Please quote as: Winkler, Rainer; Hobert, Sebastian; Salovaara, Antti; Söllner, Matthias & Leimeister, Jan Marco: Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agents. 2020. - Computer Human Interaction Conference (CHI). - Honolulu, Hawaii.

Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent

Rainer Winkler
Institute of Information Management
University of St. Gallen
rainer.winkler@unisg.ch

Sebastian Hobert
University of Goettingen
shobert@uni-goettingen.de

Antti Salovaara
Department of Design
Aalto University
antti.salovaara@aalto.fi

Matthias Söllner
Information Systems and Systems Engineering
University of Kassel
soellner@uni-kassel.de

Jan Marco Leimeister
Institute of Information Management
University of St. Gallen
janmarco.leimeister@unisg.ch

ABSTRACT
Enrollment in online courses has sharply increased in higher education. Although online education can be scaled to large audiences, the lack of interaction between educators and learners is difficult to replace and remains a primary challenge in the field. Conversational agents may alleviate this problem by engaging in natural interaction and by scaffolding learners' understanding similarly to educators. However, whether this approach can also be used to enrich online video lectures has largely remained unknown. We developed Sara, a conversational agent that appears during an online video lecture. She provides scaffolds by voice and text when needed and includes a voice-based input mode. An evaluation with 182 learners in a 2 x 2 lab experiment demonstrated that Sara, compared to more traditional conversational agents, significantly improved learning in a programming task. This study highlights the importance of including scaffolding and voice-based conversational agents in online videos to improve meaningful learning.

Author Keywords
Conversational agent; scaffolding; voice interaction; interactivity; online videos; online education; experiment

CSS Concepts
• **Human-centered computing – Natural language interfaces;**
• **Applied computing – Education.**

INTRODUCTION
Enrollment in online courses has grown rapidly [61]. In 2017, 33.1% of learners worldwide took at least one course online, compared to 24.8% in 2012 [35]. With this development, the amount of online educational content is rapidly increasing, particularly in the form of online video lectures [31]. However, the design of these learning materials is faced with a challenge: educators are ill-equipped at eliciting individualized, meaningful interactions with their learners [62]. This is problematic, since we know from learning theory that we learn best when we interact socially with others through meaningful interactions [29]. Meaningful interactions, such as scaffolding dialogs between an educator and a learner, can have a significant effect on learning outcomes, including learners' information retention [25] and their ability to apply the new knowledge to solve novel problems – commonly referred to as learners' transfer ability [8].

In the quest to find ways to imitate meaningful, individual educator–learner interactions, conversational agents (CAs) have started to receive more and more attention. A CA is a computer system intended to converse with a human [47]. In the educational domain, CAs employ text, speech, graphics, haptics, gestures, and other modes in various combinations for communication trying to help learners conduct tasks – thereby imitating the gold standard of educators [45]. The research around CAs in education was inspired by Graesser et al.'s [19] investigation of human tutoring behaviors. It was followed by many successful implementations of CAs, such as AutoTutor [54], and experiments on the effects of using CAs, for instance, using Why2-AutoTutor [67] and others. Past research also investigated CAs in online learning environments. For example, Song et al. [62] created a CA that allows learners to reflect on their weekly learning experiences during an online course. Grossman et al. [20] developed MathBot, an automated text-based tutor that explains math concepts and offers tailored feedback. These systems primarily followed an *instructional scaffolding* logic, originally introduced by Wood et al. [73] with which they described the behavior of educators, meaning that they have to analyze the individual learner in detail in order to be able to offer scaffolds that effectively help learners gain knowledge. In the wake of increasing interest in online education on the one hand and the CAs on the other, integration of scaffolded instructions in online teaching has received growing attention [9, 30]. One way of offering instructional scaffolds is to

CHI '20, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6708-0/20/04…$15.00
https://doi.org/10.1145/3313831.3376781

use CAs powered with natural language processing (NLP). This approach has recently been advancing rapidly, leading to increasingly sophisticated learning systems [14, 17]. For example, such systems enable learners to express their answers via voice in a free manner during a two-way dialog and receive adaptive answers from NLP techniques interpreting the spoken words. This way of interacting has the potential to get closer to educators' scaffolding techniques. This might lead to a more meaningful interaction and result in increased information retention and transfer ability – which are key indicators for learning success [24, 64].

The positive implication of integrating both CAs and scaffoldings have been shown mainly in classical online learning environments like online courses, learning apps or intelligent tutoring systems. However, we argue that these classical e-learning settings differ significantly from online video lectures that we address in this research study. In online video lectures, learners already receive multiple inputs consisting of (1) the educator's recorded voice and (2) the image of the recorded – usually text-based – slides. Thus, the question arises whether adding a third and fourth channel (i.e., CAs and scaffoldings) to online video lectures also achieves a positive effect on the learning success. If a positive effect can be shown, it needs to be further investigated how such a scaffolding-based CA should be integrated in this new context of online video lectures. To this end, we aim to (1) transfer and replicate the positive effects of the instructional scaffolding mechanism to the new context of online video lectures, (2) compare different implementations of CA interaction types (voice- and text-based vs. text-based), and (3) analyze the interaction effects of scaffolding and CA interaction types.

To research these aspects, we present a study on Sara, a web-based CA that can be layered on top of already existing online video lectures. Sara provides voice- and text-based scaffolds to learners and thereby helps them to remember concepts better (information retention) and to use the gained knowledge to solve novel problems (transfer ability). At predefined times Sara interrupts an online video lecture and actively intervenes in the learning process. Sara asks questions and offers detailed explanations and scaffolding questions if the learner's initial answer was wrong. This scaffolding strategy is inspired by previous studies in the field of scaffolding and intelligent tutoring systems [1, 11]. To determine the true impact of Sara's scaffolding approach on learners' information retention and transfer ability, we evaluated it in a 2 x 2 lab experiment with 182 learners. We compared Sara with three other types of CAs that are often used today: (a) a voice-based, non-scaffolding CA, (b) a text-based, scaffolding CA, and (c) a text-based, non-scaffolding CA. The findings suggest that Sara is not only able to increase learners' information retention but also helps them to apply the acquired knowledge to novel problems.

We build on previous findings that investigated the beneficial effects of CAs in education [59] and online education in particular [62]. Following this line of research, our study highlights the importance of including scaffolding and voice-based CAs in the context of online video lectures. In contrast to the work of Litman et al. [41], our paper focuses on voice-based scaffolding during video instruction when students learn about new concepts for the first time. This is different from scaffolding during the exercise phase when students solve a task. With the help of scaffolding and voice-based CAs, static online video lectures may get closer to the gold standard of educator-learner interactions. This effect may be even stronger when the CA is voice-based. To the best of our knowledge, this is the first study that designs a voice- and text-based CA with an instructional scaffolding mechanism in an online video lecture context, where the standard way of instruction is solely online videos that prevent students from actively thinking and deepening the concept they have just watched. Moreover, we empirically evaluated and rigorously compared our scaffolding and voice-based CA to other types of CAs. Furthermore, our study offers design implications based on scaffolding and multimedia learning theory for building improved future CAs to enrich online video lectures. This informs software designers and educational providers who wish to use CAs in their online education courses.

RELATED WORK AND HYPOTHESES DEVELOPMENT

The design, implementation, and strategies of CAs employed in education vary widely, which reflects the diverse nature of the evolvement of this technology field [28]. Interactions between learners and CAs are usually textually mediated (e.g., [62]), where CAs show questions or hints and learners click on buttons or type responses on the keyboard. Some systems use embodied CAs (e.g., [60]) capable of displaying emotions and gestures, whereas others use simpler avatars (e.g., [27]). Voice output, using text-to-speech synthesis, is used in some systems (e.g., [18]), and speech input systems are increasingly viable (e.g., [42]). With recent technological developments in NLP, we see more and more CAs that engage in largely free interactions with the learners resulting in knowledge gains (e.g., [40]). For example, AutoTutor and its derivations have been very effective as a learning technology [54]. AutoTutor produced learning gains that are on average about 0.8 standard deviations above controls who read static instructional materials for an equivalent amount of time [15]. Ruan et al.'s [59] mobile QuizBot helped learners gain factual knowledge and it significantly increased their knowledge compared to a more traditional flashcard application. More commercially available CAs also use voice output and input systems (e.g., Amazon's Alexa, Google's Assistant). For example, Winkler et al. [72] used Alexa to support groups of learners in solving a complex problem and showed that learners achieved solutions of better quality.

Although CAs have shown to increase learning outcomes, the potential of scaffolding and voice-based CAs layered on top of static video lectures has not yet been investigated. *This calls for the question whether scaffolding and voice-based CAs integrated in online video lectures are able to increase meaningful interactions between CAs and learners during instruction resulting in increased learning outcomes.* This

question motivated us to investigate scaffolding-based CAs in the new context of online video lectures.

Social Constructivism and the Theory of Scaffolding

Wood et al.'s [73] theory of scaffolding emerged around 1976 as a part of social constructivist theory and was particularly influenced by the work of the Russian psychologist Lev Vygotsky [69]. Vygotsky [69] argued that we learn best in a social environment, where we construct meaning through interaction with others. His *Zone of Proximal Development (ZPD) Theory*, which states that we can learn more in the presence of a knowledgeable other person, became the basis for the *Theory of Scaffolding* [22]. Among other things, this theory states that when learners strive to acquire new knowledge, they need individualized support falling within their individual ZPD. ZPD represents the potential distance the learner could reach with the help of a more knowledgeable other [69]. As they advance and become more independent in their thinking, this support can gradually fade away. Instructional scaffolding is a term used to explain the interaction between learners and their educators and is a process that enables a novice to achieve a goal or an objective that would otherwise be unattainable without assistance [73]. The main goal of the educator is to offer scaffolds, such as questions and hints, within learners' individual ZPD. Instructional scaffolding is not one-way but an interactive and reciprocal process between the learner and the educator [5]. For our study, this means that scaffolds have to be adaptive and personalized in contrast to one-size-fits-all video lectures. We refer to the concept of instructional scaffolding when we talk about scaffolds. Instructional scaffolding during instruction improves learners' ability to remember concepts [73]. However, in online learning, an educator is hardly able to offer instructional scaffolds to every learner [57]. Scaffolding-based CAs might be able to imitate the educator's scaffolding dialogs to some extent resulting in a meaningful learning process. Thus, we propose:

H1a: Learners interacting with scaffolding-based CAs show higher levels of information retention compared to those interacting with non-scaffolding-based CAs.

Past research indicates that instructional scaffolding not only helps learners to remember learned concepts but also to increase their transfer ability [2]. This has been shown both in problem-solving tasks in a computer course [71] and in solving of novel problems in physics [50]. Wang et al. [71] created iTutor to help learners to learn basic computer skills. The results indicate that learners in the iTutor group experience better learning effectiveness than those in the control group. Moreover, Murphy and Messer [50] were able to prove that learners can better transfer their knowledge to novel problems in a physics task when receiving scaffolds compared to working in a group discussion condition and a condition where learners worked alone. Thus, we propose:

H1b: Learners interacting with scaffolding-based CAs show higher levels of transfer ability compared to those interacting with non-scaffolding-based CAs.

CA's potential to imitate educators has also often hinged on their previously limited ability to translate speech to text [4]. Recent advances in NLP show a way out of this conundrum, as CAs are becoming more intelligent and users can speak almost freely with the CAs [23]. Voice-based interactions between CAs and learners have the potential to get even closer to the gold standard of educators [72]. However, there seems to be hardly any empirical study on CAs that would both offer scaffolding-based teaching and would use modern voice-based CAs in an online video lecture. In the following, we briefly discuss why voice-based technology in CAs might have advantages for the learning process.

Cognitive Theory of Multimedia Learning

The cognitive theory of multimedia learning [44] is based on three assumptions: there are two separate channels (auditory and visual) for processing information; there is limited channel capacity; and learning is an active process of filtering, selecting, organizing, and integrating information. Visual and verbal information types are processed differently and along distinct channels in the human mind, creating separate cognitive representations [43]. The ability to learn through multiple channels concurrently increases the chance of remembering knowledge in comparison to single-channel information coding [6]. This particularly applies when information is visualized and concurrently explained as in online video lectures: While the video lecture is playing, the educator's voice provides explanations, and the recorded slides illustrate the learning contents visually.

For designing and integrating a CA in this video-based setting, the most suited channel and the corresponding interaction type needs to be selected. From a theory perspective, the cognitive theory of multimedia learning draws on principles for choosing between visual and auditive information modalities, such as to "present words as speech rather than on-screen text" [48]. This modality principle is strongest when the material is complex for the learner and when the pace is fast and not under the learner's control [63]. This suggests that voice-based CAs might be generally beneficial over text-based CAs in learning settings. However, in our context of online video lectures the auditory channel is already addressed by the educator's voice while the video is playing. Thus, we argue that providing a voice-based CA might lead to an overflow of the auditory channel and thus might even perform worse than common text-based CAs. Thus, we argue that implementing a dual-channel approach consisting of a voice-based CA that also prints its interaction on the screen might overcome possible overflows and thus might outperform text-only CAs. As both channels contain exactly the same information, negative effects are not to be expected if learners only focus on one of them. Providing additional information on both channels during the interaction does not seem to be appropriate as the interaction should not be used to impart new knowledge in this setting. The interaction should rather deepen the previously learned knowledge by encouraging the learners to think about the video content. In prior CA research, Litman et al. [42] already added spoken

language capabilities to text-based dialog tutors but were not able to reveal positive effects [41]. However, more recent research shows that if learners also verbalize their knowledge (e.g., also respond to the CA via voice), they are able to improve their learning process [36, 53]. Thus, we propose:

H2a. Learners interacting with voice- and text-based CAs show higher levels of information retention compared to those interacting with textual CAs.

Past research also indicates that voice-based interactions between learners and educators help learners to transfer gained knowledge to solve novel problems [21]. For example, learners who receive audio narration solved more novel problems in a subsequent problem-solving test [21]. Thus, we propose:

H2b. Learners interacting with voice- and text-based CAs show higher levels of transfer ability compared to those interacting with textual CAs.

Voice-based instruction also has other benefits over a text-based approach. CAs' spoken output can elicit a warmer attitude among users towards the agent and lead to a richer use of language [53]. Learners also achieve a higher content average in the spoken condition and a better adjustment of communication when speaking compared to communicating via text with a CA [10]. Moreover, de Kock et al. [32] found out that learners who worked with audio/visual hints improved their information retention and solved more problems correctly in the transfer test compared to those who received text-based hints. Furthermore, Winkler et al. [72] compared voice-based SPAs with scripted human facilitators and showed that voice-based CA systems successfully approach scripted human facilitators in terms of learners' problem task. Voice-based instruction also has other benefits over a text-based approach.

CAs' spoken output can elicit a warmer attitude among users towards the agent and lead to a richer use of language [53]. Learners also achieve a higher content average in the spoken condition and a better adjustment of communication when speaking compared to communicating via text with a CA [10]. Moreover, de Kock et al. [32] found out that learners who worked with audio/visual hints improved their information retention and solved more problems correctly in the transfer test compared to those who received text-based hints. Furthermore, Winkler et al. [72] compared voice-based SPAs with scripted human facilitators and showed that voice-based CA systems successfully approach scripted human facilitators in terms of learners' problem task outcome quality. Thus, we believe that voice and text-based scaffolds provided by a CA are better able to increase learning success compared to textual scaffolds. Thus, we propose:

H3a. The scaffolding effect will be stronger for information retention when the CA is voice- and text-based rather than textual.

H3b. The scaffolding effect will be stronger for transfer ability when the CA is voice- and text-based rather than textual.

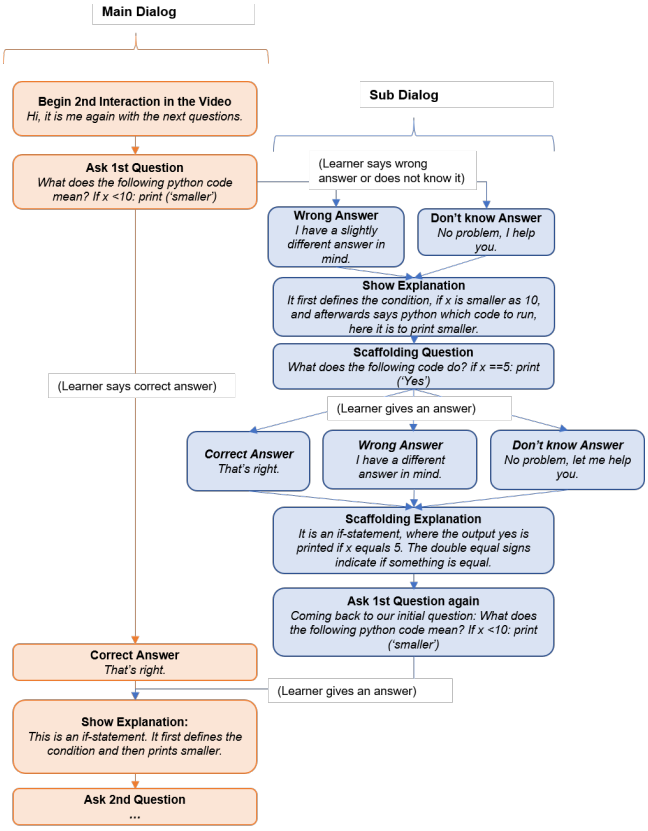

Figure 1. An example of Sara's scaffolding interaction logic.

SARA, A SCAFFOLDING-BASED CA

To investigate the hypotheses that we put forward above, we design a CA that we call Sara following three design principles: two coming from the *Theory of Scaffolding* and one design principle coming from the *Cognitive Theory of Multimedia Learning*.

The first design principle (DP1) is to include the *main- and sub-dialog* in the interaction model. It is derived from the prominent scaffolding strategy called *directed lines of reasoning*. Educators ask a series of directive questions to help the learner dive deeper into the topic [46]. If the learners' answers are correct, the educator advances to the next question. Otherwise, the educators try to scaffold the learners' understanding with additional questions that meet the learners' individual ZPD for the targeted concepts [1]. Figure 1 illustrates this design principle with the help of an example interaction between a learner and Sara. The video depicted in the interaction deals with learning how to create a conditional execution in Python. We implemented DP1 in our prototype as follows: Sara interrupts the video lecture after a specific time (i.e., a subchapter) and asks each learner a series of comprehension question to repeat the just seen and heard content. Sara detects three different kinds of learner's answers: *correct, wrong*, and *don't know* answers. When Sara detects a correct answer, she continues with the main dialog (left side of Figure 1).

When Sara detects a wrong answer or when the learner does not know how to reply, she opens a sub-dialog (right side) where she provides feedback, an explanation, and scaffolding questions within learner's ZPD that help them to close their knowledge gaps and find the answer for the initial question of the main dialog. After asking the scaffolding question, Sara explains the answer and asks the main question again. The learners have another chance to answer the question before Sara explains the answer and continues with the next question of the main dialog. Using this mechanism, Sara tries to imitate educators' scaffolding behavior when interacting with their learners. Finally, after Sara interacted with the learners for several minutes, Sara disappears, and the video lecture continues with the next piece of information. After the next section or segment, Sara interrupts the video lecture again and starts over with the next dialog.

The second design principle (DP2) is to use *appropriate diagnosis methods* to detect learners' state of knowledge. To provide appropriate scaffolds, the system has to include mechanisms that conduct some form of speech analysis that helps in deciding whether a learner's answer is right or wrong or if the learner does not know how to respond to the question [16]. To implement this principle, there are different ways of how CAs can detect the kind of answer given. In many cases, simple pattern matching is conducted to compare the inserted answer to one or multiple optimal solutions [34]. However, in open-topic natural language communication as in our context, defining all optimal solutions is hardly possible. There is not the one expression that describes the correct solution. Rather, different descriptions can show a correct solution. Thus, we used a more sophisticated solution based on a pre-trained model. We used the NLP.js framework [3] for the intent classification, which has been shown to be effective in recent benchmarks [13]. NLP.js is publicly available under https://github.com/axa-group/nlp.js. Compared to other cloud-based services, it has the advantage that all data is stored and processed in our own infrastructure, which is required in many educational settings (due to privacy regulations). For each question Sara asked the learners (see Figure 1), we trained one NLP model using a training set consisting of multiple possible *correct*, *wrong*, and *don't know* answers provided by seven learners' and three researchers' pretests resulting in a set of approx. 800 single statements. Once the learner answers a question, the voice-based answer is recorded, automatically transcribed to text using the HTML 5's Web Speech API [70], and the resulting textual data is then sent to our NLP server, where we apply our NLP models to classify the learner's answers using our pre-trained models. Using the results of this intent classification step (i.e., answer was correct, wrong, or student does not know the answer), we decide whether Sara should continue with the main dialog (correct answer, see Figure 1 left side) or if Sara should open the sub-dialog (wrong or don't know answer, see Figure 1 right side).

Our third design principle (DP3) is based on the *Cognitive Theory of Multimedia Learning* and deals with *addressing*

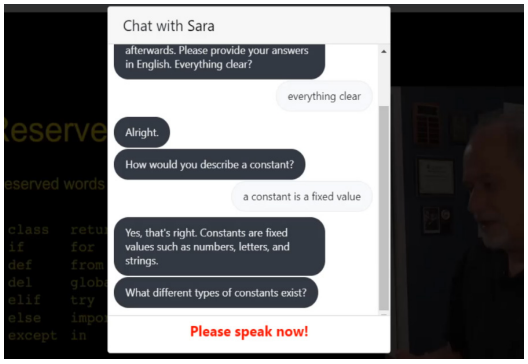

Figure 2. Screenshot of Sara during an online video lecture, shown on top the video screen.

multiple channels using voice and text communication modes. A central design principle of this theory is to reduce the load of a single processing channel by transmitting information via different channels. Moreover, the modality principle states that voice- and text-transmitted information is superior to text-transmitted information only [32]. To implement this, Sara starts talking to the learner via voice based on the HTML 5's Speech Synthesis API [70]. Sara generates a greeting statement via voice and asks the learner if he or she has understood the initial instructions. Additionally, Sara's spoken text is displayed in the chat window directly after she finishes her utterances to make sure that we use both channels one at a time (see Figure 2). As shown in Figure 2, we integrated Sara in a web-based video player. While the video lecture is playing, Sara is hidden. When a certain topic has been completed in the video lecture, Sara appears in front of it, and the video gets paused. Then the learner can interact with Sara based on her scaffolding strategy to deepen the learning content just seen. After the interaction is completed, the video continues until Sara appears again. We showed our learners two different video lectures. The timeline of video 1 includes the interaction and is displayed in Figure 3. We argue that with this educator-like type of interaction a better learning success can be achieved.

EXPERIMENTAL EVALUATION OF SARA

We designed an experiment to test our hypotheses and test Sara through a rigorous multi-comparison of different CA types. We used a 2 (non-scaffolding, scaffolding) x 2 (textual, voice- and text-based) x 2 (repeated measure, 2 random treatments per participant) design. We decided to use a repeated-measure design to show learners two videos to make sure that the results are not restricted to only one video content. We compared Sara to three other types of CAs that are

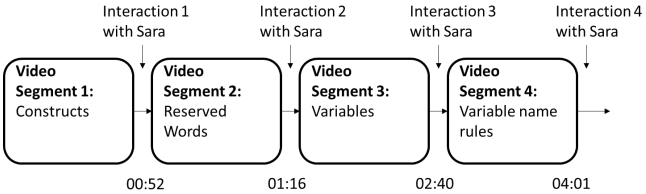

Figure 3. Timeline of video 1 and its interaction points.

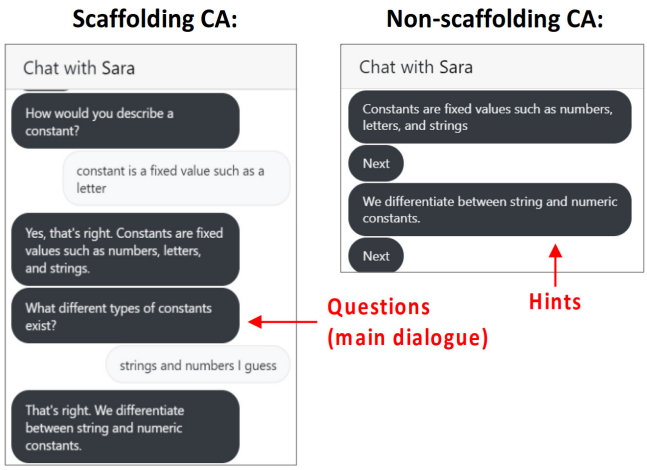

Figure 4. Scaffolding-based vs. non-scaffolding CA interaction. Left: Learners are asked for answers. Right: Learners see hints and can advance by clicking or saying "next".

often implemented in learning settings. The CA comparisons were between (a) Sara, (b) a voice- and text-based non-scaffolding CA, (c) a textual scaffolding CA and (d) a textual, non-scaffolding CA. Moreover, we included a control group where learners watched the videos only. Learners watched two online video lectures in which they interacted with one CA type per video. The CA interactions took place four times per video (except for the control group, where learners simply watched the video lecture). The learners received the treatments and the videos in a random order.

Figure 4 compares a scaffolding CA and a non-scaffolding CA. The non-scaffolding CA offered standardized, one-size-fits-all hints, where learners were able to click or say next. These kind of CAs have often been implemented in online learning settings in the past few years [33]. The hints included the same information as the main dialog of the scaffolding CA (see red rectangles in Figure 4). The only difference is that the scaffolding CA uses questions to interact with the learners and is able to open a scaffolding sub-dialog when learners provide a *wrong* or *don't know* answer. The voice- and text-based CA allows learners to provide their answers via voice. Moreover, Sara provides her questions and explanations via voice and additionally displayed the text on the screen (multi-channel, DP3). The text-based CA allows learners to type in answers via the keyboard, an input mode that is often seen in CAs in online education [59]. Our dependent variables (DVs) were the learners' information retention and transfer ability levels. We used a pretest and posttest design to measure the differences in the posttest scores of the DVs across the CA groups and the control group including the pretest scores, learning channel preference, and cognitive load as covariates (i.e., treatment scores).

Participants

We recruited undergraduate and graduate students (n = 182, 74 female, 108 male, aged 18 to 35) from the participant pool of a European business university. Table 1 shows the sample's characteristics across the four treatment groups and the control group. We randomly assigned the learners to one of the four treatment groups or the control group. They received 20 US dollars as a baseline and an optional 10 US dollars depending on their gain scores (difference between the pretest and posttest). The randomization was successful since independent samples *t*-tests revealed that all the CA groups were similar in terms of *gender, age, previous experience with CAs, personal innovativeness, language level and nationality* ($p > .05$).

	Non-scaffolding, textual	Scaffolding, textual	Non-scaffolding, voice-based	Sara	Control group (video only)
Sample size	37	36	37	37	35
Gender (M/F)	23/14	21/15	22/15	21/15	23/12
Avg. age	22.4	22.8	23.1	23.0	22.2
Prev. exp. with CAs (usage per week)	3.6	3.7	3.9	3.5	3.9
Personal innovativeness (out of 7)	4.7	4.8	4.9	4.8	4.8
Language level (out of 7, 7 = proficient)	5.4	5.6	5.2	5.3	5.5

Table 1. Sample characteristics.

Materials

Videos

Each learner watched two beginner-level videos about Python programming in a randomly assigned order. We chose these videos for two reasons. First, we wanted to make sure that the learners had little previous knowledge on the topic. Since the study participants were business school learners that do not attend a programming class, the topic was suitable. Second, we wanted to make sure that the videos were representative in terms of quality, typicality, and video format. Both videos are from a well-known and popular Python course taught by Charles R. Severance from the University of Michigan School of Information, publicly available under https://www.py4e.com/. Being about Python programming, the topic was representative among online learning material where the amount of technically oriented learning content has increased enormously in recent years [37]. The video format was web-based and picture-in-picture, which is very often used in MOOCs [31]. The videos we chose for the experiment were about constants and variables (video 1; length 4 min 52 s without CA interaction) and about conditional execution (video 2; length 4 min 48 s without CA interaction). The videos presented these contents largely independently of each other. Every video started with a 10 second resting period (black screen with white text "Resting Period"). We asked the learners in the post-survey to rate the difficulty of the two videos on a scale from 1 to 7 to make sure that the two videos were equally difficult for all the learners and that possible effects are not explained by differences in the difficulty of the videos. The learners' ratings of the videos' perceived difficulty were on the same level (Video 1: 5.43 out of 7, Video 2: 5.63 out of 7, $p > .05$).

Tests and Questionnaires

The pre-survey contained questions regarding learners' tendency to a visual vs auditive learning style (obtained from [55]). We included this variable as a covariate in our data analysis to factor out that possible differences in the CA groups came from different learning styles. The pre-stimulus test contained six multiple choice questions (3 questions per video) and two open problem questions (1 question per video) to measure learners' relevant pre-knowledge level on programming with Python. The post-stimulus test followed directly after the stimulus and contained exactly the same questions as in the pre-stimulus test. All the test questions were presented in a random order. The six multiple choice questions measured *information retention* and the two open questions measured *transfer ability*. The multiple-choice questions related to information retention asked the learners to reproduce content from the video. The open questions required learners to independently apply the acquired knowledge to a novel problem. For example, in the multiple-choice question part, we asked learners to mark the statement that best corresponds to the description of a constant. In the transfer question part, the students actively solved a small programming exercise. We asked them for instance to define and name two new variables, which assign "Martin" and "7", and to print them as output. The post-stimulus survey contained items of cognitive load (from [39]), perceived difficulty of the videos, personal innovativeness (from [66]), language proficiency, personal experience with CAs, age, gender, nationality, and one open question about how the learners experienced the CA. We used cognitive load scores as covariates in our data analysis to control for differences in cognitive load when using different CA types.

Procedure

The experimental procedure is depicted in Figure 5. The learners carried out the experiment individually and one after the other. After obtaining consent, the experimenter introduced the experiment to the learner and the learner started with the pre-survey and pre-stimulus test in room 1 in absence of the experimenter. After conducting the pre-survey and pre-stimulus test, the experimenter returned and escorted the learner into room 2, which contained a chair facing a computer monitor. The experimenter instructed the learner to pay attention to the two lectures as there would be a short quiz on the material afterwards. He then left, after which the participant watched the stimulus without pauses and was not allowed to take notes. Once the two video lectures were over, the experimenter escorted the participant to room 1 again for conducting the post-stimulus test and post-survey. After completion, the participant was asked to guess the purpose of the study and was then debriefed on the experiment. Most

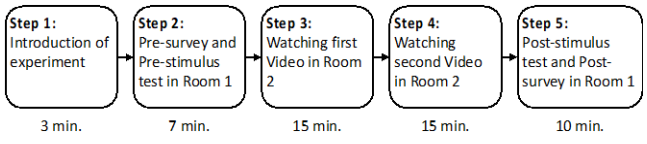

Figure 5. Experimental Procedure.

of the participants thought the study was on how to effectively learn programming with Python.

MEASUREMENT AND DATA ANALYSIS

For measuring the DVs – *information retention* and *transfer ability* – two experienced raters evaluated the results using an evaluation framework that was developed collectively in a workshop within the research team. In the multiple-choice question section, the evaluation framework provided the correct answer and specified how many points to give. For each correctly answered question, we distributed one point. In the open transfer question section, the evaluation framework specified exactly what points were to be awarded for what and how many points should be deducted for common mistakes (e.g., 2 points given for the correct syntax of the if-statement, 2 points given for the correct semantics, 0.5 point reduction if colon is missing, etc.).

The post-stimulus test results were rated individually, blinded (raters do not know the CA group), and independently from each other. Furthermore, for the transfer open question section, we checked for interrater agreement with the help of a Pearson correlation. Interrater agreement gives a score of how much homogeneity, or consensus, there is in the ratings [12]. The correlation resulted in a satisfactory result (interrater agreement = 0.94, $p > .05$). To identify differences between groups who had interacted with the CA types, we first conducted ANCOVAs to see if there exist differences across the groups including the control group. ANCOVAs help to identify whether there are significant differences in variances between the groups while controlling for covariates. We did not use ANOVA because we needed to factor out errors that were introduced by covariates and it masks the true relationship between the type of CA and the dependent variables. We thus included the pretest scores, cognitive load, and the learning channel preference (auditive or visual) as covariates. All these covariates have the potential to disturb the true effect between the CA and learning outcomes. We also verified that the data meets the assumptions of an ANCOVA (normality, equality of slopes of the covariates and the outcome variable, equality of the groups and the covariates, and homogeneity of variance) with *t*-tests and a Levene's test [38]. All assumptions were met.

After the ANCOVAs, we investigated the patterns and comparisons between specific groups in order to test our hypotheses. Although otherwise powerful, ANCOVA cannot provide such results. Thus, we conducted a post-hoc analysis using Tukey's Test [65] to test our hypotheses H1 and H2. Tukey's test compares the means of all treatments to the mean of every other treatment and is considered the best available method in cases when confidence intervals are desired or when sample sizes are unequal. To analyze the interaction effect between *non-scaffolding/scaffolding* and *text-based/text-* and *voice-based*, we used factorial ANCOVAs with both factors as independent variables to test H3. We calculated partial eta squared as a measure of the strength of an effect (0.01 = small, 0.06 = medium, 0.14 = large, [7]).

	Information retention			Transfer ability		
	Pre-test	Post-test	SD post	Pre-test	Post-test	SD post
Non-scaffolding, textual	40.3	61.6	24.6	31.5	50.6	24.3
Scaffolding, textual	39.8	73.1	22.0	264.	54.6	27.1
Non-scaffolding, voice-based	38.0	65.3	23.2	32.6	51.5	25.7
Sara	34.8	77.8	21.7	26.0	65.4	27.1
Control group (video only)	36.2	60.9	27.2	28.5	44.0	21.9

Table 2. Normalized pretest and posttest scores for information retention and transfer ability for different conditions.

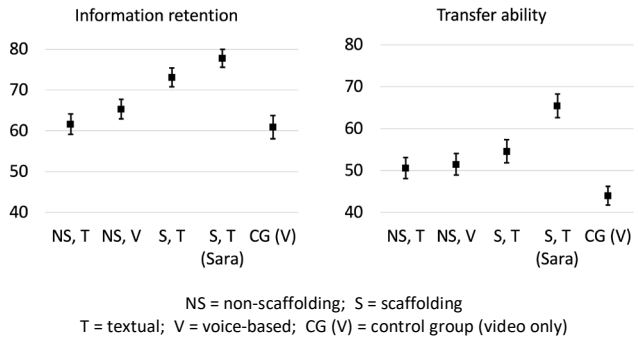

Figure 6. Normalized post-test scores for information retention and transfer ability for different conditions. 50% of the post-test scores are within the error bars.

RESULTS

Descriptive analysis

The main goal of our study was to investigate whether our scaffolding-based CA, Sara, is able to increase learners' information retention and transfer ability during online video lectures compared to other often implemented CA types. Table 2 shows a summary of the pretest and posttest scores of the groups for the two different types of learning outcomes: *information retention* and *transfer ability*. Since we had different maximum points in the tests of the two videos, we normalized all the scores between 0 and 100, where 100 corresponds to the maximum number of points scored at the pretest and post-stimulus test. Descriptive statistics for both *information retention* and *transfer ability* provide similar results: Learners interacting with Sara showed both the highest level of *information retention* (score = 77.8) and the highest level of *transfer ability* (score = 65.4). These scores were followed by a textual scaffolding-based CA (73.1 and 54.6), a voice- and text-based non-scaffolding CA (65.3 and 51.5), and a textual non-scaffolding CA (61.6 and 50.6). The control group that did not interact with a CA had the lowest score in both outcome variables (60.9 and 44.0).

Scaffolding

The ANCOVA test for *information retention* yielded a significant main effect for the scaffolding-based CA when the pre-test, cognitive load, and the learning channel preference were controlled ($F(4,359) = 29.9$, $p < .0001$***, $n = 182$). Partial eta-squared was 0.08, which is considered a medium effect strength [7]. Tukey's test revealed that the scaffolding-based CA was significantly better in supporting information retention compared to non-scaffolding CAs (*adjusted* $p < .0001$***).

Thus, we **observe support for H1a**. Furthermore, we calculated the intervention selection accuracy (ISA) for the two groups that interacted with the scaffolding-based CA. We define the ISA in our context by the number of correctly assigned scaffolding sub-dialogs (i.e., every time the classifier tagged a student answer correctly as "wrong" or "don't know" and thus opened a scaffolding sub-dialog, see right side of Figure 1) divided by the total number of all sub-dialogs given [58]. The research team tagged the interaction logs and found that in total 51 out of 73 students entered the sub-dialog part of our scaffolding mechanism at least once with 199 sub-dialogs in total (out of 1168 potentially possible sub-dialogs). In the text-based CA group, there were 96 right interventions out of 102. In the CA Sara group, there were 94 out of 97. This leads to an ISA of 95.48%. Although some students of the CA Sara group did not enter the scaffolding sub-dialogs, we argue that this aspect does not affect our results and implications: Even the learners that did not enter the sub-dialogs interacted at least with the CA during the main dialog questions (see Figure 1). Thus, we argue that there has still taken place a manipulation as a part of asking those main questions during the experiment. This interaction might have already stimulated the learners to reflect and think about the learning content even though they did not enter the sub-dialogs. Excluding those learners would disregard the effect of the main dialog interaction. Consequently, we did not exclude those learners from our analysis.

The mean differences and the 95% confidence intervals between the specific groups (control group, non-scaffolding, scaffolding) are displayed in Figure 7 (top left). If an interval does not contain zero, the corresponding means are significantly different. The ANCOVA for *transfer ability* also yielded a significant main effect for the scaffolding-based CA, again controlling for the pretest, cognitive load, and the learning channel preference ($F(4,359) = 22.3$, $p < .0001$***, $n = 182$). Partial eta-squared was 0.06, yielding again a medium effect strength. The scaffolding-based CA was significantly better than the non-scaffolding CA in terms of transfer ability (*adjusted* $p = 0.03$**). Thus, we **observe support for H1b**. The confidence intervals between the specific groups are displayed in Figure 7 (top right).

Textual vs Voice- and Text-Based Modality

The ANCOVA for *information retention* yielded a main effect for the voice- and text-based CA, with the pretest, cognitive load, and the learning channel preference controlled ($F(4, 359) = 12.5$, $p = 0.0005$***, $n = 182$). Partial eta-squared was 0.04 – a small effect strength. The voice- and text-based CA group was significantly better than the control group (*adjusted* $p = 0.001$***) but not significantly better

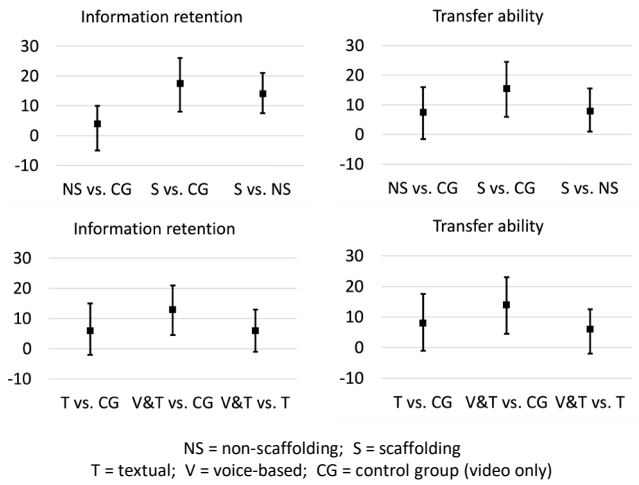

Figure 7. Mean differences in scaffolding manipulation (top) and voice manipulation (bottom).

than the textual CA (adjusted $p = 0.098$). Figure 7 shows the mean differences and the corresponding confidence intervals between the specific groups (bottom left). Thus, **H2a needs to be rejected.** Finally, the ANCOVA for *transfer ability* yielded significant effects for the voice- and text-based CA controlling for the other variables ($F(4,359)=17.2$, $p < .0001$***, $n = 182$). Eta-squared was 0.47 – a small effect. Tukey's test revealed that the voice- and text-based CA was significantly better than the control group (adjusted $p = 0.001$) but not significantly better than the textual CA. Thus, **H2b needs to be rejected**. The mean differences and confidence intervals between the specific groups are depicted in Figure 7 (bottom right).

Interaction effect

Regarding the interaction effect for *information retention*, the 2 x 2 ANCOVA revealed a significant effect between the scaffolding mechanism and the modality ($F(1,359)=11.0$, $p = 0.0001$***). Partial eta-squared was 0.03 – a small effect strength. This result suggest that the effect of a scaffolding-based CA is stronger when the CA is voice-based. Thus, **H3a is supported by our data**. For *transfer ability*, we also saw a significant interaction between non-scaffolding/scaffolding and the voice-based/modality ($p = 0.009$***). Partial eta-squared was 0.019, again a small effect. Thus, **H3b is supported by our data**.

DISCUSSION

This study aimed at investigating the effect of scaffolding and voice-based CAs layered on top of online video lectures to improve information retention and transfer ability. We proposed that a voice- and text-based CA that scaffolds learners' understanding during video instruction increases information retention (i.e., helps them remember concepts better) and transfer ability (i.e., helps them apply the acquired knowledge to solve novel problems). The underlying rationale for these hypotheses was that scaffolding and voice- and text-based CAs are able to detect knowledge gaps of learners, can react accordingly, and are able to freely interact with them similarly to the gold standard of educator-learner interactions.

Hypotheses 1a and 1b confirmed that scaffolding-based CAs can better create a meaningful interaction with the learners compared to non-scaffolding-based CAs resulting in better information retention and transfer ability. Scaffolding-based CAs might help learners to dive deeper into the topic compared to non-scaffolding-based, more traditional CAs. And although voice- and text-based CAs are not per se better than textual CAs (Hypotheses 2a and 2b), we found that voice-based CAs can further strengthen the scaffolding effect for information retention and transfer ability (Hypotheses 3a and 3b). Our study makes several implications for the design of future CAs in online education. We thereby build on previous findings that investigated beneficial effects of CAs in online learning and similar scenarios. For example, Ruan et al. [59] discovered that QuizBot yields increases in factual knowledge compared to more common learning aids, and Lin et al. [40] discovered that lessons instructed in a conversational style enhanced information retention. Litman et al. [41] compared voice-based vs. text-based tutoring and found out that adding spoken language capabilities increases the performance in human tutoring but not in computer tutoring.

Our findings advance this line of research in several ways. First, we can confirm past research that emphasized the importance of scaffolding mechanisms in CAs also in an online video lecture setting. Scaffolding mechanisms can enable CAs to identify learners' individual knowledge gaps, react to them, and bring the learner back on board again. In the open question part of our post-survey, one of the learners mentioned the following: "*Great experience! Sara helped me to think about what was said by the instructor and offered help when I was wrong. She helped me to structure my thinking processes.*" Compared to more traditional CAs, scaffolding-based CAs do not immediately show the correct answer but try to help the learner build up the missing knowledge and come up with the solution themselves. Especially in the context of online video lectures, it is very important that learners have the possibility to receive this kind of individual support right after they do not understand something. Otherwise, they may not be able to process the content that follows, which might result in frustration and not so effective learning processes.

Second, our findings recommend equipping future CA designs with speech recognition. This is because voice-based CAs might be able to build a more meaningful interaction with the user compared to textual ones. Although we did not see a main effect between textual and voice-based CAs, we were able to detect an interaction effect indicating that scaffolding is more powerful when it is voice and text-based. We thereby connect to research that compares text-based with voice-based CAs. In contrast to Litman et al. [41], we were able to show that including voice in computer tutoring dialogs is more effective than text-based only. This extends the

results of Novielli et al. [53] that showed beneficial effects of voice on a social attitude towards the CA. Moreover, we extend the modality principle [49] by indicating that voice is only beneficial when it is implemented in combination with a scaffolding mechanism in a CA context. This might happen because a voice- and scaffolding-based CA is better able to imitate a real educator-learner interaction. One learner said the following: "*T[h]ough the chatbot was not able to understand me all the time, it felt a bit like talking to a real person.*" Our results suggest that it is not enough to use voice-based CAs in online education. The effect of voice-based systems only comes to bear once the CAs have implemented an educator-like teaching strategy. For a CA to partially imitate an educator, it is essential that the learner recognizes the CA as a social actor. According to the *Computers Are Social Actors* (CASA) paradigm, people use similar social rules when dealing with computers as with people [52]. Researchers in the CASA paradigm have assumed that, because a computer agent and a human have similar features, the users' social responses are amplified, thus enabling effective interaction with computers [51]. Embedding a more human-like, voice-based interaction might allow learners to perceive a higher level of social presence. This feeling of social presence is usually missing in online learning environments and can influence the learning behavior significantly [56].

Third, our study indicates that publicly available NLP frameworks offer the possibility that CAs can interact with learners freely resulting in a higher quality of interaction. When learners need to articulate their answers in a free manner, they need to make more cognitive effort, similarly to an interaction with a human educator. This interaction mode can additionally stimulate learners' thinking processes, which ultimately leads to better learning outcomes [26]. As one learner put it: "*It was fascinating how much it understood and could say if I was correct.*" However, our research also shows that NLP still has a lot of room for improvement and that the quality of interaction becomes higher as technology advances. Some learners were not fully satisfied with the way Sara understood their answers. One learner mentioned: "*The system could not always understand my voice, this is annoying.*" With rapid advancements in NLP, future designs of CAs could use publicly available NLP modules to give learners the opportunity to interact freely with the CAs. Overall, our work provides empirical evidence that using scaffolding and voice-based CAs in online video lectures is beneficial. Nevertheless, introducing it in real online learning environments still requires severe efforts by educational institutions and lecturers. However, the possibilities for the development and integration of CAs are becoming easier. For example, today it is already possible to develop CAs without programming knowledge [68]. In addition, the integration of CAs in online video lecture content offers a very promising possibility to create individual and meaningful interactions in a scalable and personal resource-saving way.

Limitations and Future research

Our study does not come without limitations. First, although our language model almost always recognized a *right*, *wrong* or *don't know* answer, it made some mistakes, which quickly frustrated learners and possibly influenced their learning process. Moreover, the quality of interaction was heavily dependent on our training data collected from pretests. It would be very interesting to see if a CA using different NLP frameworks with different training data show similar results.

Second, interacting with CAs in a laboratory setting might be different to real life environments. Learners might be more motivated when using the new technology for the first time which might result in a higher quality of interaction. To partially compensate for that, we asked learners about their pre-experience with CAs, which resulted in a rather high experience level. Future research should investigate scaffolding-based CAs in real online courses in which learners interact with the CAs over a whole semester. With the help of qualitative interviews, additional insights can be found out.

Third, we tested our CA Sara in a rather narrow context (programming knowledge). We have tried to choose a context that is very representative today for online courses. Since the number of programming online courses has increased extremely, we have decided to choose such a course. Nevertheless, scaffolding and voice-based CAs may well achieve other effects in other contexts. Future research should try to find out to what extent the subject or domain has an influence on the effectiveness of scaffolding-based CAs in online education. This study contributes to the question of how to create meaningful interactions in large-scale online courses. The success of CAs in online learning environments will very much depend on the extent to which educational institutions and lecturers succeed in integrating CAs themselves. Hence, future research should focus on the development of appropriate authoring tools, which make it possible to create CAs without much programming knowledge. These authoring tools should make it easier to design scaffolding-based CAs including our proposed design principles.

Given that our results from scaffolding- and voice-based CAs are successful, our design principles could also have an effect on areas other than education. For example, such CAs could also be incorporated into explanatory videos, such as video tutorials on using smart systems, voting videos, etc.

CONCLUSION

In our study, we investigated to what extent a voice-based and scaffolding CA can solve the challenge of creating meaningful interactions in online video lectures. Thereby, we proposed three central design principles: main- and sub-dialog (DP1), appropriate diagnosis methods (DP2), and addressing multiple channels (DP3). The instantiation of these three design principles in our CA Sara has led to a significant increase in learners' information retention and transfer ability. This effect is even stronger in voice-based CA systems. Our results show that the use of CAs in online education and especially on top of online video lectures is promising.

REFERENCES

[1] Patricia Albacete, Pamela Jordan, Dennis Lusetich, Irene A. Chounta, Sandra Katz, and Bruce M. McLaren. 2018. Providing Proactive Scaffolding During Tutorial Dialogue Using Guidance from Student Model Predictions. In *Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. U. Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren and Benedict Du Boulay, Eds. Lecture Notes in Computer Science. Springer International Publishing, Cham, 20–25. DOI: https://doi.org/10.1007/978-3-319-93846-2_4.

[2] Arthur N. Applebee and Judith A. Langer. 1983. Instructional scaffolding: Reading and writing as natural language activities. *Language arts* 60, 2, 168–175.

[3] AXA Group Operations. 2018. *NLP.js Framework* (2018). Retrieved from https://github.com/axa-group/nlp.js.

[4] Ryan S. Baker. 2016. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education* 26, 2, 600–614.

[5] Kay S. Bull, Paul Shuler, Robert Overton, Sarah Kimball, Cynthia Boykin, and John Griffin. 1999. Processes for Developing Scaffolding in a Computer Mediated Learning Environment.

[6] James M. Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational Psychology Review* 3, 3, 149–210.

[7] Jacob Cohen. 1973. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and psychological measurement* 33, 1, 107–112.

[8] Ronald N. Cortright, Heidi L. Collins, and Stephen E. DiCarlo. 2005. Peer instruction enhanced meaningful learning: ability to solve novel problems. *Advances in physiology education* 29, 2, 107–111.

[9] Erhan Delen, Jeffrey Liew, and Victor Willson. 2014. Effects of interactivity and instructional scaffolding on learning: Self-regulation in online video-based environments. *Computers & Education* 78, 312–320.

[10] Sidney K. D'Mello, Art Graesser, and Brandon King. 2010. Toward Spoken Human-Computer Tutorial Dialogues. *Human-Computer Interaction* 25, 4, 289–323. DOI: https://doi.org/10.1080/07370024.2010.499850.

[11] Oliver Ferschke, Diyi Yang, Gaurav Tomar, and Carolyn P. Rosé. 2015. Positive Impact of Collaborative Chat Participation in an edX MOOC. In *Artificial Intelligence in Education*, Cristina Conati, Neil Heffernan, Antonija Mitrovic and M. F. Verdejo, Eds. Lecture Notes in Computer Science. Springer International Publishing, Cham, 115–124. DOI: https://doi.org/10.1007/978-3-319-19773-9_12.

[12] Joseph L. Fleiss, Bruce Levin, and Myunghee C. Paik. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions* 2, 212-236, 22–23.

[13] Git Hub. 2019. *NLP Framework BenchmarKING* (2019). Retrieved from https://github.com/axa-group/nlp.js/blob/master/docs/benchmarking.md.

[14] Arthur Graesser, Haiying Li, and Carol Forsyth. 2014. Learning by Communicating in Natural Language with Conversational Agents. *Grantee Submission*.

[15] Arthur C. Graesser. 2011. Learning, thinking, and emoting with discourse technologies. *American psychologist* 66, 8, 746.

[16] Arthur C. Graesser, Zhiqiang Cai, Brent Morgan, and Lijia Wang. 2017. Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior* 76, 607–616.

[17] Arthur C. Graesser, Xiangen Hu, and Robert Sottilare. 2018. Intelligent tutoring systems. In *International handbook of the learning sciences*. Routledge, 246–255.

[18] Arthur C. Graesser, Natalie Person, Derek Harter, and Tutoring Research Group. 2001. Teaching tactics and dialog in AutoTutor. *Int J Artif Intell Educ* 12, 3, 257–279.

[19] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Appl. Cognit. Psychol.* 9, 6, 495–522. DOI: https://doi.org/10.1002/acp.2350090604.

[20] Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny T.-Z. Wei, Joseph J. Williams, and Sharad Goel. 2019. MathBot: Transforming Online Resources for Learning Math into Conversational Interactions.

[21] Egbert G. Harskamp, Richard E. Mayer, and Cor Suhre. 2007. Does the modality principle for multimedia learning apply to science classrooms? *Learning and Instruction* 17, 5, 465–477. DOI: https://doi.org/10.1016/j.learninstruc.2007.09.010.

[22] Mariane Hedegaard. 2002. The zone of proximal development as basis for instruction. In *An introduction to Vygotsky*. Routledge, 183–207.

[23] Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science* 349, 6245, 261–266.

[24] Woei Hung. 2013. Problem-based learning: A learning environment for enhancing learning transfer. *New directions for adult and continuing education* 137, 27–38.

[25] Mohamed Ibrahim and Osama Al-Shara, Eds. 2007. *Impact of interactive learning on knowledge retention*. Springer.

[26] Fergus Im Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior* 11, 6, 671–684.

[27] Alice Kerly, Richard Ellis, and Susan Bull. 2008. CALMsystem. A Conversational Agent for Learner Modelling. *Knowledge-Based Systems* 21, 3, 238–246. DOI: https://doi.org/10.1016/j.knosys.2007.11.015.

[28] Alice Kerry, Richard Ellis, and Susan Bull. 2009. Conversational Agents in E-Learning. In *Applications and Innovations in Intelligent Systems XVI*, Tony Allen, Richard Ellis and Miltos Petridis, Eds. Springer London, London, 169–182. DOI: https://doi.org/10.1007/978-1-84882-215-3_13.

[29] Beaumie Kim. 2001. Social constructivism. *Emerging perspectives on learning, teaching, and technology* 1, 1, 16.

[30] Minchi C. Kim and Michael J. Hannafin. 2011. Scaffolding problem solving in technology-enhanced learning environments (TELEs). Bridging research and theory with practice. *Computers & Education* 56, 2, 403–417.

[31] René F. Kizilcec, Kathryn Papadopoulos, and Ladila Sritanyaratana. Showing face in video instruction: effects on information retention, visual attention, and affect. In *Proceedings of the SIGCHI14 conference systems. ACM.*

[32] Willem D. de Kock. 2016. Speech versus text supported hints in learning to solve word problems. *Computers in Human Behavior* 57, 300–311.

[33] James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems. *Review of educational research* 86, 1, 42–78. DOI: https://doi.org/10.3102/0034654315581420.

[34] Annabel M. Latham, Keeley A. Crockett, David A. McLean, Bruce Edmonds, and Karen O'shea. 2010 - 2010. Oscar: An intelligent conversational agent tutor to estimate learning styles. In *International Conference on Fuzzy Systems*. IEEE, 1–8. DOI: https://doi.org/10.1109/FUZZY.2010.5584064.

[35] D. Lederman. 2018. *Online Education Ascends* (2018). Retrieved from https://www.insidehighered.com/digital-learning/article/2018/11/07/new-data-online-enrollments-grow-and-share-overall-enrollment.

[36] Ioana Lepadatu. 2012. Use self-talking for learning progress. *Procedia - Social and Behavioral Sciences* 33, 283–287.

[37] Marina Lepp, P. Luik, P. Tauno, P. Kaspar, S. Reelika, Säde Merilin, and Edo Tõnisson, Eds. 2017. *MOOC in programming: A success story*.

[38] Howard Levene. 1960. Contributions to probability and statistics. *Essays in honor of Harold Hotelling*, 278–292.

[39] Lijia Lin, Robert K. Atkinson, Robert M. Christopherson, Stacey S. Joseph, and Caroline J. Harrison. 2013. Animated agents and learning: Does the type of verbal feedback they provide matter? *Computers & Education* 67, 239–249.

[40] Lijia Lin, Paul Ginns, Tianhui Wang, and Peilin Zhang. 2020. Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education* 143, 103658. DOI: https://doi.org/10.1016/j.compedu.2019.103658.

[41] Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt Vanlehn, Dumisizwe Bhembe, and Scott Silliman. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education* 16, 2, 145–170.

[42] Diane J. Litman and Silliman Scott, Eds. 2004. *ITSPOKE: An intelligent tutoring spoken dialogue system*. Association for Computational Linguistics.

[43] Richard E. Mayer. 2003. The promise of multimedia learning: using the same instructional design methods across different media. *Learning and Instruction* 13, 2, 125–139.

[44] Richard E. Mayer and Roxana Moreno. 1998. A cognitive theory of multimedia learning: Implications for design principles. *Journal of educational psychology* 91, 2, 358–368.

[45] Michael McTear, Zoraida Callejas, and David Griol. 2016. The conversational interface. *Springer* 6, 94, 102.

[46] Bruce Mills, Martha Evens, and Reva Freedman. 2004 - 2004. Implementing directed lines of reasoning in an intelligent tutoring system using the atlas planning environment. In *International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004*. IEEE, 729-733 Vol.1. DOI: https://doi.org/10.1109/ITCC.2004.1286554.

[47] Joao L. Z. Montenegro, Cristiano A. da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications* 129, 56–67. DOI: https://doi.org/10.1016/j.eswa.2019.03.054.

[48] David R. Moore. 2006. E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. *Education Tech Research Dev* 54, 2, 197–200.

[49] Roxana Moreno. 2006. Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning* 22, 3, 149–158.

[50] Nicola Murphy and David Messer. 2000. Differential benefits from scaffolding and children working alone. *Educational Psychology* 20, 1, 17–31.

[51] Clifford Nass, Youngme Moon, John Morkes, Eun-Young Kim, and B. J. Fogg. 1997. Computers are social actors: A review of current research. *Human values and the design of computer technology* 72, 137–162.

[52] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 72–78. DOI: https://doi.org/10.1145/191666.191703.

[53] Nicole Novielli, Fiorella de Rosis, and Irene Mazzotta. 2010. User attitude towards an embodied conversational agent. Effects of the interaction mode. *Journal of Pragmatics* 42, 9, 2385–2397. DOI: https://doi.org/10.1016/j.pragma.2009.12.016.

[54] Benjamin D. Nye, Arthur C. Graesser, and Xiangen Hu. 2014. AutoTutor and Family. A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education* 24, 4, 427–469.

[55] Lynn O'Brien. 1989. Learning styles: Make the student aware. *NASSP Bulletin* 73, 519, 85–89.

[56] Jennifer Richardson and Karen Swan. 2003. Examing social presence in online courses in relation to students' perceived learning and satisfaction. *Faculty and Staff Publications and Research - Center for Online Learning, Research and Service (COLRS)* 1, 3.

[57] Roman Rietsche, Kevin Duss, Jan M. Persch, and Matthias Soellner. 2018. Design and Evaluation of an IT-based Formative Feedback Tool to Foster Student Performance. In *39th International Conference on Information Systems (ICIS)*, San Francisco, CA, USA, 1–17.

[58] Carolyn Rosé and Kurt Vanlehn. 2005. An evaluation of a hybrid language understanding approach for robust selection of tutoring goals. *Int J Artif Intell Educ* 15, 4, 325–355.

[59] Sherry Ruan, Liwei Jian, Justin Xu, B. Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[60] Perez M. Santos, Eva Gonzalez-Parada, and M. Jose. 2013. Mobile embodied conversational agent for task specific applications. *IEEE Transactions on Consumer Electronics* 59, 3, 61-614. DOI: https://doi.org/10.1109/TCE.2013.6626246.

[61] Julia E. Seaman, I. E. Allen, and Jeff Seaman. 2018. *Grade Increase: Tracking Distance Education in the United States* (2018). Retrieved from https://www.onlinelearningsurvey.com/highered.html.

[62] Donggil Song, Eun Y. Oh, and Marilyn Rice. 2017. Interacting with a conversational agent system for educational purposes in online courses. In *2017 10th International Conference on Human System Interactions (HSI). Proceedings : International Hall, University of Ulsan, Ulsan, Republic of Korea, July 17-19, 2017*. IEEE, Piscataway, NJ, 78–82. DOI: https://doi.org/10.1109/HSI.2017.8005002.

[63] Huib K. Tabbers, Rob L. Martens, and Jeroen J. G. van Merriënboer. 2004. Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British journal of educational psychology* 74, 1, 71–81.

[64] Chih-Hsiung Tu. 2000. On-line learning migration: from social learning theory to social presence theory in a CMC environment. *Journal of Network and Computer Applications* 23, 1, 27–37. DOI: https://doi.org/10.1006/jnca.1999.0099.

[65] John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5, 2, 99–114.

[66] Erik M. van Raaij and Jeroen J. L. Schepers. 2008. The acceptance and use of a virtual learning environment in China. *Computers & Education* 50, 3, 838–852.

[67] Kurt Vanlehn, Arthur C. Graesser, G. T. Jackson, Pamela Jordan, Andrew Olney, and Carolyn P. Rosé. 2007. When are tutorial dialogues more effective than reading? *Cognitive science* 31, 1, 3–62. DOI: https://doi.org/10.1080/03640210709336984.

[68] VoiceFlow. 2019. *Design, prototype and build voice apps* (2019). Retrieved from https://www.voiceflow.com/.

[69] Lev S. Vygotsky. 1978. *Mind in society. The development of higher mental process*. Cambridge, MA: Harvard University Press.

[70] W3C Community and Business Groups. 2019. *Speech API* (2019). Retrieved from https://w3c.github.io/speech-api/.

[71] Dongqing Wang, Hou Han, Zehui Zhan, Jun Xu, Quanbo Liu, and Guangjie Ren. 2015. A problem solving oriented intelligent tutoring system to improve students' acquisition of basic computer skills. *Computers & Education* 81, 102–112.

[72] Rainer Winkler, Matthias Söllner, Neuweiler Maya Lisa, Flavia C. Rossini, and Jan M. Leimeister. 2019. Alexa, can you help us solve this problem? How conversations with smart personal assistant tutors increase task group outcomes. In *CHI'19 Conference on Human Factors in Computing Systems Extended Abstract*. SIGCHI, 1–6.

[73] David Wood, Jerome S. Bruner, and Gail Ross. 1976. The role of tutoring in problem solving. *Journal of child psychology and psychiatry* 17, 2, 89–100.