How do Pedagogical Conversational Agents affect Learning Outcomes among High School Pupils: Insights from a Field Experiment

Sarah Waldner
University of Innsbruck
s.waldner@student.uibk.ac.at

Isabella Seeber
Grenoble Ecole de Management
isabella.seeber@grenoble-em.com

Lena Waizenegger
Auckland University of Technology
lena.waizenegger@aut.ac.nz

Ronald Maier
University of Innsbruck, University of Vienna
ronald.maier@univie.ac.at

Abstract

Pedagogical conversational agents (CA) support formal and informal learning to help students achieve better learning outcomes by providing information, guidance or fostering reflections. Even though the extant literature suggests that pedagogical CAs can improve learning outcomes, there exists little empirical evidence of what design features drive this effect. This study reports on an exploratory field experiment involving 31 pupils in commercial high schools and finds that students achieved better learning outcomes when preparing for their tests with a pedagogical CA than without. However, the drivers of this effect remain unclear. Neither the use frequency of the design features nor the pupils' expectations towards the CA could explain the improvement in marks. However, for the subjective perception of learning achievement, pupils' expectations was a significant predictor. These findings provide support for the use of pedagogical CAs in teaching but also highlight that the drivers of better learning outcomes still remain unknown.

Keywords: chatbot, conversational agent, education, learning outcomes, pedagogical agent

1. Introduction

Pedagogical conversational agents (CAs) as a form of intelligent tutoring systems (ITS) [1] are virtual characters in an online environment that support students' learning process [2]. Particularly, with the vast development of AI-enabled conversational abilities, pedagogical CAs experience a new hype [3], [4] and are introduced in the education environment as a means to relieve the workload of human teachers and support learners similar to teaching assistants or tutors [4], [5]. Particularly, messenger-like, chatbot-based CAs have gained increased popularity for education purposes [1] as they can be easily accessed anywhere and anytime. Furthermore, in comparison to other passive or interactive technology-based learning aids such as explainer videos, podcasts, blogs, learning games, or gamified quizzes [6], pedagogical CAs have a more sophisticated offering due to their comprehensive functionalities and anthropomorphic cues. Pedagogical CAs provide (customized) guidance and help students to acquire new knowledge, outline procedures and principles, of theoretical frameworks for instance, and enrich learning content with examples to improve understanding of the theoretical concepts [2].

Past research on ITS in general and pedagogical CAs in particular, which is mainly published in the education domain, showed that these technologies can improve learning outcomes [7]–[9]. However, it is somewhat unclear how strong this effect is; while some studies report big learning improvements of almost one grade [7], others report small to medium effect sizes or no effects at all [2], [8], [10]–[12]. These findings from (meta-)reviews and comprehensive empirical studies provide valuable insights and show that pedagogical CAs can be useful learning aids (in addition to traditional learning tools) that are here to stay. Yet, past research also acknowledges the high heterogeneity of the effects of pedagogical CAs, suggesting additional (moderator) variables may play a significant role [8].

Much of this past research considers individual dispositions or the learning environment as influencing factors but leaves the potential effects of a CA's technological features largely unaddressed. Yet, it has been suggested that atomic design features need to be taken into account when [13] developing pedagogical CAs as they might have a considerable effect on the learning outcomes. In fact, design features can directly affect learners' motivation, attitudes [14] and behaviors [2]. These design features include those related to the agent's appearance (e.g., gesture, facial expression, presence) or the agent's role (e.g., navigator, mentor, feedback) [2]. Even though these investigations of design features have created valuable insights, we lack an understanding of how the actual use of pedagogical CAs' more content-oriented design features affects learning outcomes. The goal of this research is to open the black box of CAs and improve our understanding on which content-related CA features instead of CA's role and appearance drive learner's success. Moreover, our cumulated knowledge on technology adoption suggests that positive beliefs of users towards the technology are indicative of its use [15]. Yet, we know little about how

URI: https://hdl.handle.net/10125/79379
978-0-9981331-5-7

such expectations affect the use of pedagogical CAs, which in turn may impact the user's learning outcomes. The aim of this paper is to address these gaps and try to understand how expectations and actual engagement with pedagogical CA features foster learning outcomes. Therefore, we state the following research question: *How does the use of pedagogical CA features and expectations towards CAs of high-school pupils affect their learning outcomes?*

To answer this question, we conducted an exploratory field experiment with 31 pupils from commercial high schools learning for tests in business administration. Our findings suggest that pedagogical CAs improve learning outcomes significantly compared to learning with traditional learning materials (e.g., books). However, this effect can neither be explained by the use of the CAs features (i.e. knowledge function and quiz function) nor by pupils expectations towards the CA. Only students' subjective perception of learning achievement has been found to be significantly and positively driven by pupils' expectations of the CA, emphasizing the relevance of pupils' belief towards technologies. This leaves open a gap in our understanding and calls for more research on how specific features need to be designed and used to improve concrete learning outcomes and perceived learning achievements. With this, we contribute to the emergent IS research on pedagogical CAs to better understand how the relationships between technology capabilities, psychological processes, and learning emerge [3], [16].

2. Background

Pedagogical CAs are part of the research stream of ITS [1] and comprise technologies such as chatbots, virtual agents, or other digital agents that aim at supporting students' learning practices and processes. They provide individualized and personalized support by providing learners with instructions or feedback. They commonly consist of a communication interface for presenting and receiving information, a domain model that contains the information to teach, a student model that has the students' learning states and a pedagogical model that represents instructional strategies [8].

Pedagogical CAs can support formal learning settings like at school or university and informal settings like at home [1]. They can be used anytime and anywhere as the agents are usually mobile and web-based [1]. Pedagogical CAs can fulfill various functions and roles [13], [17]. They can act as navigators, facilitators [2] or as education agents to reinforce learning [4]. For example, they can show how a student can successfully complete a task, give individualized feedback, provide information material to the student, and ask questions to test their knowledge [17].

Research on ITS and, more specifically, pedagogical CAs is not new and has been conducted for at least a decade [4]. Several reviews and meta-reviews have been conducted on ITS and, in particular, on pedagogical CAs and found that these technologies impact learning success positively with small to moderate effects [2], [8], [10]–[12]. Pedagogical CAs have also been studied across several subject domains, including maths, computer science, languages etc. In the subject domain of accounting, a domain that is considered close to the business administration topic under investigation in this study, positive and moderate effect sizes have been found for ITS [12]. For example, a recent meta-analysis investigated the effects of the chatbot Aleks on student learning outcomes. Aleks is an adaptive learning system and assesses the knowledge of any given student to provide personalized learning paths to master the expected concepts in a certain knowledge domain (e.g., Algebra 1). Half of the studies in the meta-analysis showed a positive effect on students' learning, and the other half of the studies had a negative effect. The analysis using a random-effects model suggests a positive influence; however, given the high heterogeneity of studies, the authors deduce that additional moderator variables may play a significant role [8]. What these moderators or mediators could be is open for research to elicit.

In order to understand the underlying dynamics between pedagogical CA use and learning outcomes and be able to explain the effect relationships, we need to apply a more fine-grained approach and zoom in on the actual design features of the CA and if and how they affect the learner's achievements. Thus far, we lack a good understanding of pedagogical CAs' design features and their effects.

On the one hand, pedagogical CAs provide affective responses using their speech, body gestures, or facial expressions [18]. We refer to this as emotion-related design features. AutoTutor is an example of an affect-sensitive and interactive CA that infers affective states from students' eye movements, body postures, and keyboard and mouse behaviors. Given a student's affective state, AutoTutor would then provide appropriate responses, which increased students' engagement and learning motivation [18]. Consequently, CA features that include uplifting CA responses (e.g. amusing sentences or pictures) or provide words of encouragement [19] can lighten the mood of pupils and keep them motivated.

On the other hand, pedagogical CAs also deliver, recommend, or assess content. We refer to this as content-related design features. For example, a recent study involving ten students showed that participants

perceived the pedagogical CA as useful as it helped them improve their argumentation quality [20].

While these preliminary findings are promising, we see a lack of research focusing on content-related design features and their effects. Particularly, the actual usage of such design features on learning outcomes has thus far not been explored [2]. Yet, gaining such knowledge is relevant in order to advance our understanding how technology use can enhance learning and improve learning outcomes [16].

3. Hypotheses Development

Pedagogical CAs can help students learn [11]. They support students in constructing more elaborate answers by engaging them in conversations and drawing out the students' knowledge [7]. When evaluating their knowledge, students with chatbot support outperformed students with textbook support [7]. Based on the results of several experiments, Graesser and colleagues argue that the use of chatbots leads to learning gains no matter which form of assessments was used, including essays, multiple-choice questions, or tasks that require problem-solving [7]. The main rationale is that pedagogical CAs provide a more individual learning experience for pupils that fosters their learning process and thus results in improved learning outcomes [4], [8]. Therefore, we propose:

Hypothesis 1: Pupils' learning outcomes will be better with a pedagogical CA than without a pedagogical CA.

Related research on pedagogical CAs has claimed that it remains largely unclear how pedagogical CAs actually induce learning gains [1], [3], [13]. It is probable that emotion-related and content-related design features of CAs can foster a student's motivation to learn, resulting in improved learning outcomes.

The use of gamification features, such as exercises or quizzes [4] has shown that pupils actively engage by answering, for example, content-related questions [21]. We classify this content-related design features as quiz feature. Such interactive engagements should help students recall, retrieve, reflect, and strengthen students' domain knowledge [13], [22]. However, empirical evidence for this theoretical relationship is scarce. Benotti et al. (2018) evaluated a chatbot that helped teenagers learn computer science using a quiz feature as a means of formative assessment. Data from the 34 pupils that took part in the field experiment revealed that pupils did not consider the CA with a quiz function more useful than a CA without the quiz feature, but they reported a higher task completion [23]. However, it remains unclear if a quiz feature can foster learning gains. Besides quiz features, pedagogical CAs can also present the learning content, which can be considered a form of passive engagement [21]. We classify this content-related design features as knowledge features. In contrast to quiz features, the knowledge feature repeats the class content [24], [25]; not with the goal to test the knowledge, but with the goal to let the pupil (re-)acquire information.

Given these findings from past research, we propose that the frequent use of knowledge features (passive engagement) and quiz features (interactive engagement) should increase learners' motivation, which should result in improved learning outcomes. We speculated that besides the more commonly studied emotion-related design features of CAs (e.g., speech, gestures, appearances), explanatory power should be detectable in the actual interactions with the learning content, achieved through the use of content-related design features. This should be a reasonable heuristic as for the average student, learning occurs through repetition. Pedagogical CAs can provide this repetition with ease, which, in turn, can be measured through the frequency a pupil used the CA. Thus, we claim

Hypothesis 2: More frequent use of (H2a) knowledge features and (H2b) quiz features of the pedagogical conversational agent will be associated with better learning outcomes.

People are more likely to continue using the CA service and experience an emotional uplift when the initial expectations towards the CA were positive and have been met [26]. Also, in the context of pedagogical CAs, a positive attitude towards new technologies is believed to foster pupils' perceived usefulness of the pedagogical CA, which makes it more likely that the pedagogical CA will get adopted [19]. It is unclear, however, how positive expectations regarding the CA link to learning outcomes. One rationale is that CA technologies are still quite novel for pupils, which creates extra interest to learn [4]. Consequently, pupils that have positive expectations towards CA technology should have a higher motivation to use such tools in the learning process and thus perform better. We speculate that when pupils hold high expectations towards the CA, the CA could act like a placebo. Placebo effects do not only occur in the medical context but have also been identified in the consumer research context. Related research found that brands that promise performance improvements (e.g., jogging shoes) can lead to not only subjective improvements but also objectives as these promises decrease anxiety and increase self-esteem in prospective users [27]. Similar psychological mechanisms may also exist in the context of education and learning with CAs. Pupils that are promised a CA

that will help them learn may form high expectations towards that technology, which gives them extra motivation to learn, resulting in subjective and objective learning improvements. Therefore, we propose

Hypothesis 3: Higher positive expectations towards conversational agents will be associated with better learning outcomes.

Thus far, we have argued that the actual use frequency of a quiz and knowledge feature as well as the positive expectation of pupils towards CA technologies should drive learning outcomes. We propose that the use frequency should mediate the relationship between a pupil's positive expectations towards a CA and the ensuing learning outcome. Pupils that have positive expectations towards CA technology should have a higher motivation to use such tools in the learning process. More frequent use of the CA should help the pupil to retain more knowledge and thus achieve better learning outcomes. Thus, higher positive expectations should lead to more usage, which in turn should result in better performance as the positive attitude should give an extra boost to the pupil's motivation to repeatedly use the CA features. Thus, we claim

Hypothesis 4: The frequency of use will positively mediate the positive relationship between expectations towards the pedagogical conversational agents and learning outcomes.

4. Methods

We applied experimental methods in a field setting to explore the effects of using a pedagogical CA on learning outcomes. We used a within-subject study design involving pupils from higher-level secondary commercial colleges (grades 9 and 10). Data was collected between March and June 2019. All pupils were exposed to both, the treatment condition with pedagogical CA and the control condition without pedagogical CA support. In the following, we will detail how we manipulated the treatment condition, the characteristics of the subjects, explain the procedure and measures, including reliability and validity checks.

Manipulation. The pupils in the treatment condition had the possibility to learn with the help of a CA in the form of a text-based, online chatbot while also using their traditional learning material. The pupils in the control condition learned with their traditional learning materials, which is their usual way of preparing for a test. We refer to this as the baseline. The goal of the chatbot was to provide pupils with explanations about the learning content, for which it used two features: a knowledge feature and a quiz feature. The knowledge feature provided topical content (e.g. what are the characteristics of a stock company) to help students revise the theory and better understand the learning material. The quiz feature tested the pupils' knowledge with knowledge and comprehension questions (see Figure 1). The user could respond via buttons or free text input.

Figure 1: Example of quiz feature

While all classes followed the same national curriculum for commercial colleges in Austria, the progress of each class was slightly different at the time of data collection. Therefore, the learning content for the knowledge and quiz feature of the CA had to be adapted for each class to ensure relevance for the pupils. For this purpose, the first author frequently interacted with the teachers to discuss, evaluate, and verify the content for the CA. To mitigate any timing effects, the exposure to the treatment (i.e. learning with CA support) was counter-balanced. This means that within a certain class, 50% of pupils were supported by a pedagogical CA during the preparation for their first test (learning phase 1) but did not use the pedagogical CA during the preparation for their second test (learning phase 2) and vice versa for the other 50% of the class. Overall, eleven chatbots covering the topics marketing, warehouse organization, procurement policy, sales contract, trade, advertisement, price elasticity and logistics were programmed using the chatbot platform Snatchbot.me.

The appearance of chatbots plays an important role [28] and therefore, several design choices were made: the chatbots were given a personality as related research suggests that personality increases user acceptance [26], [29]. Hence, each chatbot carried the name Hansi – the schoolbot, had a friendly face (implemented with a smiley and thumbs up) and greeted the pupils with a

welcome message. To increase engagement and personal connection, the chatbot addressed the pupils repeatedly with their username and interacted with them in a more juvenile language. To emphasize the juvenile language, chatbot responses included smileys and humorous and uplifting statements. The chatbot also sent images to provide pupils with a visual overview of the topics and subtopics it covered. All chatbot responses were pre-programmed and thus under the control of the experimenter.

Subjects and Sampling. Pupils attended the first and second grade of commercial college (9^{th} and 10^{th} grade) and their age ranged from 14 to 18 years. Of the 78 pupils that participated in the pre-survey of the field experiment, 42 actually used the CA at least once. This means that about half of the students used the CA support voluntary to prepare for their tests during the treatment phase. Of these 42 students, we had data on only 31 test scores with and without CA use. Of these 31 students, only 26 pupils provided their answer to the propensity to CA use language question. 18 of the pupils were female and 8 were male. Pupils were invited by using a convenience sampling strategy. We approached head principals and teachers, who, in turn, provided us with the opportunity to invite students during their classes. Pupils' participation was voluntary. The first author visited each class to present the overall study. To avoid any bias, we refrained from highlighting the variables under consideration. All participating students signed a consent form before taking part in the field experiment. Overall, six classes from three commercial colleges in Tyrol (Austria) participated. Students did not receive any (non-)monetary reward for their participation.

Procedure. The experimental duration varied slightly from school to school, with a minimum of 15 days. The first author visited each class, presented the field experiment, and informed them about their rights (e.g., withdraw from the study at any time without any consequences). Pupils were invited to the field experiment per e-mail, where they found the link to the CA and the link to the pre-survey. The assignment of classes to the treatment or control condition was random. The pre-survey informed pupils once again about the field experiment, and students could only start the survey once they gave informed consent. The learning phase with and without the chatbot was always five days. During this time, 50% of the pupils could learn with the chatbot as well as traditional learning materials, whereas the other 50% of the pupils had no chatbot support and thus relied on their traditional learning materials. Since we were interested in pupils' self-regulated learning engagement, we refrained from sending pupils in the treatment group notifications about the chatbot's availability. After the first learning unit (LU), teachers assessed pupils' knowledge with a test. For pupils in the treatment conditions, the first author sent pupils an e-mail after the test and before the results were released with an invitation to participate in the post-survey, where we collected data on their perceived learning outcome. Between the first and second learning phase, a break of a minimum of 5 days followed. Then, the second learning phase started where students had again five days to prepare for the test. Here, students that had chatbot support in the first learning phase were not provided with a chatbot in the second learning phase and vice versa. As in the previous round, the learning phase ended with a test that was administered and assessed by the teacher. Pupils that received the treatment in the second learning phase were invited to the post-survey. For the post-survey, we sent e-mail reminders to pupils and asked teachers to encourage pupils to take part in the last survey.

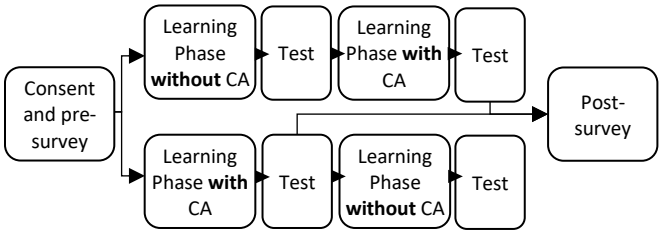

Figure 2: Overview of the experimental procedure

Measures. *Learning outcome*. We assessed the learning outcome with two variables; an objective measure *achieved test score* as well as a subjective measure *perceived learning achievement*. Pupils had tests after each learning unit. Thus, *achieved test score* measures the decrease or increase in points from a student's test result without CA support to the test result with CA support. To keep test scores comparable across classes and schools, the first author, who also holds a degree in business education (a pre-requisite degree for teachers at such business high schools) met with the teachers to discuss the test content and the assessment grid. All tests were assessed by the teachers who were blind to the allocation of students to the control or treatment condition as this was also randomized.

We also collected a more subjective measure of learning outcomes in terms of the pupils' *perceived learning achievement*. This variable measures whether the students think they were able to learn better and achieve better results when preparing with a chatbot compared to traditional learning materials (e.g., books). Items were derived from previous research [30]–[32] and adapted to focus on the performance and

effectiveness dimension of measurable benefits besides, e.g., cost savings or time savings [30]. We deemed this focus on performance and effectiveness relevant as we did not necessarily expect savings in time.

Expectations towards conversational agents: This variable measures the perceived importance of chatbots in daily life, business, and the school context in the future. The items were self-developed in order to address the beliefs towards a specific technology, that is CAs. Before using the scale in the experiment, we assessed its reliability and validity after the pre-tests of the experiment. All reliability and validity tests were satisfactory, and therefore, we included the scale in the experiment.

Frequency of quiz feature use. This variable describes the extent that pupils used the quiz feature during the experimental duration. We operationalized the measure by counting the number of exchanges between the pupil and the CA for dialogues that linked to the quiz feature.

Frequency of knowledge feature use. This variable describes the extent that pupils used the knowledge feature during the experimental duration. We operationalized the measure by counting the number of exchanges between the pupil and the CA for dialogues that linked to the knowledge feature. The classification of the quiz or knowledge feature was accomplished according to the dialogue model of the CA. Hence, no assessment of intercoder reliability for the code assignment was deemed necessary.

Control variables. We controlled for pupils' tendency to use a pedagogical CA when it interacts in a youthful language and gender. Past research found that pupils' learning outcomes when learning with a CA are better for female students [33]. Please refer to Table 2 for an overview of the items.

Reliability, validity, and assumption checks. We assessed reliability with Cronbach's Alpha. All multi-item constructs reached the threshold of 0.7 [34] thus supporting reliability. We assessed the validity of post-survey constructs learning achievement and chatbot expectations with exploratory factor analysis (rotation = Promax). The analysis showed that all factor loadings reached the threshold of 0.6 and the indicators loaded onto their expected factors without considerable cross-loading. Hence, we consider the results of exploratory factor analysis to support convergent validity.

We proceeded with assessing the statistical assumptions for the analyses. For the repeated measures ANOVA and multiple regression, the dependent variable should be normally distributed, and there should be no outliers. The inspection of QQ-Plots and Distribution of Student Residuals Plots [35] revealed no problematic (statistical) outlier. In addition, we assessed homoscedasticity with Breush-Pagan Test (homoscedasticity cannot be assumed if $p < 0.05$) and multi-collinearity using Variance Inflated Factor (VIF, potential multicollinearity problem if VIF > 4.0) [36]. All common thresholds were fulfilled so that we proceeded with hypotheses testing.

5. Results

5.1 The effects of pedagogical conversational agents on learning outcomes?

Hypothesis 1 suggested that pupils that use pedagogical CA will have better learning outcomes. We performed a repeated-measures ANOVA. As within-factor, we included the received test points with ($TestPts_{CB}$) and without CA use ($TestPts_{noCB}$). Gender and Propensity CA use language were specified as covariates. Since no interaction effect was assumed, a Sum of squares Type II model was selected. Our findings suggest that test performances of pupils were significantly better with CA support than without ($F(2, 28) = 25.181$, $p < 0.001$, $\eta 2 = .188$). In other words, pupils that prepared for their tests without a chatbot achieved on average a "satisfactory" mark with 6.339 points (SD= 2.067) compared to pupils who learned with a chatbot, who achieved on average a "good" grade with 8.161 (SD=1.793). The effect size with an η^2 of 0.188 suggests a moderate effect [37], [38]. Thus, H1 is supported.

Even though the pedagogical CA support seems to have considerably helped pupils to prepare for their test, it is unclear if it's actual use led to the improved test results, which was tested next.

5.2 What drives learning outcomes?

Hypothesis 2 and 3 suggested that when pupils have positive expectations towards CAs and actually use the pedagogical CA more frequently this would be associated with better learning outcomes. Table 1 provides the corresponding results from the regression analyses for achieved test score and perceived learning achievement.

The frequency of CA use in terms of knowledge (H2a) as well as the quiz feature use (H2b) were not significantly related to both the learning outcome measures. Thus, H2 is not supported. The pupils' positive CA expectation (H3) was a significant predictor for their perceived learning achievement (β 0.626, $p < 0.01$), but not for the actually achieved test score (β -0.356, $p < 0.10$). This means that pupils that had higher positive expectations towards CAs also believed that they were better in their tests. Since, we only found

support for the perception-based learning outcome variable, H3 is partly supported.

It should be noted that the control variables "Gender" and "Propensity CA use language" were significantly associated with the achieved test scores. This means that girls (β 1.458, $p < 0.05$) and pupils that would use CAs more if they had a juvenile language (β 0.811, $p < 0.01$) achieved higher test scores. Overall, the models could explain with 44% and 57% a considerable amount of variance in the dependent variables (see Table 1).

Finally, H4 suggested that the effect of pupils' expectations towards CAs on learning outcomes is mediated by the pupils' use frequency of the knowledge or quiz function. We performed parallel mediation analysis using the R plugin Process (Model 4) (Hayes) for R. No mediation effect of the theory or quiz feature could be found for *achieved test score* and *perceived learning achievement* (see Table 1).

Table 1: Results of parallel mediation analysis

	DV: Achieved test score	DV: Perceived learning achievement
Constant	1.955 (1.835)	-1.262 (1.469)
Control		
Gender	1.458 (0.580)*	-0.091 (0.464)
Prop. CA use language	0.811 (0.246)**	0.278 (0.197)
Independent Variable (H3)		
Expectation towards CAs	-0.356 (0.206)°	0.626 (0.165)**
Mediators (H2)		
UF Theory	0.011 (0.008)	-0.002 (0.006)
UF Quiz	0.013 (0.057)°	0.009 (0.005)
R2, F statistic	*0.441, 3.467** (df = 5; 22)*	*0.568, 5.790** (df = 5; 22)*
Direct effect	*-0.356° [-0.782, 0.071]*	*0.626** [0.285, 0.967]*
Indirect effects (H4)		
UF Theory	*0.036 [-0.268, 0.211]*	*-0.006 [-0.110, 0.164]*
UF Quiz	*0.166 [-0.023, 0.464]*	*0.107 [-0.041, 0.331]*

*Note: ° p< 0.10; * p < 0.05; ** p < 0.01; *** p < 0.001 UF = use frequency*

6. Discussion and Implications

This study set out to increase our understanding of the effects of pedagogical CAs' content-related design features and pupils' expectations towards the CA on learning outcomes. Based on our findings, we make the following contributions:

First, *pedagogical CAs improve learning performance*. We found that pupils achieved a higher test score when they prepared with a pedagogical CA compared to a test preparation phase with their traditional learning materials only. This allows us to conclude that the idiosyncratic functionalities that present and test knowledge as well as the social cues of CAs, which sets this technology apart from other offline and online learning aids proof to be successful in improving the learner's academic achievements. Consequently, with this finding, we add additional support to the extant literature [2], [8], [11], [12], that the usage of chatbot-based pedagogical CAs is beneficial for helping pupils learn.

Second, *pupils' positive expectations towards CAs make them believe to have achieved better learning outcomes, even though those with more positive expectations did not have better test scores.* Our findings show that when pupils expected CAs to be important in their future life, school, or work, they also thought to have performed better in their tests. Yet, the actual test scores did not confirm this. We theorized that a pupil's positive initial expectation could act as a placebo [27], providing the necessary motivation to learn. Our findings imply that a positive expectation is not enough to see actual learning gains. Nonetheless, *pupils' positive expectations towards CAs* evoke positive feelings about their learning performance. Hence, CA support might help pupils with positive expectations to reduce stress and anxiety after tests and thus have more psychological benefits than performance benefits.

Third, the *frequency of CAs features use (knowledge or quiz function) did not foster actual or perceived learning outcomes nor were they found to be the reason why positive expectations towards CA were associated with better learning performance.* Our non-findings keep the conundrum on the role of content-related design features of CAs open. While we did not find that content-related design features lead to better learning outcomes, we found a positive association between pupils that preferred to use CAs due to their juvenile language and their actual test scores. Thus, emotion-focused design features might play a more important role than content-oriented design features. This implies that research on pedagogical CAs should consider both content-related and emotion-focused design features when studying CA learning effectiveness. In this study, we focused on the quantity of use of content-related design features. Yet, the quality of the CA usage might be an even stronger predictor than the quantity of CA interactions.

7. Conclusion, Future Research, and Limitations

The aim of this study was to investigate how the use frequency of pedagogical CA features and pupils' expectations towards CAs affect their learning outcomes. We found that the use of pedagogical CAs could in fact improve learning outcomes and that learners' expectations are drivers of their subjective belief of better performance. However, the role of design feature usage remains unclear, requiring future research. It could be interesting to explore if usage patterns of chatbot-based learning and traditional learning materials differ. Our anecdotal evidence suggests that mind-wandering or getting distracted is a common experience with schoolbooks. Yet, we have experienced less mind-wandering and distractions when using, for example, language apps such as Duolingo or Babbel. It might be that chatbots foster more effective and attentive learning strategies given their didactic models in the background than traditional methods. Therefore, we encourage future research to compare pedagogical CAs not only against the use of traditional learning materials such as books or power point slides, which was the focus of our study, but explore other interactive technology-based learning aids such as learning games, apps, or gamified quizzes.

Moreover, we consider the further investigation of expectations towards CA as a fruitful avenue for future research. In this study, we theorized that an underlying placebo effect is the causal mechanism for subjective learning achievement, without directly testing it. Given our finding that expectations drive subjective learning achievement, future research could dive deeper into the underlying causes. In addition to comparing pedagogical CAs with a baseline condition, future research could also employ multiple treatment conditions that decompose several aspects of pedagogical agents, such as calendar and reminder functions, the personal addressing of pupils, or a personalized learning model.

Furthermore, while the goal of our research was to open the black box of CAs and explore the effectiveness of using different CA features, our IT artefact was rather simple. For example, we pre-programmed the responses of the CA, instead of relying on AI techniques. Future research could employ more sophisticated CAs, that could detect and accommodate the pupils' skill level and emotional states, e.g., feeling attentive or fatigued and provide personalized learning content and test material. However, with more advanced CAs it is important to bear potential cost, implementation, and data privacy issues in mind that are associated with autonomous technology that stores and shares sensitive information such as a person's emotional or cognitive states.

Several limitations should be considered when assessing the contributions of this study. First, the sample size was small but comparable with other related studies [24], [39]. A larger sample size might have detected direct or even mediation effects. Our measures to motivate more students to participate and take part in the survey were modest in order to not overburden students and teachers. A repetition of the study was considered in the following year, but due to the COVID19-pandemic not conducted. Therefore, we encourage future research to conduct similar studies focusing on CA design features and expectations towards CA with larger sample sizes as well as in different settings, including different schools or higher education institutions.

Second, the study was conducted in six classes and in the 9th and 10th grades. Hence, some variability of the data may stem from the specific learning unit. We mitigated this problem by keeping the programming and test creation with one person; the first author. A statistical test if the class influenced the learning outcome measures was not significant. Nonetheless, future research could reduce the inherent variability in the data by exploring larger class settings where everyone is exposed to the same learning content and completes the same assessment.

Third, the survey construct to measure CA expectations was self-developed and thus cannot yet be considered a mature measurement scale. The corresponding reliability and validity tests were acceptable, which increases our trust in the measurement of the concept. Since, CA expectations played an important role for perceived learning achievement, we encourage future research to further investigate technology expectancy as a placebo.

References

[1] S. Hobert and R. M. Von Wolff, "Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents," *14th Int. Conf. Wirtschaftsinformatik*, pp. 301–314, 2019.

[2] A. S. D. Martha and H. B. Santoso, "The design and impact of the pedagogical agent: A systematic literature review," *J. Educ. Online*, vol. 16, no. 1, 2019, doi: 10.9743/jeo.2019.16.1.8.

[3] R. Winkler and M. Söllner, "Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis," 2018, [Online]. Available: https://www.alexandria.unisg.ch/254848/1/JML_699.pdf.

[4] J. Q. Pérez, T. Daradoumis, and J. M. M. Puig, "Rediscovering the use of chatbots in education: A systematic literature review," *Comput. Appl. Eng. Educ.*, no. August, pp. 1–17, 2020, doi: 10.1002/cae.22326.

[5] K. Crockett, A. Latham, and N. Whitton, "On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees," *Int. J. Hum. Comput. Stud.*, vol. 97, pp. 98–115, 2017, doi: 10.1016/j.ijhcs.2016.08.005.

[6] D. R. Sanchez, M. Langer, and R. Kaur, "Gamification in the classroom: Examining the impact of gamified quizzes on student learning," *Comput. Educ.*, vol. 144, p. 103666, 2020, doi: https://doi.org/10.1016/j.compedu.2019.103666.

[7] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, "Auto tutor: An intelligent tutoring system with mixed-initiative dialogue," *IEEE Trans. Educ.*, vol. 48, no. 4, pp. 612–618, 2005, doi: 10.1109/TE.2005.856149.

[8] Y. Fang, Z. Ren, X. Hu, and A. C. Graesser, "A meta-analysis of the effectiveness of ALEKS on learning," *Educ. Psychol.*, vol. 39, no. 10, pp. 1278–1292, 2019, doi: 10.1080/01443410.2018.1495829.

[9] N. L. Schroeder, O. O. Adesope, and R. B. Gilbert, "How effective are pedagogical agents for learning? a meta-analytic review," *J. Educ. Comput. Res.*, vol. 49, no. 1, pp. 1–39, 2013, doi: 10.2190/EC.49.1.a.

[10] J. A. Kulik and J. D. Fletcher, "Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review," *Rev. Educ. Res.*, vol. 86, no. 1, pp. 42–78, 2016, doi: 10.3102/0034654315581420.

[11] S. Steenbergen-Hu and H. Cooper, "A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning," *J. Educ. Psychol.*, vol. 105, no. 4, pp. 970–987, 2013, doi: 10.1037/a0032447.

[12] W. Ma, O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent Tutoring Systems and Learning Outcomes : A Meta-Analysis," *J. Educ. Psychol.*, vol. 106, no. 4, pp. 901–918, 2014.

[13] N. Wellnhammer, M. Dolata, S. Steigler, and G. Schwabe, "Studying with the Help of Digital Tutors: Design Aspects of Conversational Agents that Influence the Learning Process," *Proc. 53rd Hawaii Int. Conf. Syst. Sci.*, vol. 3, pp. 146–155, 2020, doi: 10.24251/hicss.2020.019.

[14] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions," *Comput. Human Behav.*, vol. 85, pp. 183–189, 2018, doi: 10.1016/j.chb.2018.03.051.

[15] V. Venkatesh, M. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Q.*, vol. 27, no. 3, pp. 425–478, 2003, [Online]. Available: http://www.jstor.org/stable/30036540.

[16] M. Alavi and D. Leidner, "Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues," *MIS Q.*, vol. 25, no. 1, pp. 107–136, 2001, Accessed: Feb. 18, 2014. [Online]. Available: http://www.jstor.org/stable/3250961.

[17] N. L. Schroeder and C. M. Gotch, "Persisting issues in pedagogical agent research," *J. Educ. Comput. Res.*, vol. 53, no. 2, pp. 183–204, 2015, doi: 10.1177/0735633115597625.

[18] Y. R. Guo and D. H. L. Goh, "Affect in embodied pedagogical agents: Meta-analytic review," *J. Educ. Comput. Res.*, vol. 53, no. 1, pp. 124–149, 2015, doi: 10.1177/0735633115588774.

[19] J. Kim, K. Merrill, K. Xu, and D. D. Sellnow, "My Teacher Is a Machine: Understanding Students' Perceptions of AI Teaching Assistants in Online Education," *Int. J. Hum. Comput. Interact.*, no. online first, pp. 1902–1911, 2020, doi: 10.1080/10447318.2020.1801227.

[20] T. Wambsganss, S. Guggisberg, and M. Söllner, "ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback," *16th Int. Conf. Wirtschaftsinformatik*, no. January, 2021.

[21] S. Hobert, "Say hello to 'Coding Tutor'! Design and evaluation of a chatbot-based learning system supporting students to learn to program," *40th Int. Conf. Inf. Syst.*, no. 1, pp. 1–17, 2019.

[22] S. Tegos and S. Demetriadis, "Conversational agents improve peer learning through building on prior knowledge," *Educ. Technol. Soc.*, vol. 20, no. 1, pp. 99–111, 2017.

[23] L. Benotti, M. C. Martínez, and F. Schapachnik, "A Tool for Introducing Computer Science with Automatic Formative Assessment," *IEEE Trans. Learn. Technol.*, vol. 11, no. 2, pp. 179–192, 2018, doi: 10.1109/TLT.2017.2682084.

[24] H. L. Chen, G. Vicki Widarso, and H. Sutrisno, "A ChatBot for Learning Chinese: Learning Achievement and Technology Acceptance," *J. Educ. Comput. Res.*, vol. 58, no. 6, pp. 1161–1189, 2020, doi: 10.1177/0735633120929622.

[25] K. vanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educ. Psychol.*, vol. 46, no. 4, pp. 197–221, 2011, doi: 10.1080/00461520.2011.611369.

[26] L. Waizenegger, I. Seeber, G. Dawson, and K. Desouza, "Conversational Agents - Exploring Generative Mechanisms and Second-hand Effects of Actualized Technology Affordances," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020, pp. 5180–5189.

[27] A. M. Garvey, F. Germann, and L. E. Bolton, "Performance brand placebos: How brands improve performance and consumers take the credit," *J. Consum. Res.*, vol. 42, no. 6, pp. 931–951, 2016, doi: 10.1093/jcr/ucv094.

[28] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "A Taxonomy of Social Cues for Conversational Agents," *Int. J. Hum. Comput. Stud.*, vol. 132, pp. 138–161, 2019, doi: 10.1016/j.ijhcs.2019.07.009.

[29] H. U. I. Liao, D. Liu, and R. Loi, "Looking at Both Sides of the Social Exchange Coin: A Social Cognitive Perspective on the Joint Effects of Relationship Quality and Differentiation on Creativity," *Acad. Manag. J.*, vol. 53, no. 5, pp. 1090–1109, 2010, doi: 10.5465/amj.2010.54533207.

[30] A. Hassanzadeh, F. Kanaani, and S. Elahi, "A model for measuring e-learning system success in universities." pp. 10959–10966, 2012.
[31] C.-M. Hung, I. Huang, and G.-J. Hwang, "Effects of digital game-based learning on students' self-efficacy, motivation, anxiety, and achievements in learning mathematics," *J. Comput. Educ.*, vol. 1, no. 2–3, pp. 151–166, 2014, doi: 10.1007/s40692-014-0008-8.
[32] D. Al-Fraihat, M. Joy, R. Masa'deh, and J. Sinclair, "Evaluating E-learning systems success: An empirical study," *Comput. Human Behav.*, vol. 102, no. March 2019, pp. 67–86, 2020, doi: 10.1016/j.chb.2019.08.004.
[33] G. Ozogul, A. M. Johnson, R. K. Atkinson, and M. Reisslein, "Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions," *Comput. Educ.*, vol. 67, pp. 36–50, 2013, doi: 10.1016/j.compedu.2013.02.006.
[34] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis: a global perspective*. Pearson Education, 2010.
[35] M. Kozak and H. P. Piepho, "What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions," *J. Agron. Crop Sci.*, vol. 204, no. 1, pp. 86–98, 2018, doi: 10.1111/jac.12220.
[36] G. D. Garson, *Testing statistical assumptions: Blue Book Series*. Asheboro, NC, 2012.
[37] R. Bakeman, "Recommended effect size statistic for repeated measures designs," *Behav. Res. Methods*, vol. 37, no. 3, pp. 379–384, 2005, [Online]. Available: https://link.springer.com/content/pdf/10.3758%2FBF03192707.pdf.
[38] C. O. Fritz, P. E. Morris, and J. J. Richler, "Effect size estimates: Current use, calculations, and interpretation," *J. Exp. Psychol. Gen.*, vol. 141, no. 1, pp. 2–18, 2012, doi: 10.1037/a0024338.
[39] W. Huang, K. F. Hew, and D. E. Gonda, "Designing and evaluating three chatbot-enhanced activities for a flipped graduate course," *Int. J. Mech. Eng. Robot. Res.*, vol. 8, no. 5, pp. 813–818, 2019, doi: 10.18178/ijmerr.8.5.813-818.

Appendix

Table 2: Variables in the study

Variables	Items and description
Achieved test score	Represents the number of points the student achieved in the test ($TestPts_{CB}$, $TestPts_{noCB}$) after a study phase with or without the CA
Perceived learning achievement (7-pt Likert Scale, Cronbach's Alpha = 0.94) (adapted from [30]–[32])	I could learn better with a CA than with the traditional learning materials.
	I could prepare better with a CA than with the traditional learning materials.
	I could achieve better results with a CA than with the traditional learning materials.
	I could achieve more points when learning with the CA than with the traditional learning materials.
	I had a better feeling when learning with the CB than with the traditional learning materials.
Perceived CA expectations (7-pt Likert Scale, Cronbach's Alpha = 0.94)	I think that …. CAs will become more important in the future.
	CAs can enrich life.
	Make my life more interesting.
	Will become more and more similar to humans in the way they communicate.
	Can be used successfully in companies.
	Become more and more important in schools.
	Will make life easier.
	I will use a chatbot frequently in the future.
Frequency of theory interaction	Sum of exchanges with the chatbot using the knowledge feature
Frequency of quiz interaction	Sum of exchanges with the chatbot using the quiz feature
Propensity to CA use language	I tend to use a chatbot when it communicates in a youthful language.