International Journal of Artificial Intelligence in Education (2025) 35:791–822
https://doi.org/10.1007/s40593-024-00409-x

ARTICLE

Evaluating the Impact of Learner Control and Interactivity in Conversational Tutoring Systems for Persuasive Writing

Thiemo Wambsganss[1] · Ivo Benke[2] · Alexander Maedche[2] · Kenneth Koedinger[3] · Tanja Käser[4]

Accepted: 25 April 2024 / Published online: 31 May 2024

Abstract

Conversational tutoring systems (CTSs) offer a promising avenue for individualized learning support, especially in domains like persuasive writing. Although these systems have the potential to enhance the learning process, the specific role of learner control and inter- activity within them remains underexplored. This paper introduces *WritingTutor*, a CTS designed to guide students through the pro- cess of crafting persuasive essays, with a focus on varying levels of learner control. In an experimental study involving 96 students, we evaluated the effects of high-level learner control, encompassing con- tent navigation and interface appearance control, against a benchmark version of *WritingTutor* without these features and a static, non- interactive tutoring group. Preliminary findings suggest that tutoring and learner control might enhance the learning experience in terms of enjoyment, ease-of-use, and perceived autonomy. However, these differences are not significant after pair-wise comparison and appear not to translate to significant differences in learning outcomes. This research contributes to the understanding of learner control in CTS, offering empirical insights into its influence on the learning experience.

Keywords Conversational agents · Learner control · Writing support · Design

Introduction

Individual support for learning has not yet been solved in many contemporary learning processes (Kulik & Fletcher, 2016; Sharples, 2023). The gold standard for helping learners acquire and comprehend knowledge is still a personal, human tutor. However, educational institutions, such as high schools or universities, struggle to offer this kind of individual support due to financial and organizational constraints (Seaman et al., 2018). As Winkler et al. (2021) state, "..the growing number of classroom sizes in high schools, mass lectures at universities with more than 100 students per lecturer, and massive open online courses (MOOCs) with more than

Extended author information available on the last page of the article

1000 participants make individual interaction with a teacher or tutor even more difficult" (Winkler et al., 2021). Several studies have found that this lack of individualized support leads to poor learning outcomes, high dropout rates, and unsatisfactory learning experiences (Brinton et al., 2015; Eom et al., 2006; Hone & El Said, 2016).

Conversational tutoring systems (CTSs) are one promising way to address this challenge of scalability and to engage students in meaningful and individual interactions with an artificial tutor (Weber et al., 2021; Han et al., 2023). CTSs are learning tools that communicate with users through dialog-based interfaces using natural language (Weber et al., 2021; Winkler & Söllner, 2018). They have been successfully designed and deployed to accompany learners individually on their learning paths in various domains, such as developing factual knowledge (Ruan et al., 2019), programming skills (Winkler et al., 2020), or problem-solving (Winkler et al., 2021). One exemplary application of leveraging CTS is to individualize learning in writing assignments, e.g., in persuasive writing (Sharples, 2023; Wambsganss et al., 2021a, b). In particular, with the recent introduction of ChatGPT (https://chat.openai.com/), a novel generative pretrained trans- former model with a conversational interface, research and practice have been calling for more in-depth empirical investigations of the influence of CTS on students writing assignments and learning processes (e.g., (Sharples, 2023; Holden Thorp, 2023; Roscoe et al., 2014; Baidoo-Anu & Owusu Ansah, 2023)). Despite the growing research interest, many CTSs still yield only mixed effects on students' learning outcomes and experiences in empirical studies (Weber et al., 2021; Han et al., 2023; Winkler & Söllner, 2018; Følstad & Brandtzæg, 2017; Zierau et al., 2020).

For example, as Weber et al. (2021) mentioned, CTSs "…seem to positively impact affective outcome variables such as motivation or engagement" (Weber et al., 2021), but when it comes to learning performance through CTS, only "…25.0% of all relevant papers identified and analyzed […] significantly higher learning performance" (Weber et al., 2021). As a result, Weber et al. (2021) recommend that researchers and practitioners design CTS more systematically according to relevant pedagogical design dimensions and characteristics (Weber et al., 2021).

In particular, learner control has been shown to enhance learning experiences in interactive learning systems by providing learners with control over the learning process (Soloway et al., 1994). We follow the definition of Jumonville (2012), which characterizes learner control as" instructional strategies through which learners can exercise some level of control over the events of instruction. It means that learners make their own decisions regarding the sequence, pace, flow, amount, and review of instruction" (Jumonville, 2012). Several studies have investigated the effect of increased learner control in interactive learning systems (Brown et al., 2016). For example, a high degree of control of learning feedback has been shown to positively impact skill-based learning outcomes among interviewee-based learners (Brown et al., 2016; Fisher et al., 2017). Lower levels of learning control, in turn, were beneficial for novice students (Carolan et al., 2014; Kraiger & Jerden, 2007). Although dialog-based CTS provides a natural level of control over interaction, research on embedding learner control principles and its effects on learners' experiences and learning outcomes remains limited (Weber et al., 2021; Winkler & Söllner, 2018; Følstad & Brandtzæg, 2017; Brown et al., 2016; Sorgenfrei & Smolnik, 2016).

This study aims to address that gap by investigating how learner control embedded in a CTS impacts students' learning outcomes and experiences through the design of *WritingTutor*, a novel CTS with learner control features. *WritingTutor* guides learners through an essay-writing exercise by providing theoretical instructions, input, and text-based feedback on the essays' objectivity, polarity, and text quality.

We conducted a laboratory experiment with 96 students to investigate the impact of learner control in CTS. In this experiment, students were asked to engage in an essay-writing exercise. Students in treatment group 1 (TG1) used *WritingTutor* with high-level learner-control (i.e., surface- and deep-level) features based on our design studies, including content navigation and interface appearance control. Students in treatment group 2 (TG2) used a version of *WritingTutor* without additional learner control features for the same writing exercises. Students in the control group received static tutoring with no learner control. Based on this study, we tested two hypotheses: (H1) *In comparison to a static tutoring system, an interactive CTS with different levels of learner control improves learning experience and learning outcomes in an essay-writing task;* and (H2) *In comparison to low levels of learner control, high levels of learner control in a CTS improve learning experience and learning outcomes in an essay-writing task*. While H2 is shedding light on learner control features in CTS, H1 is investigating the effects of the design of our interactive CTS *WritingTutor* in comparison to static learning exercises on learning outcomes and learning experiences (e.g., as Weber et al. (2021) have called for). Whereas H2 is based on the diverse literature on learner control, the foundation for H1 is based on the interactive, constructive, active, and passive (ICAP) framework by Chi and Wylie (2014). According to Chi and Wylie (2014) the design of interactive learning environments has a positive impact on a learner's engagement.

Our work makes three contributions. First, we provide design rationales for a CTS system for persuasive essay writing. With *WritingTutor*, we provide a CTS that can be used domain-independently to engage students in an interactive persuasive writing task (without expensive retraining). Our findings shed light on the advantages of conversational tutoring and possibilities with learner control to enhance learning experiences, such as ease-of-use and enjoyment.

Second, our study is one of the first to thoroughly evaluate learner control features embedded into a CTS. Although we see indications of potential positive effects on the perceived learning autonomy of students' having content navigation and interface appearance control when comparing the means with a two-tailed Welch's t-test, our results do not reveal any significant differences after a pairwise comparison. Hence, our study also indicates the limitations of learner control embedded in CTS. Third, our qualitative analysis of user comments reveals that CTS appears to already provide a natural level of control through dialog-based interaction, which may further our understanding of how to design interactive CTS. Future research is needed to dig deeper into the effects of learner control features for different skill levels of students to enhance interactive tutoring across domains.

Related Work and Conceptual Background

Our work is inspired by previous studies on conversational learning systems and CTS, as well as by literature on learner control, which serves as an underlying conceptual model for our hypotheses.

Conversational Tutoring Systems

Conversational user interfaces are software applications designed to interact with users through natural conversation (Han et al., 2023; Shawar & Atwell, 2005). They are increasingly deployed and utilized in multiple user domains such as customer service (Xu et al., 2017), healthcare (Kowatsch et al., 2017; Laumer et al., 2019) or education (Han et al., 2023, Kerly et al., 2007). Weber et al. (2021) define conversational interfaces in education as pedagogical conversational agents. CTSs are special forms of learning systems that can provide learners with individual and personalized interaction (Winkler & Söllner, 2018). CTSs offer a natural and intuitive way for students to express them- selves in computer-based educational settings, such as in large-scale learning scenarios or massive open online courses. This can, in turn, impact the learner experience and learning outcomes (e.g., Ruan et al. (2019), Winkler et al. (2020), Wambsganss et al. (2021a)). CTSs are able to imitate a human-tutor conversation and allow for an adaptive interaction path that goes beyond a static user interaction experience, thereby increasing student engagement and enjoyment (Weber et al., 2021; Winkler et al., 2021). The development of CTS for education dates back to the research stream on intelligent tutoring systems more than 50 years ago (Atkinson & Shiffrin, 1968; Suppes & Morningstar, 1969). Similar to human tutors, the systems were designed to instruct learners, ask questions, or provide immediate feedback (Kulik & Fletcher, 2016). Intelligent tutoring systems have progressed from impersonal abstractions with few technological options to anthropomorphized computer systems capable of interacting with students through a variety of channels, displaying social abilities, and taking on various roles (e.g., a human tutor Payr, 2003; Wambsganss et al., 2021a), or a learning peer (Kim & Baylor, 2008; Winkler et al., 2020)). Especially in artificial intelligence for education (AIED) research, various CTSs have been designed and evaluated for, e.g., argumentation skill learning (Wambsganss et al., 2021a), conducting course evaluations (Xiao et al., 2019), information retrieval (Sumikawa et al., 2019), programming skills (Winkler et al., 2020), problem-solving abilities (Winkler et al., 2021), and mathematical skills (Grossman et al., 2019), or to foster factual knowledge learning (Ruan et al., 2019).

Despite these developments, there is a recognized need for more comprehensive empirical studies exploring the multifaceted effects of CTS on student learning experiences and learning outcomes (Weber et al., 2021; Winkler & Söllner, 2018; Kuhail et al., 2023). In fact, we see a lack of empirical studies that consider the interplay between different elements of CTS, including personalized feedback, dialogue flow, and learner control, and their collective impact on learning experiences and outcomes. The current body of literature indicates that the effects of CTS on students' learning outcomes and experiences have been mixed. As pointed out by Weber et al. (2021), many existing CTS designs are not grounded in solid evidence or learning theory paradigms. Although the studies have reported positive impacts on affective outcome variables such as motivation or engagement, only 25.0% of the papers analyzed found significantly higher learning performance linked to CTS use (Weber et al., 2021). Consequently, the call for systematic designs for CTS grounded in pedagogical design dimensions and characteristics has grown louder (Kuhail et al., 2023). Hence, in our paper, we investigate the pedagogical concept, namely learner control, in a general conversational-learning scenario based on our conversational agent, *WritingTutor*.

Learner Control for Interactive Learning

One paradigm to enhance learning in certain scenarios is to provide learners with control over the learning process (Brown et al., 2016; Fisher et al., 2017). Past research has distinguished between two dimensions of learner control: 1) no learner discretion (con- trolled by educational characteristics), also called program control (Brown et al., 2016; Sorgenfrei & Smolnik, 2016; and 2) learner discretion. In the case of no learner discretion, the "program" (tutor) guides and instructs learners through exercises and determines the pace, sequence, content, timing, and feedback given to the learner (Sorgenfrei & Smolnik, 2016). Because the program is in control of the learning process, this approach is called the "instruction-centric" view. Complete learner discretion is a substantive learning process that sees the learner as the center of the learning process, which they can control. Therefore, this approach is termed "learner-centric" (Brown et al., 2016). Merging these two strands, Brown et al. (2016) outlines, "*learner control can be conceptualized along a continuum from completely instruction-controlled (i.e., no learner discretion) to completely learner controlled (i.e., complete learner discretion)*" (Brown et al., 2016).

Learner control in digital learning scenarios usually becomes inherent in specific design features. These features can be categorized into five different dimensions: 1) control of timing of instructional events, 2) control of the physical environment of instructional events, 3) control of information dis- play and processing capabilities, 4) control of the type and degree of contact between the participants, and 5) control of resources and content available to the learner (Sorgenfrei & Smolnik, 2016). Hence, several studies have investigated different modalities of learner control in digital learning scenarios. Jung et al. (2019), for instance, investigated the influence of instructional design in MOOCs on learner control and perceived effectiveness with 664 participants. Their results indicated that age and instructional design factors of structure, transactional interaction between students and content, and assessment are stronger indicators for learner control than content itself. Chen et al. (2021) conducted a study where students were able to control the pace during video instructions in online learning. Among other outcomes, they assessed intrinsic motivation, learning satisfaction, perceived learning and transfer learning in a post-test. They found that the learner control features in the form of controlling the pace of video instructions, as opposed to program control, do increase factual learning outcomes for learners with negative emotions, but do not influence motivation, learning satisfaction, or other perceived measures. Detailed literature reviews on learner control in digital learning scenarios can be found in Sorgenfrei and Smolnik (2016) or Brown et al. (2016).

The existing literature on learner control features in digital learning has produced mixed results, with some studies reporting positive effects and others negative effects (Brown et al., 2016; Sorgenfrei & Smolnik, 2016). However, most research has been conducted in non- adaptive and static interaction paradigms, e.g., when reading a text, receiving instructions, or watching a video (Sorgenfrei & Smolnik, 2016; Weber et al., 2021). Although a digital learning interface with learner control elements gives users the opportunity to use them (i.e., the objective degree of learner control), this does not guarantee that users will do so (i.e., real learner control) (Brown et al., 2016; Kraiger & Jerden, 2007). Additionally, the existence of learner control features may have a negative impact on learning outcomes, for example, due to an increase in cognitive load (Brown et al., 2016; Orvis et al., 2009).

To sum up, past work suggests that research must take an interactive approach to the role of learner control in digital learning by examining the qualities of the learning exercises as well as those of the learner, and how well they match to learner control (Fisher et al., 2017). Design-related research in the field of AIED about these specific digital learning scenarios is rather rare. As Fisher et al. (2017) mentioned a research gap in studies about *“optimal degree of learner control”* and the lack of research on *“how the degree of learner control determines e-learning effectiveness”* (Fisher et al., 2017). In particular, when it comes to more modern learning scenarios (for instance, with interactive learning tools based on ML), past literature on learner control is scarce indeed. In fact, we have not found a single study from recent years that investigated the effect of learner control on learning experience and learning outcomes in ML-based conversational tutoring.

Hypothesis Development About Learner Control in Conversational Tutoring Systems

The combination of both literature streams, e.g., to design novel conversational learning scenarios where students have different forms of surface- and deep-level control has been rarely investigated by past works. We believe that higher-levels of learner control could increase the interactivity of the learning exercises. We know from existing research on conversational user interfaces, that it can be beneficial for users to have control over the conversational flow (e.g., Benke et al., 2022). We base this hypothesis on the ICAP framework by Chi and Wylie (2014). Accordingly, the design of interactive vs active learning environments has positive impacts on a learner’s engagement with the learning materials; the same accounts for changing the interaction “from passive to active, constructive, to interactive” (Chi & Wylie, 2014). In contrast to passive engagement, which involves students merely receiving or consuming learning materials (such as reading a text), active engagement involves students actively managing the presentation of the subject (e.g., by highlighting important text paragraphs). Students expand their contact in the two types of interaction that Chi and Wylie (2014) described as the most engaging ones, for example, comparing the learning materials with prior knowledge (constructive engagement), engaging in debate with others, or posing and responding to questions (interactive engagement). Each ICAP mode predicts a different learning result by correlating with various types of behaviors and knowledge-change processes (Chi & Wylie, 2014). Hence, based on prior literature on CTS, learner control, and the ICAP framework, we derive the following hypothesis:

> H1: In comparison to a static tutoring system, an interactive CTS with different levels of learner control improves learning experience and learning outcomes in an essay-writing task.

Following this hypothesis and framework, interactive CTS with learning control could help to foster the students' engagement as they not only add dialoguing but also the control over the dialogue or the appearance of the CTS to the learning experience. Based on prior literature on learner control and the ICAP framework, we derive the following second hypothesis:

> H2: In comparison to low levels of learner control, high levels of learner control in a CTS improve learning experience and learning outcomes in an essay-writing task.

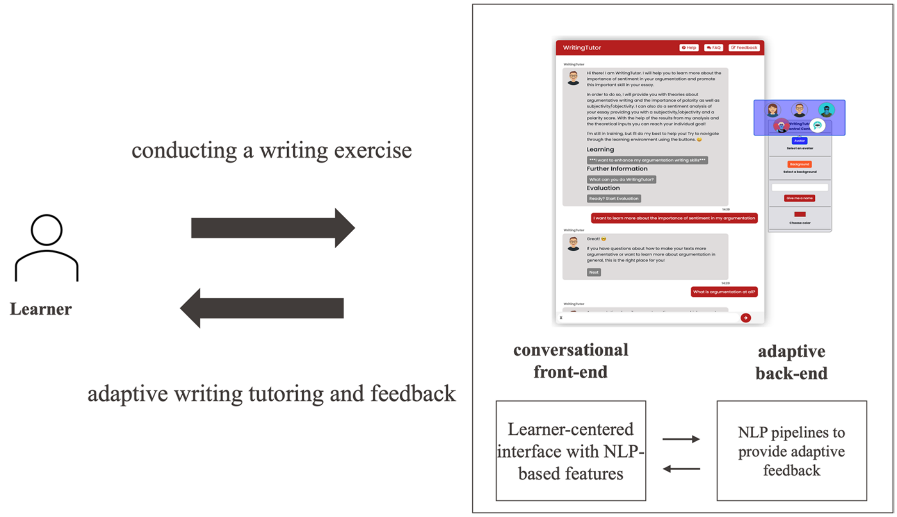
By investigating these hypotheses, we answer the call of Fisher et al. (2017) to "examine the dimensionality and meaning of various learner control features to help move the field forward both in theory and in practice." (Fisher et al., 2017). Hence, with our research, we aim to narrow this literature gap and investigate the influence of dialogue-based deep-level control (e.g., having control over the conversational learning process with a CTS vs. not having control over the conversational learning process) and surface-level control (e.g., personalizing the appearance features of dialogue-based learning tools vs. not personalizing the appearance features of dialogue-based learning tools) on students' learning outcomes and experiences.

Design of a Conversational Tutoring System with Learner Control Features

To investigate 1) the effects of an interactive CTS on students' learning experiences and learning outcomes and 2) the differences between low levels of learner control and high levels of learner control in a CTS, we designed a CTS called *WritingTutor*.

WritingTutor guides students through an essay-writing exercise and offers learners the opportunity to control the learning path sequence as well as the interface design appearance. *WritingTutor* consists of three main components:

1) a learner-centered conversational user interface, 2) distinct learner control features, and 3) an NLP and ML back end to intelligently guide the students through a natural tutoring conversation and provide adaptive feedback (i.e., in the form


conducting a writing exercise
Learner
adaptive writing tutoring and feedback
WritingTutor
Help
FAQ
Feedback
WritingTutor
Learning
I want to enhance my argumentation writing skills
Further Information
What can you do WritingTutor?
Evaluation
Ready? Start Evaluation
14:19
WritingTutor
Next
14:20
WritingTutor
WritingTutor Control Center
Avatar
Select an avatar
Background
Select a background
Give me a name
Choose color
conversational front-end
adaptive back-end
Learner-centered interface with NLP-based features
NLP pipelines to provide adaptive feedback

Hi there! I am WritingTutor. I will help you to learn more about the importance of sentiment in your argumentation and promote this important skill in your essay.

In order to do so, I will provide you with theories about argumentative writing and the importance of polarity as well as subjectivity/objectivity. I can also do a sentiment analysis of your essay providing you with a subjectivity/objectivity and a polarity score. With the help of the results from my analysis and the theoretical inputs you can reach your individual goal!

I'm still in training, but I'll do my best to help you! Try to navigate through the learning environment using the buttons. 😀

I want to learn more about the importance of sentiment in my argumentation

Great! 🤓

If you have questions about how to make your texts more argumentative or want to learn more about argumentation in general, this is the right place for you!

What is argumentation at all?

Fig. 1 Overview of the design architecture of *WritingTutor*: A learner engages in an essay- writing exercise and is adaptively tutored with theory input and feedback through the learning exercises. The user is able to control **a**) the content (through pre-defined buttons, thus being able to control the pace and the instructional input of the learning exercise), and **b**) the interface appearance of *WritingTutor* (avatar, name, main color theme, background)

of individual recommendations) on the text quality of their essays. The basic interaction concept of *WritingTutor* is illustrated in Fig. 1.[1]

Requirements for Learner Control in Conversational Tutoring Systems

We aimed to build on the learner control literature to equip *WritingTutor* with a high level of learner control during the conversational tutoring exercise. To build a theory-infused CTS supporting learner control, we followed Cooper (1988) to conduct a literature review with the aim of deriving a set of theory requirements and design principles for learner control features. We searched various databases with keywords such as ("learner control" OR "student control") AND ("e-learning" OR "technology" or "computer-supported"). We found 65 papers from the fields of psychology, education, and educational technology that have investigated various traits of learner control features in e-learning exercises (e.g., (Brown et al., 2016; Fisher et al., 2017). A summary of the most relevant articles can be found in the related work section. Detailed literature reviews on learner control in digital learning scenarios can be found in Sorgenfrei and Smolnik (2016) or Brown et al. (2016).

Having reviewed the selected literature, we aggregated similar study topics as literature issues, of which we formed two clusters; these served as design requirements for the design of learner control features in our CTS. The theory

[1] We provide interaction videos of *WritingTutor* in the supplementary material of this submission.

requirements were then instantiated and iteratively refined in the course of our study. Based on prior research, we aimed to equip *WritingTutor* with a feature that allows students to control the content sequence as well as the pace of the learning path. Controlling the learning pace and the content in e-learning exercises has been found to increase students' motivation and engagement (Brown et al., 2016; Sorgenfrei & Smolnik, 2016). Hence, we equipped *WritingTutor* with a sequence of pre-defined buttons, where students could skip certain learning steps (e.g., the theory explanations, or jump directly from and to different steps in the interaction path). Providing learners with the ability to control the learning path, content, and pace is considered a deep-level control feature (Brown et al., 2016; Fisher et al., 2017).

Moreover, based on prior studies, higher-level learner control, such as con- trolling superficial elements of the appearance of an e-learning tool, has been found to increase user experience. For instance, Behrend and Thompson (2012) found that giving students control over program design improved learning outcomes in research on students' Microsoft Excel courses. Studies typically suggest that this has favorable implications on cognitive results when learners have control over numerous parameters (i.e., high learner control). This is also reflected in literature on gamification (e.g., Schöbel et al., 2021). Enabling students to customize their "own" learning tool can create enjoyment and motivation, ultimately leading to increased engagement (Schöbel et al., 2021). Hence, we developed an interface design panel next to the chatbot to enable students to control the interface appearance of *WritingTu- tor*. We gave students control over the 1) avatar, 2) name, 3) general color scheme, and 4) background picture (see F3 in Fig. 3).

Next, our objective was to collect general requirements from learners for the design of a CTS for a typical persuasive essay-writing task. To do so, we conducted twelve semi-structured interviews with students to build an initial understanding of learners' needs and requirements with regard to essay-writing exercises in general and conversational tutoring tools in particular (Gläser & Laudel, 2010). Our interview book consisted of 28 questions. We asked the students about their experiences with conversational user interfaces in general and in education. Moreover, we asked specifically about the requirements of a conversational tutoring tool and the kind of instructions, support, and feedback they would like to receive in a typical essay-writing exercise (and that would make them more likely to continuously use such a tool). The interviewees were a subset of students at a university, all of them potential users of a conversational essay tutoring system. The interviewed students had a mean age of 25.5 years (SD = 1.95), and all were pursuing bachelor's or master's degrees in economics or law. Nine were male and three were female. The interview transcripts were evaluated using qualitative content analysis to code the data and form abstract categories. Open coding was used to form a uniform coding system during the evaluation (Gläser & Laudel, 2010). We derived several user stories in each interview and aggregated the most common ones following (Cohn, 2004), yielding ten common topics.

The user stories contained requirements for the user interface design, conversational flow, input formats, adaptive feedback types and illustrations, level of anthropomorphization, and efficient usage of a conversational tutoring tool. All in all, the interviews, as well as the user stories, gave us an overview of what users need and require from a CTS for essay-writing exercises.

The first user story was derived from students' need to have a familiar and easy-to-use layout. They described several existing layout structures such as those of WhatsApp or Facebook Messenger, which should be used as a template. The interviewees' statements also align with the literature, which reports better user experiences when familiar concepts and layouts are used (Feine et al., 2019; Zierau et al., 2020). Moreover, the interviewees wanted to have a guided learning experience. This desire stems from the fact that because individual interviewees' learning behaviors are often chaotic, they would like to see an increase in efficiency from a guided and tutored learning process. The desire for a guided learning experience is also corroborated by the existing literature (Roth, 1970; Wambsganss et al., 2021a; Weber et al., 2021). Moreover, almost all students stated that they would like to receive feedback on their written texts as well as integration of theoretical content into the learning process. The integration of learning materials should make it possible to react to the received feedback as, according to the interviewees, this would also make the feedback more comprehensible. Another user story describes the interviewees' wish that not only plain text, but also other media elements should be embedded in the tutoring tool. This involved the integration of learning videos as well as the integration of external sources to supplement theoretical content. During the interviews, the interviewees also expressed the wish that, ideally, a count function of the written text should also be displayed. This is because in the learners' experiences, writing exercises often come with a maximum word limit. Next, interviewees expressed the desire for a tool that encourages and supports them in an iterative learning process. Concretely, the learner should be able to access the learning material and also start the feedback process at any time. This should ensure maximum flexibility in using the CTS. Moreover, the feedback should be presented in a single dashboard overview. The students expressed a desire to have the option of contacting a real human tutor during the learning interaction in case they need assistance with more complex questions. Finally, the interviewees wanted a CTS that could talk like a learning peer but, at the same time, be able to deliver important learning content in a serious way.

All in all, the interviews, as well as the user stories, gave us an overview of users' needs and requirements of a CTS for essay-writing exercises.

User Interaction of *WritingTutor*

Following the requirements enumerated within the literature and by users, *WritingTutor* is built as a responsive web-based application that can be used on all kinds of devices. A screenshot of *WritingTutor* and its core function- alities (e.g., *F1—F5*) can be seen in Fig. 3). *WritingTutor* consists of two modes: 1) adaptive theory and instruction mode and 2) essay writing and feedback mode. The interaction flow of *WritingTutor* with the two interaction modes can be seen in Fig. 2.

The interaction starts with *WritingTutor* providing the user with an overview and explanation of the upcoming essay-writing task and the various functions. Right from the beginning, *WritingTutor* provides students with an overview of predefined

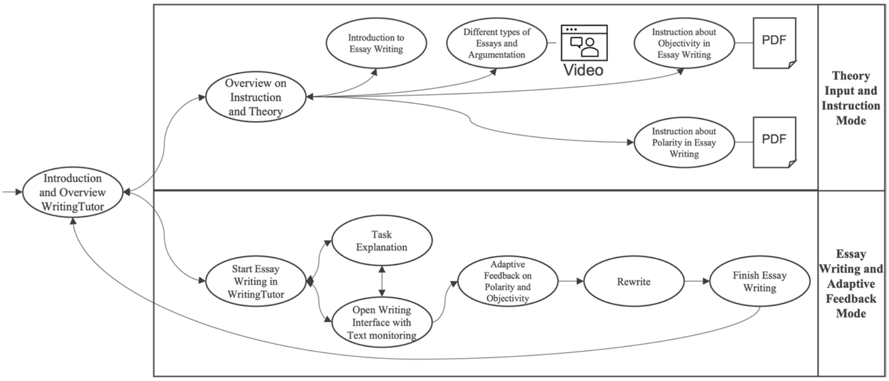

Introduction and Overview WritingTutor
Overview on Instruction and Theory
Introduction to Essay Writing
Different types of Essays and Argumentation
Video
Instruction about Objectivity in Essay Writing
PDF
Instruction about Polarity in Essay Writing
PDF
Theory Input and Instruction Mode
Start Essay Writing in WritingTutor
Task Explanation
Open Writing Interface with Text monitoring
Adaptive Feedback on Polarity and Objectivity
Rewrite
Finish Essay Writing
Essay Writing and Adaptive Feedback Mode

Fig. 2 Overview of the two different interaction modes for *WritingTutor*: 1) theory input and instruction mode and 2) essay writing and adaptive feedback mode. Learners can switch between the modes anytime they want, reflecting a high level of learner control

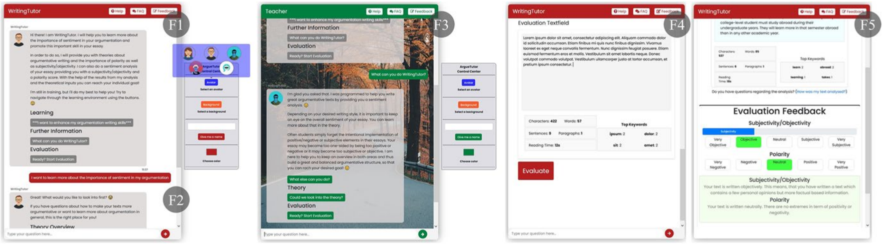


Fig. 3 Screenshot of our conversational tutoring system *WritingTutor:* a user conducts an essay-writing exercise and receives interactive tutoring that includes individual feedback on their written texts. Different learner control features enable the learner to control the content and the appearance of the interface

buttons that enable them to directly jump to different parts of the learning exercises, granting them a high level of learner control as explained in the previous section (see F1 in Fig. 3). *WritingTutor* then guides the learner through an essay-writing exercise with the aim of imitating a human tutor. The learner interaction starts with a lecture on persuasive essay writing, including videos and reading exercises about the typical structure of an essay, logical argumentation, and the importance of taking a certain standpoint.

In every part of the *theory input mode*, as well as in the *essay writing and feedback mode*, the user is able to control the pace and the sequence of the learning path and can skip certain elements. Moreover, the learner can always modify the appearance of *WritingTutor* according to their preferences.

After the theory lecture is finished, *WritingTutor* asks the user whether they would like to revisit any content or practice essay writing. *WritingTutor* gives the user a typical essay-writing task based on the ETS General Writing Exam (GRE (https://www.ets.org/gre)), such as *"Every country in the world has problems with pollution and damage to the environment. Do you think these problems can* *be solved? Evaluate the question within a 200 to 250-word text about the pros and cons."* The user is now able to elaborate on their standpoint on the given topic in a text input field directly embedded in the CTS (see F4 in Fig. 3). Based on the user requirements mentioned before, *WritingTutor* provides the learner with an always-on self-monitoring of general text metrics (such as the number of words, number of sentences, reading time, and top keywords in their text).

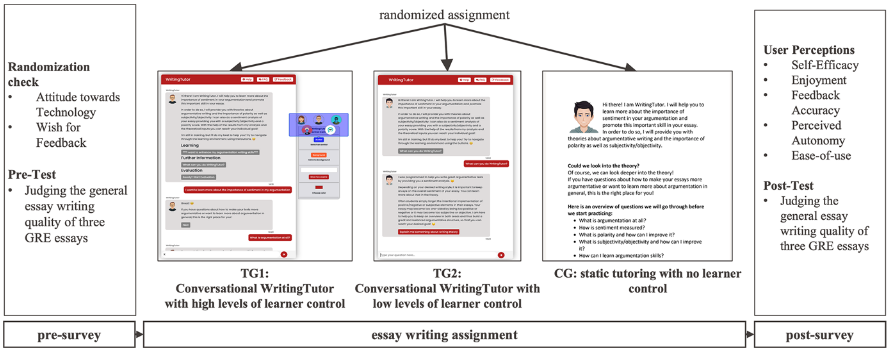
If the user has finished writing, they can click an "evaluate" but- ton to receive immediate feedback and a summary of the standpoint of their essay. The feedback dashboard provides an overview of a) the objectivity of the users' essays and b) the polarity of their texts. Together with an adaptive feedback prompt for both categories, the user receives adaptive scaffolding on the objectivity/subjectivity of their essay and the standpoint taken. Both factors are important factors for general persuasive essay writing.

Furthermore, the user always has the option to a) receive help, b) view frequently asked questions about the learning tool and the learning exercise, or 3) provide feedback about their interaction with *WritingTutor* (see Fig. 3). The three functions are incorporated as always-visible buttons in the right- top corner of *WritingTutor* (see Fig. 3). In this vein, the user can always provide feedback about wrongly classified intents, or to receive assistance if the user modeling is ever wrong.

Moreover, *WritingTutor* incorporates a user-initiated casual chat function. To imitate the students' having a personal learning session with a human tutor, we incorporated several more functions to incorporate real-world conversational elements into *WritingTutor*'s design. For example, we provided a wide variety of different responses to common conversation states (e.g., "how are you?") as well as positive reinforcement feedback typical of a study partner.

Back End of *WritingTutor*

For the back-end functionality of *WritingTutor*, we utilized a combination of different NLP- and ML-based techniques. In general, *WritingTutor* is built as a web app in HTML5 with CSS and JavaScript. The front end is connected to a Python script that a) processes incoming user intents and b) provides pre- defined answers based on the incoming classifications. For the conversational logic of *WritingTutor*, our aim was to model a common student tutoring session. To do so, we used a "Naive Bayes classifier" in combination with semantic similarity matching, as has been done in, e.g., (Ruan et al., 2019). The Naive Bayes classifier is a probabilistic machine learning algorithm that is often used for text classification tasks. In *WritingTutor*, it was trained to classify user intents based on the text of their input. The algorithm works by calculating the probability of each intent given the words in the input, then selecting the intent with the highest probability as the classification. To train the classifier, a dataset of user inputs and their associated intents were modeled. In total, we created 62 intents, including the introduction, the theory input, and the casual dialogue of a student tutoring session on persuasive essay writing with our dataset. The Naive Bayes classifier was trained on a training set using the text of the inputs as features and the intents as labels.


randomized assignment
Randomization check
• Attitude towards Technology
• Wish for Feedback
Pre-Test
• Judging the general essay writing quality of three GRE essays
TG1:
Conversational WritingTutor with high levels of learner control
TG2:
Conversational WritingTutor with low levels of learner control
CG: static tutoring with no learner control
User Perceptions
• Self-Efficacy
• Enjoyment
• Feedback Accuracy
• Perceived Autonomy
• Ease-of-use
Post-Test
• Judging the general essay writing quality of three GRE essays
pre-survey
essay writing assignment
post-survey

Hi there! I am WritingTutor. I will help you to learn more about the importance of sentiment in your argumentation and promote this important skill in your essay. In order to do so, I will provide you with theories about argumentative writing and the importance of polarity as well as subjectivity/objectivity.

Could we look into the theory?
Of course, we can look deeper into the theory!
If you have questions about how to make your essays more argumentative or want to learn more about argumentation in general, this is the right place for you!

Here is an overview of questions we will go through before we start practicing:
- What is argumentation at all?
- How is sentiment measured?
- What is polarity and how can I improve it?
- What is subjectivity/objectivity and how can I improve it?
- How can I learn argumentation skills?

Fig. 4 Overview of the experimental setup of our study. Students in treatment group 1 (TG1) received conversational tutoring with high levels of learner control with *WritingTutor* during an essay-writing assignment. Participants in treatment group 2 (TG2) used *Writ- ingTutor* with low levels of learner control for the exact same task. Students in the control group (CG) receive static tutoring with no learner control

The model input of the Naive Bayes classifier is the text of the user input, which is processed by the classifier to output a predicted intent. The output of the classifier is the predicted intent label, which is used to retrieve the appropriate predefined answer from the database and deliver it to the user.

The conversational back end and the classifier are implemented by utilizing the frameworks *chatterbot* (ChatterBot, n.d.) and *spacy*. The confidence level of the prediction is set to 70%. Moreover, as mentioned above, we implemented several other NLP-based functions with *spacy*, such as a tokenizer to display to the user the number of words. The polarity and objectivity feedback are calculated using *the textblob library* (TextBlob, n.d.).

Experimental Setup

To investigate the effects of CTS with different levels of learner control on students' learning experience and evaluate the impact of high-level learner control features in CTS, we designed an experiment in which participants were asked to conduct a persuasive essay-writing exercise. Participants were randomly assigned to two treatment groups and one control group (see Fig. 4). Treatment group 1 (TG1) used *WritingTutor* featuring high levels of learner control to do the writing exercise, whereas participants in treatment group 2 (TG2) used *WritingTutor* with no extra levels of learner control; participants in the control group received static tutoring and theory input. In the following, we will describe the experimental design. The results are then analyzed in a hierarchical evaluation, where we compare CG vs. TG1 and CG vs. TG2 for H1, and TG1 and TG2 for H2.

Participants

We recruited 99 students from our university to take part in the experiment. The experiment was conducted as a web experiment facilitated by a laboratory for behavioral experimental science and was therefore designed and approved by the ethics committee of the lab and the hosting university. After randomization and removal of failed attention checks, we counted 96 valid results overall with 32 participants in each group (TG1, TG2, CG).[2] Participants of the TG1 had an average age of 23.86 years (SD = 2.95); 13 were female, 19 were male. Participants in TG2 were on average 23.47 years old (SD = 2.99); 13 were female, 19 were male. In the control group, the average age was 23.38 years (SD = 2.64); 9 were female, 23 were male. All participants were compensated with an equivalent of about USD 12 for a 25- to 30-min experiment.

Manipulations

To control for the differences and similarities in the design among the three different groups, we also designed and implemented the alternative tools our- selves. For TG2, we used our design of *WritingTutor* and deactivated both learner control features such as the surface-level control (aka the interface appearance dashboard; see F3 in Fig. 3) and the deep-level learning path control (aka the content navigation buttons; see F1 in Fig. 3). To ensure consistency in the conversational tutoring, all other functions of *WritingTu- tor* were the same in TG2 as in TG1. In both treatment groups, learners received conversational tutoring as well as adaptive writing feedback on their essay-writing assignments. For the control group, we used a traditional bench- mark approach, implementing static tutoring through an online platform. This involved providing students with digital worksheets that included instructions on the theory, an overview of the necessary input, and identical video content on persuasive essay writing as was incorporated in *WritingTutor*. For further clarification, it is important to understand that this control group scenario is akin to a simple online tutoring tool. It provided exactly the same content on the instructional topics as its conversational counterparts in treatment groups TG1 and TG2. However, unlike the dynamically interactive nature of the CTS in TG1 and TG2, the control group's learning environment was static, being offered through a simple online website with anthropomorphized design elements, such as an avatar (see CG in 4). This static tutoring approach included no dynamic conversational interaction. Students were exposed to the content passively and had no opportunity for interactive engagement with the learning materials. Additionally, unlike the TG1 and TG2 groups, the control group received no personalized feedback and were unable to control their learning experience. In essence, the control group received traditional, linear online

[2] We incorporated only complete data points in our sample. Additionally, we tested several control variables to ensure that there are no significant differences between the randomized subsets (explained further below).

instruction without the benefits of the interactive and adaptive features we designed for the treatment groups.

Design and Procedure

The study consisted of three main phases: 1) a pre-survey, 2) an essay-writing assignment, and 3) a post-survey. All subjects underwent the same pre- and post-phases. In the essay-writing phase, the treatment group used *WritingTutor* to conduct a persuasive essay-writing exercise, whereas participants in the control group conducted the same exercise using the alternative tool.

1) Pre-survey (about 5–10 min): Eleven questions made up the pre- survey that preceded the experiment. Here, we evaluated the effectiveness of the randomization using three different constructs. First, we tested each participant's level of personal innovation in the area of information technology using four questions following (Agarwal & Karahanna, 2000). Example items were: *"I like to experiment with new information technologies"* or *"If I heard about a new information technology, I would look for ways to experiment with it"*. Second, we tested the construct of individuals' feedback-seeking following (Ashford, 1986). Example items were: *"It is important for me to receive feedback on my performance.*" or "*I find feedback on my performance useful."* Both constructs were measured with a seven-point Likert scale (1 = totally disagree; 7 = totally agree, with 4 indicating a neutral statement). Third, we constructed a pre-test to capture students' essay-writing assessment qualities. The test is described together with the post-assessment in the next section.

2) Essay-writing assignment (about 10–20 min): In the essay-writing part of the experiments, we asked the participants to engage in a persuasive essay-writing task, simulating a typical homework assignment that students enrolled in higher education should be able to complete within 10–15 min. The essay prompt was: *"Every country in the world has problems with pollution and damage to the environment. Do you think these problems can be solved? Evaluate the question within a 200- to 250-word text about the pros and cons."* TG1 used *WritingTutor* to write the review with higher levels of learner control, according to our design, whereas TG2 used *WritingTutor* without extra learner control features. The control group used a static essay-tutoring reference tool. All students in the treatment groups were adaptively tutored through the writing exercise, e.g., through theory input, individual recommendations, and adaptive feedback based on our feedback algorithm. Participants using *WritingTutor* in TG1 could control the learning sequence and the appearance of the CTS. Students in TG2 were also adaptively tutored; however, they did not have opportunities to control the learning process or adjust the appearance of *WritingTutor* (see Fig. 4).

3) Post-survey (about 5–10 min): In the post-survey, we measured the students' perceived learning experiences using different constructs from the literature. First of all, we measured each student's level of enjoyment during the learning task, which has a major influence on the adoption of IT tools (Lee et al., 2005)

as well as on students' learning success (Pekrun & Stephens, 2012). To this end, we asked the students to rate their agreement or disagreement with the following prompts: *"The interaction with the learning tool was exciting"* and *"It is fun to interact with the learning tool"* (Kim et al., 2019). Also, we captured learning autonomy based on (Jung et al., 2019). The items included *"I had complete control over my learning and writing process in this exercise," "How I experienced completing the task was entirely up to me",* and *"There is nothing that prevented me from experiencing the learning and writing process the way I want to".* Moreover, we measured the perceived ease-of-use of the participants following the technology acceptance model of (Venkatesh & Bala, 2008; Venkatesh et al., 2003) and captured the demographics. The items for ease-of-use were *"Learning how to use the argumentation tool would be easy for me"*, *"I find the argumentation tool easy to interact with,"* or *"It would be easy for me to become skilled in using the argumentation tool."* All post-survey constructs were measured using a seven-point Likert scale (1 = totally disagree; 7 = totally agree, with 4 indicating a neutral statement). Finally, we asked four qualitative questions: *"How has WritingTutor impacted your control over your learning and writing process?"*, "*What did you particularly like about the use of the learning tool?"*, "*What else could be improved?"* and "*Do you have any other ideas?"* In total, we asked fourteen questions.

Pre- and Post-Assessment Test

Understanding students' ability to comprehend and critically assess essay quality is an integral learning outcome in persuasive writing tasks, especially with novice learners (David Smith & Brooker, 1999). Before students are able to master the task of persuasive essay writing itself, it is of the utmost importance that they be able to com- prehend, assess, and understand the core principles of argumentation, essay structuring, and language clarity (David Smith & Brooker, 1999). As highlighted in the related work, the lack of individualized support for learning in general, but also for essay-writing scenarios in specific, is common in many contemporary learning processes and often leads to poor learning outcomes and unsatisfactory learning experiences, also with regards to a fundamental understanding of knowledge components (Brinton et al., 2015; Eom et al., 2006; Hone & El Said, 2016). One of the primary goals of introducing interactivity and learner control mechanisms, as defined by Jumonville (2012) is to empower learners to make decisions regarding their learning process, controlling input or pace, thereby enhancing their engagement and comprehension of knowledge concepts. To measure the impact of interactivity and learner control features on students' ability to comprehend persuasive writing, essay structuring, and language clarity, we aimed to evaluate their skills by asking them to assess essays' general persuasive writing quality. By doing so before and after the intervention, we aimed to determine not only the learners' initial proficiency but also any improvements in their ability to discern the quality of essays.

This, in turn, provides a measure of the effectiveness of the tutoring system in enhancing students' understanding and application of persuasive writing principles, as a core out- come in mastering essay writing. If a student can accurately judge the quality of an essay, it indicates a deeper understanding of the principles of persuasive writing, which is the core focus of our designed learning task as described in "Design and Procedure" section.

To this end, as part of the pre- and post-surveys, we conducted an essay assessment test. Students were asked to judge the essay-writing quality on a given topic using a five-point Likert scale (1 = low essay-writing quality; 5 = high essay-writing quality). The task was to *"please assess the general writing quality of this essay for the above-mentioned topic."*. We chose the ETS General Writing Exam (GRE) as our source for the essays because they resemble typical essay-writing assignments for students and because they are standardized, thus ensuring a consistent quality benchmark. The essay topic, the essays, and the grading were retrieved from the GRE. After the students made their assessments, we compared their assessments of essay quality with the grades given by the ETS GRE test. This comparison allowed us to control whether students were accurately assessing the general writing quality of the essay. In total, we asked all students to assess three GRE essays as part of the pre-test and three GRE essays as part of the post-test. The essays from the pre-test had the same difficulty level as the essays from the post-test. With these results, we calculated the learning gains of each participant using pre- and post-test comparisons.

Results

To present our findings, we follow a hierarchical evaluation set-up as done in Wambsganss et al. (2022). This allows us to shed a nuanced light on the different modalities of interactivity and conversational tutoring (H1) and learner control (H2). First, to validate H1, we compare TG1 with CG and TG2 with CG. Afterwards, we compare TG1 with TG2 to shed light on the learner control features in conversational tutoring and thereby investigate H2. To ensure that the randomization indeed yielded randomized groups, and to control for any potential effects of interfering variables with our sample size, we compared the differences of the construct of individuals' personal innovativeness and feedback-seeking. For both, we received p-values greater than 0.05 between the two treatment groups and the control group by applying ANOVA analysis.

Next, we combined the perceptive measures from our questionnaire as well as the evaluation of the assessment abilities for essay-writing quality before and after the participants' experience with the CTS. To assess the participants' perceptions, we analyzed their responses to the questionnaire items. To assure the internal consistency of latent constructs, we assessed outer factor loadings and Cronbach's alpha with cutoffs at 0.7 and 0.6 (Hair et al., 2021; Van Griethuijsen et al., 2015). Afterward, the scales were mean-scored.

To assess the directed effect of the experimental treatment conditions (TG1, TG2, CG) in our hypotheses, we calculated two-tailed Welch's t-tests to evaluate whether the constructs' means are significantly different. After that, we conducted ANOVA analysis and pair-wise comparisons; we will report on both. In addition

Table 1 Descriptive statistics and analyzed results of perceptive measures and learning outcomes for conversational tutoring with high levels of learner control (TG1) and the static tutoring condition (CG)

Dependent variable	Details	TG1 ($n=32$)	CG ($n=32$)	Analysis results	Bonferroni adjusted p- value
Learning Enjoyment	Mean (SD)	4.875 (1.606)	4.094 (1.542)	t(61.9) = 1.98, $p=0.052$, 95% CI[-0.006, 1.568], d = 0.49	0.206
Learner Autonomy	Mean (SD)	5.521 (0.946)	5.229 (1.092)	t(60.77) = 1.141, $p=0.258$, 95% CI[-0.219, 0.803], d = 0.29	1.0
Ease-of-Use	Mean (SD)	5.552 (1.272)	4.833 (1.071)	t(60.26) = 2.446, $p=0.017$, 95% CI[0.131, 1.307], d = 0.61	0.069
Assessment Quality	Mean (SD)	0.052 (0.605)	-0.052 (0.704)	t(60.64) = 0.635, $p=0.528$, 95% CI[-0.224, 0.432], d = 0.16	1.0

to the hierarchical evaluation set-up, we aim to provide maximum transparency by also providing the analysis of our overall independent variable (the condition, with three values: static tutoring (CG), conversational interaction with higher levels of learner control (TG1), and conversational tutoring with lower levels of learner control (TG2)) as well as the four dependent variables measured in our study (learning enjoyment, learner autonomy, ease- of-use, and assessment quality) in Table 4. Finally, we analyzed the learners' qualitative feedback to gain a deeper understanding of the underlying drivers of their experience.

H1: Results of Conversational Tutoring on Learning Experience and Learning Outcomes

To evaluate our first hypothesis *(H1: In comparison to a static tutoring system, an interactive CTS with different levels of learner control improves learning experience and learning outcomes in an essay-writing task)*, we compared the learner experience by testing perceived ease-of-use, perceived learning enjoyment, and perceived learning autonomy between TG1 and CG participants and then between TG2 and CG participants. For learning outcomes, we evaluated learners' essay-quality assessment abilities by letting them assess the quality of three essays both before and after the learning experience. We assessed learners' performance by computing the absolute deviation of the student's rating from the expert's rating, termed as expert-rating—student-rating. This measure was then subtracted from 5 to create a flipped scale, resulting in the student's grade, formulated as student-grade = 5—expert-rating—student-rating. Thus, a perfect student grade of 5 indicates an alignment with the expert's evaluation, while a lower score reflects a larger deviation from the expert's judgment. Then, we calculated the average performance for the pre-test and post-test separately. The learning gain for each group was computed by subtracting the pre-assessment average from the post-assessment average. A positive value indicates a learning gain. The answers were provided on a five-point Likert scale (1 = low essay-writing quality; 5 = high essay-writing quality). All means and standard deviations are displayed in Tables 1 and 2.

Table 2 Descriptive statistics and analyzed results of perceptive measures and learning outcomes for conversational tutoring with low levels of learner control (TG2) and the static tutoring condition (CG)

Dependent variable	Details	TG2 ($n=32$)	CG ($n=32$)	Analysis results	Bonferroni adjusted p- value
Learning Enjoyment	Mean (SD)	4.359 (1.135)	4.094 (1.542)	t(61.95) = 0.732, $p=0.47$, 95% CI[-0.459, 0.99], d = 0.183	1.0
Learner Autonomy	Mean (SD)	5.073 (1.191)	5.229 (1.092)	t(61.55) = -0.547, $p=0.709$, 95% CI[-0.633, inf], d = -0.136	1.0
Ease-of-Use	Mean (SD)	5.24 (1.1149)	4.833 (1.071)	t(61.69) = 1.462, $p=0.149$, 95% CI[-0.149, 0.961], d = 0.366	0.594
Assessment Quality	Mean (SD)	-0.125 (0.626)	-0.052 (0.704)	t(61.18) = -0.437, $p=0.668$, 95% CI[-0.351, inf], d = 0.287	1.0

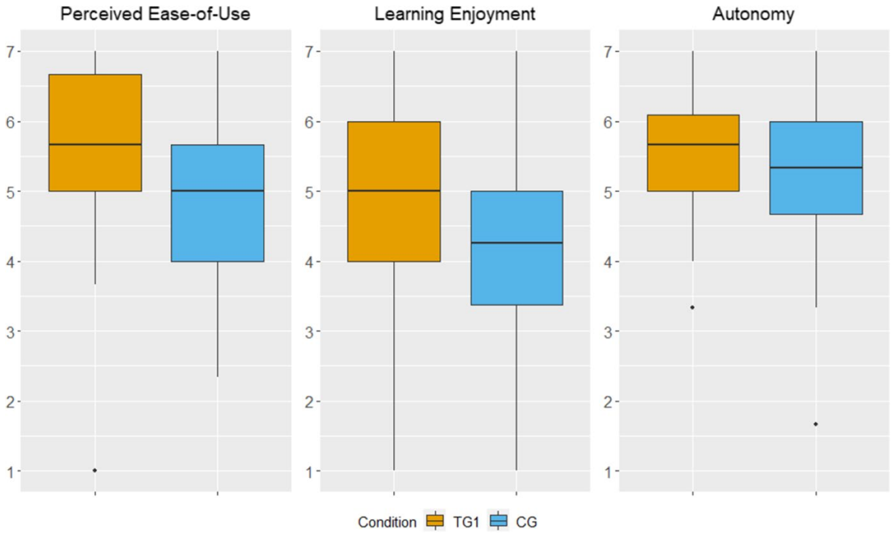


Fig. 5 Boxplots of perceptions of ease-of-use, learning enjoyment, and autonomy between TG1 and CG

First, we compared the ease-of-use of *WritingTutor* with the results of the static control group (see Fig. 5). The perceived ease-of-use of *WritingTutor* was rated with a mean value of 5.552 (SD = 1.272). The value is significantly higher than the results of the alternative static tutoring approach (mean = 4.833, SD = 1.071, $p=0.017^*$) according to a two-tailed Welch's t-test.

Second, we assessed the perceived learning enjoyment of *WritingTutor*. Participants judged enjoyment with a mean value of 4.875 (SD = 1.606). For the control group, we observed a mean value of 4.094 (SD = 1.542). The results of the t-test confirm that the difference is statistically significant ($p \leq 0.05$) according to a two-tailed Welch's t-test. The confidence interval for TG1, with a range of (95% CI[-0.006, 1.568]), indicates a potential higher mean for the measured learning enjoyment level compared to the CG.

Third, we analyzed students' learning outcomes and perceived learner autonomy between the pre- and post-tests. Students using *WritingTutor* rated their perceived learner autonomy with a mean value of 5.521 (SD = 0.946) and showed an improvement in learning outcomes of 0.052 between the pre- and post-tests (SD = 0.605). Students in the control group rated their perceived learner autonomy with a mean of 5.22 (SD = 1.09); however, they did not show an improvement in learning outcomes between the pre- and post- tests (mean = -0.052, SD = 0.704). The results show no significant difference between TG1 and CG and allow no clear indication based on the confidence interval (see Table 1). We also conducted an ANOVA analysis between the pre-and post-tests of each group's results and found no significant difference between pre-and post-tests within any of the groups ($p \geq 0.05$). Finally, we corrected the results of the questionnaires for multiple comparisons with a Bonferroni correction. The adjusted p-values are: for learning enjoyment, $p = 0.206$; for learning autonomy, $p = 1.0$, for perceived ease-of-use, $p = 0.069$, and for assessment quality, $p = 1.0$.

Also, we compare the values from the learning enjoyment, learning autonomy, and ease-of-use to the midpoints (neutral value of 4), and we see indications of positive learning experiences for *WritingTutor*. To provide a standardized metric for comparison against known benchmarks in human- computer interaction, we normalized the scores. Specifically, the normalization was calculated using the formula:

$$\text{normalized} - \text{score} = \frac{\text{raw} - \text{score}}{\text{maximum} - \text{score}}$$

where the maximum score for our scale is 7. Using this approach, students utilizing our CTS from TG1 had normalized scores of approximately 0.79 for perceived ease-of-use, 0.69 for learning enjoyment, and 0.78 for perceived learner autonomy. All these values are at or above the benchmark of 0.7, as established in prior HCI research, such as by Wambsganss et al. (2021a), Lehmann et al. (2016) and Wambsganss et al. (2020). It is important to note that these normalized scores do not represent percentages of enjoyment or ease-of-use, but instead are used for comparison against the established benchmarks (0.7). These results suggest that incorporating learner control into CTS may lead to a good student learning experiences. Nevertheless, we aim to mention that the normalized learning experience of the control group was similarly high for learner control (0.74), but lower for learning enjoyment (0.58) and ease-of-use (0.69).

For the comparison between TG2 and CG, the results are detailed in Table 2. The statistical analysis between conversational tutoring with lower levels of learner control (TG2) and static tutoring (CG) did not reveal any significant differences in learning enjoyment, learner autonomy, ease-of-use, or assessment quality.

Our findings suggest that conversational tutoring with higher levels of learner control may enhance certain aspects of the learning experience in essay-writing assignments, such as ease-of-use and learning enjoyment, when juxtaposed with a static tutoring group. However, conversational tutoring with lower levels of learner control did not demonstrate a significant impact on learner experiences compared to the static control group.

Table 3 Descriptive statistics and analyzed results of high and low learner control in conversational tutoring systems (TG1 and TG2)

Dependent variable	Details	TG1 ($n=32$)	TG2 ($n=32$)	Analysis results	Bonferroni Corrected p
Learning Enjoyment	Mean (SD)	4.875 (1.606)	4.359 (1.351)	t(60.24) = 1.39, $p=0.17$, 95% CI[-0.227, 1.258], d = 0.35	0.679
Learner Autonomy	Mean (SD)	5.521 (0.946)	5.073 (1.191)	t(58.84) = 1.66, $p=0.101$, 95% CI[-0.090, 0.986], d = 0.42	0.404
Ease-of-Use	Mean (SD)	5.552 (1.272)	5.24 (1.149)	t(61.37) = 1.031, $p=0.306$, 95% CI[-0.293, 0.918], d = 0.26	1.0
Assessment Quality	Mean (SD)	0.052 (0.605)	-0.125 (0.626)	t(61.93) = 1.15, $p=0.255$, 95% CI[-0.108, 0.509], d = 0.29	1.0

It is important to clarify that although we observed a directional change in learning gains between TG1 and CG, this difference was not statistically significant. Specifically, TG1 showed a slight positive learning gain (0.052), whereas the CG exhibited a slight negative change (-0.052).

In the pre-test, the mean scores were as follows: CG = 4.167 (SD = 0.542), TG1 = 4.042 (SD = 0.347), and TG2 = 4.167 (SD = 0.456). This suggests that the students began with comparable levels of expertise across the groups (ANOVA analysis $p=0.611$). For the post-test, the mean scores were: CG = 4.115 (SD = 0.512), TG1 = 4.094 (SD = 0.473), and TG2 = 4.042 (SD = 0.534), indicating a consistent level of competency in assessing essay quality across the groups at the study's conclusion (ANOVA analysis $p=0.0.528$). Table 4 provides the results from the ANOVA that compares the learning experience and learning outcome variables among the three groups.

H2: Results of Learner Control in Conversational Tutoring on Learning Experience and Learning Outcomes

To investigate our second hypothesis *(H2: in comparison to low levels of learner control, high levels of learner control in a CTS improve learning experience and learning outcomes in an essay-writing task)*, we compared the results of three variables (perceived ease-of-use, perceived learning enjoyment, and perceived autonomy) as a learning experience of the participants from TG1 (i.e., treatment with high learner control) and TG2 (i.e., treatment with low learner control). For learning outcomes, we evaluated learners' essay-quality assessment abilities by letting them assess the quality of three essays both before and after the experience (see Tables 3 and 4).

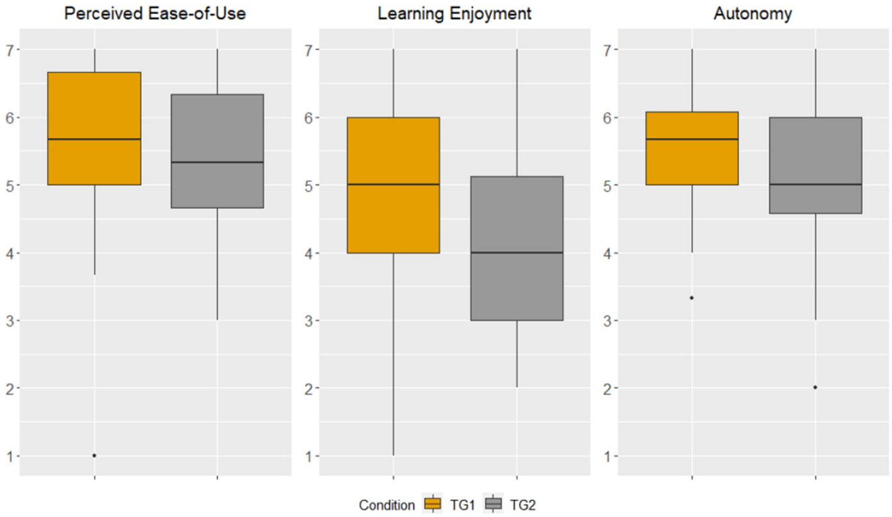


Fig. 6 Boxplots of perceptions of ease-of-use, learning enjoyment, and learning autonomy

Table 4 Comparison of learning experience and gains across conditions

Type	Construct	Mean (SD)			ANOVA Result	PW Sign. Comparisons
		CG	TG1	TG2		
Learning Experience	Learning Enjoyment	4.094 (1.542)	4.875 (1.606)	4.359 (1.351)	$p=0.08735$*	CG-TG1: 0.11, CG-TG2: 0.68, TG1- TG2: 0.16
	Learner Autonomy	5.229 (1.092)	5.521 (0.946)	5.073 (1.191)	$p=0.248$	-
	Ease-of-Use	4.833 (1.071)	5.552 (1.272)	5.24 (1.149)	$p=0.01992$*	CG-TG1: 0.017, CG-TG2: 0.175, TG1-TG2: 0.209
Learning Gains	Assessment Quality	-0.052 (0.704)	0.052 (0.605)	-0.125 (0.626)	$p=0.548$	-

*Results based on Kruskal–Wallis test and Wilcoxon rank sum test for the pair-wise comparison with continuity correction

Again, we compared the ease-of-use of *WritingTutor* including high levels of learner control (TG1) against the results of the benchmark version of *WritingTutor* with no additional learner control features and, in consequence, a lower level of learner control (see Fig. 6). The perceived ease-of-use of TG1 was rated with a mean value of 5.552 (SD = 1.272); for TG2 it was rated with a mean value of 5.24 (SD = 1.149). The results of the statistical analysis show no significant difference ($p=0.306$). For perceived learning enjoyment ($p=0.17$), we also see no statistically significant.

For learning autonomy, we observed a medium effect size of d = 0.42 between TG1 and TG2. Although the p-value suggests a trend ($p \leq 0.1$) according to a two-tailed Welch's t-test, it is important to note that after adjusting for multiple comparisons, this result was not statistically significant (adjusted $p = 0.404$). The confidence interval indicates a potentially higher level of learning autonomy for TG1 compared to TG2 (95% CI[-0.090, 0.986]), but this should be interpreted with caution given the lack of statistical significance. Participants assessed learning enjoyment in TG1 with a mean value of 4.875 (SD = 1.606) and learner autonomy with a mean of 5.52 (SD = 0.946). In TG2, participants rated learning enjoyment with a mean of 4.359 (SD = 1.351) and learner autonomy with a mean of 5.073 (SD = 1.191). Although the results are not statistically different, there is a consistent trend with participants of TG1 perceiving the interaction more favorably. In terms of learning outcomes, participants in TG1 showed a positive learning gain between the pre- and post-tests (mean = 0.052, SD = 0.605), whereas TG2 showed a small decrease (mean = -0.125, SD = 0.626). This difference is not statistically significant ($p = 0.255$), but the confidence interval suggests a potentially higher level in TG1 (95% CI[-0.108, 0.509]). After applying a Bonferroni correction for multiple comparisons, the adjusted p-values are: for learning enjoyment, $p = 0.679$; for learning autonomy, $p = 0.404$; for perceived ease-of-use, $p = 1.0$, and for assessment quality, $p = 1.0$. The results suggest that high learner control, in the form of content navigation and interface appearance control, may positively influence perceived learning enjoyment and learner autonomy. However, these findings are not statistically significant. Furthermore, participants from both *WritingTutor* groups rated the ease-of-use very high (normalized > 0.7, cf. (Venkatesh et al., 2003)). In conclusion, although there is a descriptive difference in the learning gains between TG1 and TG2, this difference is not statistically significant, though the confidence interval indicates a positive trend.

Qualitative Analysis of Effects on Learning Experience

To further investigate these results, and to understand the effects of learner control in CTS on students' learning experiences in more detail, we analyzed all students' qualitative responses. We identified three main topics among the treatments: 1) conversational tutoring and user experience, 2) adaptive writing feedback, and 3) learner control and autonomy.

In the following, we briefly present the highlights from the qualitative interviews and further elaborate on them in the discussion section. For the analysis, we translated the responses from German to English and categorized the most representative responses for TG1 in Table 5.

1) Conversational tutoring and user experience: Across all treat- ment groups, we see positive comments on the general instructions and the embedding of the learning intervention in the students' writing processes. Also, students from the control group (who received only static tutoring) reported that they liked the provided information and that the instructions, video tutorials, and static feedback categories helped them to improve their persuasive essay-writing

Table 5 Representative examples of qualitative user topics from TG1

Topic	Exemplary User Response from TG1
Conversational Tutoring and User Experience	*"I liked that I could decide for myself how much time I want to spend on each step, e.g. theory. In addition, the operation is very easy to understand."* *"It would be great to understand based on which word- s/sentence fragments the tool comes to its judgment. This would help to gain further insight. Also, more features could be added than just the current two."*
Adaptive Writing Feedback	*"I found the feedback very useful and gave a quick overview."* *"I liked the character and word count tracking, as well as enumeration of reused words to reduce repetitive writing."*
Learner Control and Autonomy	*"I had full control and could select the individual steps (and further information, e.g. theory) at my own convenience. I liked this."* *"I had greater control because I got feedback directly." "The different options rather quickly distracted, because of many possibilities to click on something."* *"I found modifying the UI more of a gimmick and not that important. What I found best was the display of the number of words or most used words,* etc.*"* *"That you could give the chatbot a name.... Was very nice and made the usage process very personal."* *"The color settings were a nice touch, but not really helpful or enriching."* *"I would improve the step by step approach of the writing tool. With the many different choices you don't really know where to start."*

skills. Exemplary quotations include: *"WritingTutor has given me confidence in how best to structure and craft a persuasive text. It is a good tool to hold on to when writing an argumentative text"* and *"It has helped me recall the theory of argumentation during the writing process."* Nevertheless, most students in the control group (around 80.0%) also mentioned a lack of interactivity within the learning exercise. For example, as this student reported:*"For me, the learning exercise wasn't interactive at all."* In contrast, most of the subjects in the treatment groups (around 90.0% of students in TG1 and TG2) did report positive impressions of the tool regarding the general user interaction and the conversational tutoring. For example, students from TG2 stated: *"WritingTutor gave me some new ideas regarding my approach to writing texts"*; *"I found the linking of relevant videos very good, as this helped significantly in writing"*; and *"WritingTutor can give us a bit more confidence in writing."*

2) Adaptive writing feedback: A strong difference in qualitative comments emerged when it came to the second topic, adaptive writing feedback. Several students in the control group mentioned the absence of feedback as an opportunity to improve the tool's interactivity.

In both treatment groups, most students expressed positive attitudes about the text feedback, such as on 1) the general function (*"Very simple tool with high information content and exciting analysis of my text"*), 2) the word count (*"The field with the much-used words below helped me a lot to see what is highlighted in my text. Moreover, this awareness that the text needs to become argumentatively strong gave me an extra motivation"*), or 3) the polarity (*"The evaluation at the end about the objectivity/subjectivity of the text was very nice. Such aspects are often overlooked during the writing process"*). Nevertheless, some negative aspects of the learning interaction and the feedback were mentioned, such as *"There were no specific suggestions"* (TG2), *"Grammar feedback was missing"* and *"More explanations in the feedback dashboard*" (TG1), or *"The theory should be taught a little more interactively" (TG1).*

3) Learner control and autonomy: Qualitative results show rather divergent comments on the autonomy and control of the learning experience, in particular for the interaction in TG1. Interestingly, some participants mentioned the adaptivity of writing feedback as a great way to control their learning experience, such as: *"I had greater control because I got feedback directly."* Finally, in TG2 (*WritingTutor* without any levels of learner control), students mentioned that they would have liked more personalization options for the tool, such as: *"The tutor could be made more individual, e.g. with a name."*

Discussion

Theoretical and Practical Contributions

The objectives of our study were to investigate the effects of learner control embedded in a CTS on students' essay-writing skills. In particular, we focused on how the interactions with a CTS impact the learning experience and learning outcomes. To do so, we designed *WritingTutor*, a CTS that guides learners through an essay-writing exercise by providing theoretical instructions, input, and text-based feedback on the essays' objectivity, polarity, and text quality. Our findings suggest that students who had a high level of learner control in the CTS, such as navigating the learning content through predefined buttons and customizing the interface appearance, tended to perceive higher learning enjoyment and ease-of-use compared to those using a static tutoring tool when comparing the means with a two-tailed Welch's t-test. Notably, students using our CTS with these control features perceived a trend toward higher learning autonomy than those without these features. However, it is essential to emphasize that none of the results reached levels of statistical significance after correction with multiple comparisons with a Bonferroni correction ($p \leq 0.05$). Our results from comparing the means with a two-tailed Welch's t-test partially support hypothesis H1, suggesting that embedding learner control in a CTS may enhance the digital learning experience in essay-writing exercises.

Both CTS groups (TG1, TG2) perceived high ease-of-use, reflecting the general positive impact of CTS on students' learning experiences. For instance, students in TG1 had normalized scores of approximately 0.79 for perceived ease-of-use and 0.78 for perceived learner autonomy, surpassing the benchmark of 0.7.

With the help of personalized adjustments and content navigation, students appeared to perceive the same learning interaction as more enjoyable than without learner control. Students across both CTS treatment groups reported that they were motivated by an easy-to-use learning tool that combined theory explanation, video instructions, and essay-writing criteria. These are important factors for motivation and the long-term usage of CTS and digital learning tools in general.

Our results from a comparison of means with two-tailed Welch's t-tests hint at the possibility that high-level learner control may positively influence learning autonomy when compared to a CTS without additional control features (H2). However, it is crucial to note that these findings are not statistically significant after adjusting for multiple comparisons. Thus, although there is a trend, it should be interpreted with caution. The participants' perceived high learning autonomy aligns with the findings of prior literature that has highlighted the benefits of CTS with high learner control.

Interestingly, feedback on learner control and autonomy, particularly between TG1 and TG2, showed that although students in TG1 appreciated the diverse control opportunities, many in TG2 also felt a strong sense of control through the conversational writing exercises. This suggests that a CTS, even without added features, may already offer students satisfactory levels of learning autonomy for their exercises. The quantitative results on learning outcomes appear to support this observation. Participants valued the additional features, implying higher learning control, but this did not translate to superior learning outcomes.

Although the quantitative data of our sample do not allow for the derivation of more precise implications, the qualitative data suggest explanations for the effects of learner control in CTS on students' learning experience. The participants reported feeling more informed in CTS treatments (TG1, TG2) (e.g., *"I found the feedback very useful and gave a quick overview"*). Also, they connected the positive effect and the personalization of high-level control CTS (TG1) (e.g., *"That you could give the chatbot a name…qas very nice and made the usage process very personal"*).

With these insights, our study expands on prior research around two main literature streams. First, we provide empirical data and design findings for the emerging field of CTS (Weber et al., 2021; Han et al., 2023; MacLellan & Koedinger, 2020). By investigating H1, we find evidence of the positive impact of conversational tutoring with high levels of learner control on learner experience in essay-writing exercises when comparing the means. However, we cannot definitively prove them with significance levels of $p \leq 0.05$. Second, we provide insights for the literature on learner control and the embedding of high-level learner control features in novel CTSs. Especially for learner control, our study is one of the first to shed light on design implications and empirical investigations of different learner control features in a controlled experiment. Our results lay the foundation for future research to investigate the potential interfering effects with student knowledge status and the innovative learning form of CTS in more detail.

Our study also offers several practical implications for researchers and educators to design general educational scenarios or intelligent tutoring systems. First, we provide an archetypal example of a CTS for general essay-writing skills. Our user study with *WritingTutor* provides other researchers with insights for how best to embed a CTS into a general writing exercise. Because *WritingTutor* is built entirely on open-source frameworks, adapting it to other scenarios or languages would be relatively easy compared to previously published tutors, which often require that new datasets be trained (e.g., Wambsganss et al. (2021a)).

Limitations and Future Work

Our study's quantitative findings revealed no statistically significant differences in learning outcomes. Notably, although there was an observable difference in the essay-writing assessment ability among students using the *WritingTutor* with enhanced learner control, this difference did not reach statistical significance when juxtaposed against both benchmark groups. The lack of statistical significance may be attributable to various external factors in our experimental setup. These factors encompass the sample size, potential confounding variables, boundary conditions, experimental design, and the measurement tools employed. For example, the instrument we employed to measure learning might not have been optimally designed to capture the specific knowledge improvements targeted by the tutoring systems under examination. Although past literature highlights the importance of students' ability to com- prehend and critically assess essay quality as an integral learning outcome in persuasive writing tasks (especially with novice learners (David Smith & Brooker, 1999)), a more tailored measurement construct might have yielded different insights with regard to our pedagogical scenario. Second, the duration and scope of the intervention may play a major role in the measured constructs. The student's limited exposure to a single essay-writing task may not have been sufficient to induce notice- able knowledge gains and the power of learner control and interactivity of a CTS over a longer period of time. Extending the intervention period would allow students to engage with a broader range of essay-writing tasks over a longer period of time. This extended interaction might enhance the chances of observing significant knowledge gains and provide a more comprehensive understanding of the learning tool's impact. Third, the ICAP framework, which is foundational to our CTS design, may encounter challenges when trans- posed to the realms of learner control and conversational tutoring. The ICAP framework, as described by Chi and Wylie (2014), categorizes learning activities based on their level of engagement: passive, active, constructive, and interactive. Although robust for traditional learning environments, its direct application to a conversational tutoring system with varying levels of learner control may not be straightforward. For example, the dynamics of human-to-AI interaction in our CTS may differ from the peer-to-peer interaction described in ICAP as "interactive" (e.g., Xu et al.(2021)). Additionally, the ICAP framework does not explicitly address learner control. Hence, our attempt to map levels of learner control to ICAP's engagement levels may have introduced nuances not captured by the framework. Fourth, beyond the potential limitations of the ICAP framework itself, our application of its principles in our design may not have been optimal. The unique challenges of integrating the ICAP modes into a CTS environment, especially when introducing learner control, may have led to unforeseen complexities in student engagement levels. Despite these challenges and uncertainties, our study's findings align with existing literature on the topic: Prior research has yielded mixed results concerning the efficacy of learner control features (Brown et al., 2016; Fisher et al., 2017; Sorgenfrei & Smolnik, 2016). It is noteworthy that not every study has identified a direct, positive impact of learner control on learning outcomes. In this con- text, our findings contribute to this ongoing discourse, emphasizing the need for further exploration and more nuanced understanding.

Moreover, our study entails some limitations concerning the implementation and sample group. Although deploying learner control features in a CTS had positive effects on the learner experience of the CTS, it did not yield significant benefits for learning outcomes. From previous research, we know that the level of learning progress has interaction effects on the learning outcome. For example, novice learners need more structure and should not get too much control over the core pedagogical features of the CTS. Our participants also belong to the same participant sample, and we cannot naturally control for differences in the level of learning progress. Consequently, we do not distinguish between novice participants, who are just beginning their learning path, and more advanced learners. Future research should dig deeper and investigate the effect of different learner control features (i.e., surface-level and deep-level) in detail. Furthermore, in future studies, we intend to include multiple participant samples with different levels of learning progress. For example, we plan to include participants of different age levels in future evaluations.

In our study, we conducted a laboratory experiment with an online participant sample. In the future, we plan to conduct field studies in secondary schools that integrate *WritingTutor* into participating students' English language learning curriculum.

Conclusion

In our study, we investigated whether high-level learner control embedded in a CTS helps improve students' learning outcomes. To do so, we designed a novel conversational agent called *WritingTutor* as a CTS that tutors students through an essay-writing exercise. We compared *WritingTutor* with high levels of learner control against a benchmark version with low levels of learner control, as well as a static tutoring group, in a study sampling 96 students. We found indications that students who were able to control the interface appearance and the learning sequence with *WritingTutor* reported higher levels of perceived ease-of-use and learning enjoyment in the essay-writing task. How- ever, participants with high levels of learner control did not show significantly higher learning outcomes in their post-tests. Our study, one of the first to explore learner control in CTS, suggests that simple learner control features may potentially enhance the learning experience in traditional learning tasks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-024-00409-x.

Funding Open access funding provided by Bern University of Applied Sciences. The work of the first author was supported by a grant from the Swiss State Secretariat for Education, Research and Innovation.

Declarations

Disclosure The authors have no relevant financial or non-financial interests to disclose.

References

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: cognitive absorption and beliefs about information technology usage. *MIS Quarterly, 24*(4), 665. https://doi.org/10.2307/3250951

Ashford, S. J. (1986). Feedback-seeking in individual adaptation: a resource perspective. *Academy of Management Journal, 29*(3), 465–487. https://doi.org/10.2307/256219

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation - Advances in Research and Theory, 2*(C), 89–195. https://doi.org/10.1016/s0079-7421(08)60422-3

Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Behrend, T. S., & Thompson, L. F. (2012). Using animated agents in learner-controlled training: The effects of design control. *International Journal of Training and Development, 16*(4), 263–283.

Benke, I., Gnewuch, U., & Maedche, A. (2022). Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior, 129*, 107122. https://doi.org/10.1016/j.chb.2021.107122

Brinton, C. G., Rill, R., Ha, S., Chiang, M., Smith, R., & Ju, W. (2015). Individualization for education at scale: MIIC design and preliminary evaluation. *IEEE Transactions on Learning Technologies, 8*(1), 136–148. https://doi.org/10.1109/tlt.2014.2370635

Brown, K. G., Howardson, G., & Fisher, S. L. (2016). Learner control and e-learning: Taking stock and moving forward. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 267–291. https://doi.org/10.1146/annurev-orgpsych-041015-062344

Carolan, T. F., Hutchins, S. D., Wickens, C. D., & Cumming, J. M. (2014). Costs and benefits of more learner freedom: Meta-analyses of exploratory and learner control training methods. *Human Factors, 56*(5), 999–1004. https://doi.org/10.1177/0018720813517710

ChatterBot. (n.d.). Documentation. Retrieved February 29, 2024, from https://chatterbot.readthedocs.io/en/stable/.

Chen, L., Zeng, S., & Wang, W. (2021). The influence of emotion and learner control on multimedia learning. *Learning and Motivation, 76*, 101762. https://doi.org/10.1016/j.lmot.2021.101762

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cohn, M. (2004). *User Stories Applied: For Agile Software Development*. USA: Addison Wesley Longman Publishing Co. Inc.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104–126. https://doi.org/10.1007/bf03177550

David Smith, J. C., & Brooker, R. (1999). The impact of students' approaches to essay writing on the quality of their essays. *Assessment & Evaluation in Higher Education, 24*(3), 327–338. https://doi.org/10.1080/0260293990240306

Eom, S. B., Wen, H. J., & Ashill, N. (2006). The determinants of students' perceived learning outcomes and satisfaction in university online education: An empirical investigation. *Decision Sciences Journal of Innovative Education, 4*(2), 215–235. https://doi.org/10.1111/j.1540-4609.2006.00114.x

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies, 132*, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Fisher, S., Howardson, G., Wasserman, M. E., & Orvis, K. (2017). How do learners interact with e-learning? Examining patterns of learner control behaviors. *AIS Transactions on Human-Computer Interaction, 9*(2), 75–98.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions, 24*(4), 38–42. https://doi.org/10.1016/j.chb.2018.02.025

Gläser, J., & Laudel, G. (2010). Experteninterviews und Qualitative Inhaltsanalyse: Als Instrumente Rekonstruierender Untersuchungen.

Grossman, J., Lin, Z., Sheng, H., Wei, J. T. Z., Williams, J. J., & Goel, S. (2019). MathBot: Transforming online resources for learning math into conversational interactions. *AAAI 2019 Story-Enabled Intelligence.*

Hair, J. F., Jr., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook* (p. 197). Springer Nature.

Han, S., Liu, M., Pan, Z., Cai, Y., & Shao, P. (2023). Making FAQ chatbots more inclusive: An examination of non-native English users' interactions with new technology in massive open online courses. https://doi.org/10.1007/s40593-022-00311-4.

Holden Thorp, H. (2023). ChatGPT is fun, but not an author. *Science, 379*, 313–313. https://doi.org/10.1126/science.adg7879

Hone, K. S., & El Said, G. R. (2016). Exploring the factors affecting MOOC retention: A survey study. *Computers and Education, 98*, 157–168. https://doi.org/10.1016/j.compedu.2016.03.016

Jumonville, A. (2012). Encyclopedia of the Sciences of Learning. *Reference Reviews, 26*(8), 23–24.

Jung, E., Kim, D., Yoon, M., Park, S., & Oakley, B. (2019). The influence of instructional design on learner control, sense of achievement, and perceived effectiveness in a supersize mooc course. *Computers & Education, 128*, 377–388. https://doi.org/10.1016/j.compedu.2018.10.001

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems, 20*(2), 177–185. https://doi.org/10.1016/j.knosys.2006.11.014

Kim, C. M., & Baylor, A. L. (2008). A virtual change agent: Motivating pre-service teachers to integrate technology in their future classrooms. *Educational Technology and Society, 11*(2), 309–321.

Kim, S., Lee, J., & Gweon, G. (2019). Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Conference on Human Factors in Computing Systems - Proceedings*, 1–12. https://doi.org/10.1145/3290605.3300316

Kowatsch, T., Nißen, M., Shih, C.-h.I., Rüegger, D., Filler, A., Künzler, F., Barata, F., Haug, S., Brogle, B., Heldt, K., Gindrat, P., Farpour-lambert, N., Allemand, D. (2017). Text-based healthcare chatbots supporting patient and health professional teams: Preliminary results of a randomized controlled rrial on childhood obesity. *Persuasive Embodied Agents for Behavior Change (PEACH2017) Workshop 1(Iva 2017)*, 1–10.

Kraiger, K., & Jerden, E. (2007). A meta-analytic investigation of learner control: Old findings and new directions. In S. M. Fiore & E. Salas (Eds.), *Toward a science of distributed learning* (pp. 65–90). American Psychological Association. https://doi.org/10.1037/11582-004

Kuhail, M. A., Alturki, N., Alramlawi, S., et al. (2023). Interacting with educational chatbots: A systematic review. *Education Information and Technologies, 28*, 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Laumer, S., Maier, C., & Gubler, F.T. (2019). Chatbot acceptance in healthcare: explaining user adoption of conversational agents for disease diagnosis. *Twenty-Seventh European Conference on Information Systems (ECIS2019)*, Stockholm-Uppsala, Sweden, 0–18.

Lee, M. K. O., Cheung, C. M. K., & Chen, Z. (2005). Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information and Management, 42*(8), 1095–1104. https://doi.org/10.1016/j.im.2003.10.007

Lehmann, K., Söllner, M., & Leimeister, J.M. (2016). Design and evaluation of an IT-based peer assessment to increase learner performance in large-scale lectures. *ICIS 2016 Proceedings*.

MacLellan, C.J., & Koedinger, K.R. (2020). Domain-general tutor authoring with apprentice learner models. *32*(1), 76–117. https://doi.org/10.1007/s40593-020-00214-2. Accessed 2023-02-03.

Orvis, K. A., Fisher, S. L., & Wasserman, M. E. (2009). Power to the people: Using learner control to improve trainee reactions and learning in web-based instructional environments. *Journal of Applied Psychology, 94*(4), 960.

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

Pekrun, R., & Stephens, E.J. (2012). Academic emotions. In *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors*, 3–31. https://doi.org/10.1037/13274-001

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Roth, H. (1970). P¨adagogische Psychologie des Lehrens und Lernens. https://issuu.com/audio2brain/docs/name6bce04.

Ruan, S., Jiang, L., Xu, J., Tham, B. J.-K., Qiu, Z., Zhu, Y., Murnane, E. L., Brunskill, E., & Landay, J. A. (2019). QuizBot: A dialogue-based adaptive learning system for factual knowledge. *Chi*, 1–13. https://doi.org/10.1145/3290605.3300587.

Schöbel, S., Saqr, M., & Janson, A. (2021). Two decades of game concepts in digital learning environments–a bibliometric study and research agenda. *Computers & Education, 173*, 104296.

Seaman, J. E., Allen, I. E., & Seaman, J. (2018). Higher education reports - Babson survey research group. http://www.onlinelearningsurvey.com/highered.html.

Sharples, M. (2023). Automated essay writing: An AIED opinion. 32(4), 1119–1126. https://doi.org/10.1007/s40593-022-00300-7.

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics, 10*(4), 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

Soloway, E., Guzdial, M., & Hay, K. E. (1994). Learner-centered design: The challenge for HCI in the 21st century. *Interactions, 1*(2), 36–48.

Sorgenfrei, C., & Smolnik, S. (2016). The effectiveness of e-learning systems: A review of the empirical literature on learner control. *Decision Sciences Journal of Innovative Education, 14*(2), 154–184.

Sumikawa, Y., Fujiyoshi, M., Hatakeyama, H., & Nagai, M. (2019). Supporting creation of FAQ dataset for e-learning Chatbot. *Smart Innovation, Systems and Technologies, 142*(February), 3–13.

Suppes, P., & Morningstar, M. (1969). Computer-assisted instruction. *Science, 166*(3903), 343–350. https://doi.org/10.1126/science.166.3903.343

TextBlob. (n.d.). Documentation. Retrieved February 29, 2024, from https://textblob.readthedocs.io/en/dev/.

Van Griethuijsen, R. A., van Eijck, M. W., Haste, H., Den Brok, P. J., Skinner, N. C., Mansour, N., SavranGencer, A., & BouJaoude, S. (2015). Global patterns in students' views of science and interest in science. *Research in Science Education, 45*, 581–603.

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Leimeister, J.M., & Handschuh, S. (2020). AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. Association for Computing Machinery, New York, NY, USA.

Wambsganss, T., Kueng, T., Söllner, M., & Leimeister, J.M. (2021a). ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445781.

Wambsganss, T., Guggisberg, S., & Söllner, M. (2021b). ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback. In *16th International Conference on Wirtschaftsinformatik*, Essen, Germany.

Wambsganss, T., Zierau, N., Söllner, M., Käser, T., Koedinger, K.R., & Leimeister, J.M. (2022). Designing conversational evaluation tools: A comparison of text and voice modalities to improve response quality in course evaluations. *Proc. ACM Hum.-Comput. Interact., 6*(CSCW2). https://doi.org/10.1145/3555619

Weber, F., Wambsganss, T., Rüttimann, D., & Söllner, M. (2021). Pedagogical agents for interactive learning: A taxonomy of conversational agents in education. In Forty-Second International Conference on Information Systems. Austin, Texas, pp. 1–17.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3313831.3376781

Winkler, R., Söllner, M., & Leimeister, J. M. (2021). Enhancing problem-solving skills with smart personal assistant technology. *Computers & Education*, *165*. https://doi.org/10.1016/j.compedu.2021.104148.

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings*. https://doi.org/10.5465/ambpp.2018.15903abstract

Xiao, Z., Zhou, M.X., Liao, Q.V., Mark, G., Chi, C., Chen, W., & Yang, H. (2019). Tell me about yourself: using an ai-powered chatbot to conduct conversational surveys with open-ended questions, 27*(3). https://doi.org/10.1145/3381804.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on social media. *Conference on Human Factors in Computing Systems - Proceedings, 2017*, 3506–3510. https://doi.org/10.1145/3025453.3025496

Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2021). *From human-computer interaction to human-AI Interaction: New challenges and opportunities for enabling human-centered AI* (pp. 1–73)

Zierau, N., Wambsganss, T., Janson, A., Schöbel, S., & Leimeister, J.M. (2020). The Anatomy of User Experience with Conversational Agents: A Taxonomy and Propositions of Service Clues. In *International Conference on Information Systems (ICIS)*, pp. 1–17.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Thiemo Wambsganss[1] · Ivo Benke[2] · Alexander Maedche[2] · Kenneth Koedinger[3] · Tanja Käser[4]

✉ Thiemo Wambsganss
thiemo.wambsganss@bfh.ch

Ivo Benke
ivo.benke@kit.edu

Alexander Maedche
alexander.maedche@kit.edu

Kenneth Koedinger
kk1u@andrew.cmu.edu

Tanja Käser
tanja.kaeser@epfl.ch

[1] Institute of Digital Technology Management, Bern University of Applied Sciences, Bern, Switzerland

[2] Human-Centered Systems Lab (H-Lab), Karlsruhe Institute of Technology, Karlsruhe, Germany

[3] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA

[4] Machine Learning for Education Laboratory, EPFL, Lausanne, Switzerland