based Systems 2/07, pp. 177–185, 2007.

[LHS08] Laugwitz, B.; Held, T.; Schrepp, M.: Construction and Evaluation of a User Experience Questionnaire. In (Holzinger, A. ed.): HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings. Springer, Berlin, i.a., pp. 63–76, 2008.

[MHH12] Martin, B.; Hanington, B.; Hanington, B. M.: Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions. Rockport Publishers, 2012.

[Mi09] Mikic, F. A. et al.: CHARLIE: An AIML-based chatterbot which works as an interface among INES and humans. In: 2009 EAEEIE Annual Conference, pp. 1–6, 2009.

[MR13] Minge, M.; Riedel, L.: meCUE - Ein modularer Fragebogen zur Erfassung des Nutzungserlebens. In (Boll, S., Maaß, S. & Malaka, R. ed.): Mensch & Computer 2013: Interaktive Vielfalt. Oldenbourg Verlag, München, pp. 89-98, 2013.

[Pe16] Pereira, J.: Leveraging chatbots to improve self-guided learning through conversational quizzes: In: Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality, pp. 911–918, 2016.

[PP13] Pérez-Marín, D.; Pascual-Nieto, I.: An exploratory study on how children interact with pedagogic conversational agents. Behaviour & Information Technology 9/13, pp. 955–964, 2013.

[PPR08] Pilato, G.; Pirrone, R.; Rizzo, R.: A KST-BASED SYSTEM FOR STUDENT TUTORING. Applied Artificial Intelligence 4/08, pp. 283–308, 2008.

[TP12] Tamayo, S.; Pérez-Marín, D.: An agent proposal for Reading Understanding: Applied to the resolution of maths problems. In: 2012 International Symposium on Computers in Education (SIIE). IEEE, Piscataway, NJ, pp. 1–4, 2012.

[TP16] Tamayo-Moreno, S.; Perez-Marin, D.: Adapting the design and the use methodology of a pedagogical conversational agent of secondary education to childhood education. In: 2016 International Symposium on Computers in Education (SIIE). IEEE, Piscataway, NJ, pp. 1–6, 2016.

[vo15] vom Brocke, J. et al.: Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. Communications of the Association for Information Systems 1/15, pp. 205–224, 2015.