Received: 28 November 2021 | Revised: 20 April 2022 | Accepted: 26 May 2022

DOI: 10.1111/jcal.12704

ARTICLE

Journal of Computer Assisted Learning WILEY

The influences of a virtual instructor's voice and appearance on learning from video lectures

Zhongling Pi[1,2] | Lixia Deng[2] | Xu Wang[3] | Peirong Guo[3] | Tao Xu[3] | Yun Zhou[2]

[1]Key Laboratory of Modern Teaching Technology (Ministry of Education), Shaanxi Normal University, Xi'an, Shaanxi Province, PR China

[2]Faculty of Education, Shaanxi Normal University, Xi'an, Shaanxi Province, PR China

[3]School of Software, Northwestern Polytechnical University, Xi'an, Shaanxi Province, PR China

Correspondence

Tao Xu, School of Software, Northwestern Polytechnical University, 127 West Youyi Road, Beilin District, Xi'an 710072, Shaanxi Province, PR China.
Email: xutao@nwpu.edu.cn

Yun Zhou, Faculty of Education, Shaanxi Normal University, South Chang'an Road 199, Yanta District, Xi'an 710062, Shaanxi Province, PR China.
Email: zhouyun@snnu.edu.cn

Funding information

Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, Grant/Award Number: 2021-05-029-BZPK01; National Natural Science Foundation of China, Grant/Award Numbers: 62007023, 62077036; National Key Research and Development Program of China, Grant/Award Number: 2018AAA0100501

Abstract

Background: Video lectures which include the instructor's presence are becoming increasingly popular. Presenting a real human does, however, entail higher financial and time costs in making videos, and one innovative approach to reduce costs has been to generate a virtual speaking instructor.

Objectives: The current study examined whether the use of a virtual instructor in video lectures would facilitate learning as well as a human instructor, and whether manipulating the virtual instructor's characteristics (i.e., voice and appearance) might optimize the effectiveness of the virtual instructor.

Methods: Our study set four conditions. In the control condition, students watched a human instructor. In the experiment conditions, students watched one of (a) a virtual instructor which used the human instructor's voice and an AI image, (b) a virtual instructor which spoke in an AI voice with an AI image made to speak using text-to-speech and lip synthesis techniques, or (c) a virtual instructor with used an AI voice and an AI likable-image of an instructor.

Results and Conclusions: The AI likable instructor condition had a significant positive effect on students' learning performance and motivation, without decreasing the attention students paid to the learning materials.

Implications: Our findings suggest that instructional video designers can make use of AI voices and AI images of likable humans as instructors to motivate students and enhance their learning performance.

KEYWORDS

a virtual instructor, appearance, video lectures, voice

1 | INTRODUCTION

Video lectures which present an instructor alongside the informational slides have becoming popular learning material. Presenting a virtual instructor (i.e., a pedagogical agent) in video lectures has been shown to reduce costs of video production, both financially and in terms of time (Edwards et al., 2019; Li, Kizilcec, et al., 2016). Advances in artificial intelligence (AI) have made it possible to create virtual instructors with a high quality of voice and appearance (Craig & Schroeder, 2017; Domagk, 2010). Research shows that a virtual instructor can achieve the same levels of facial and vocal expression as a human instructor (i.e., a human; Hsieh & Sato, 2021; Lawson et al., 2021).

1.1 | Possible benefits and costs of using a virtual instructor in video lectures

It is commonly assumed that the addition of a virtual instructor in a video lecture may provide social enrichment to motivate students to engage more in the cognitive process, and thus facilitates learning from video lectures (Horovitz & Mayer, 2021; Lawson & Mayer, 2021; Mayer, 2014; Mayer et al., 2003). A virtual instructor rendered using computer graphic software is a digital character with features of speech, gesture, movement, and human-like behaviour, all intended to facilitate the learning process (Domagk, 2010; Li, Kizilcec, et al., 2016; Li, Oksama, & Hyönä, 2016). However, this idea that the virtual instructor's ability to generally motivate or facilitate learning has been questioned. Numerous studies have failed to reveal any learning benefits for students learning from a virtual instructor on motivation (Alyahya, 2021; Domagk, 2010; Lin et al., 2013) or on learning performance (Choi & Clark, 2006; Shiban et al., 2015). A literature review has also drawn a discouraging picture concerning the overall advantages of a virtual instructor on learning (Heidig & Clarebout, 2011). It found that the majority of the reviewed experiments (39 in total) yielded non-significant results on learning performance when comparing students learning from a virtual instructor to those learning from no-virtual instructor.

One possible reason for this might be that an instructor is a salient but unnecessary part of the lecture (Sweller et al., 2011). It could be that the virtual instructor distracted students' attention from the learning materials. While adding a virtual instructor into video lectures might motivate students to learn, it may also increase the attention they give to irrelevant information (i.e., instructor characteristics). Based on the theoretical considerations mentioned here, then, as well as on findings from previous studies, the current study set out to explore the consequences of adding a virtual instructor in video lectures in terms of students' motivation, attention, and learning performance.

Existing research has focused mainly on the impact of the presence of a virtual instructor by comparing virtual instructor and no-virtual instructor groups and using self-reported scales after learning (Lin et al., 2013; Shiban et al., 2015). Little attention has been given to whether a virtual instructor is equally capable as a human instructor in facilitating the learning process (e.g., measuring students' attention given to the instructor and learning materials) or examining students' outcomes (e.g., motivation, retention, and transfer) from video lectures. According to the equivalence principle, as recently proposed by Horovitz and Mayer (2021), a virtual instructor can play a similar role as a human instructor in video lectures, for example, expressing emotions just as well as a human instructor (Horovitz & Mayer, 2021; Lawson et al., 2021). The present study was interested in doing a direct comparison of participants' learning experience from video lectures which used either a human or a virtual instructor teaching the same content.

1.2 | Voice and appearance effects of a virtual instructor

An instructor's characteristics have been shown to have the potential to impact students' motivation, attention, and thereby their learning performance (Chiou et al., 2020; Lawson & Mayer, 2021). Thus, a considerable amount of research has already been conducted into the importance of a virtual instructor's characteristics (Chiou et al., 2020; Lawson & Mayer, 2021). For example, previous studies have shown that a virtual instructor's vocal characteristics can affect students' motivation and interaction with the instructor (Edwards et al., 2019; Lawson & Mayer, 2021). Other earlier studies have found that students learn better and report better social rapport with an onscreen instructor when they use a human voice rather than a machine-generated one (Atkinson et al., 2005; Mayer et al., 2003; Mayer & DaPra, 2012). Interestingly, however, more recent studies have shown that modern AI can produce an appealing human-like voice which can facilitate learning performance even more than a human instructor's voice (Craig & Schroeder, 2017; Lawson & Mayer, 2021).

Some studies have shown that a virtual instructor's appearance (e.g., age, gender, race, clothing, realism, and likableness) also affects students' motivation and learning performance (Domagk, 2010; Johnson et al., 2013; Shiban et al., 2015). For example, a study conducted by Domagk (2010) compared the effects of the perceived appeal of a virtual instructor's appearance and voice on students' motivation and learning performance (i.e., retention and transfer). Both experiments found that a likable virtual instructor enhanced students' transfer, but with the advantage being only at a medium effect level ($\eta^2 = 0.06$).

Findings have also shown that a virtual instructor in video lectures with an appealing appearance attract students' attention more and promote the maintenance of that attention (Li, Oksama, & Hyönä, 2016; Liu & Chen, 2012; Maner et al., 2007; Sui & Liu, 2009). However, students' attention to the appealing appearance of an instructor might distract some attentional resources and trigger students' shift in focus from the learning materials to the instructor. Some researchers have proposed, however, that the increased motivation evoked by appealing appearances might have a moderating positive effect on learning via maintaining students' attention towards the video lectures overall (Domagk, 2010; Shiban et al., 2015).

Despite the many studies into virtual instructors, however, research has tended to focus mainly on either the role of a virtual instructor's voice or the instructor's appearance by comparing different virtual instructor groups (Mayer & DaPra, 2012; Shiban et al., 2015). The current study therefore takes a comprehensive view to combine both of these characteristics in order to determine how best to design a virtual instructor to optimize the effectiveness of the instructor in video lectures. Specifically, which virtual instructor characteristics might prime students' motivation and attention on learning materials, thus improving their learning performance?

1.3 | Using eye-tracking technology to understand the effects of voice and appearance of a virtual instructor on students' attention

Eye tracking technology provides a direct and objective way of recording eye movements in real time, allowing the visual attentional processes to be tracked, measured, and interpreted (Wang et al., 2019; Zhang et al., 2021). An instructor should attract students' attention while they watch video lectures (Pi & Hong, 2016; Zhang et al., 2021). Previous studies on video lectures have often employed eye tracking technology to capture students' point of attention or focus (van Wermeskerken et al., 2018; van Wermeskerken & van Gog, 2017). Pi and Hong (2016) used an eye-tracker to test whether an on-screen instructor attracted students' visual attention, and found that while students paid a great amount of attention to the instructor, they also switched focus frequently between the instructor and the slides. Following this, we determined that eye movement tracking technology could be an appropriate tool to use to capture students' visual attention as they viewed video lectures presenting various virtual instructor versions, different in voice and appearance.

1.4 | The present study

The current study examined whether a virtual instructor would be equally capable at facilitating learning from video lectures as a human instructor, and what virtual instructor characteristics (i.e., voice and appearance) might optimize the effectiveness of the instructor in the video lectures in terms of learning performance (i.e., retention and transfer), self-reported motivation after the video lectures, and attention while viewing the lectures. Participants were asked to watch video lectures about English vocabulary words. In the control condition, students watched a video lecture which used a human instructor. In the experiment conditions, students watched video lectures which used various versions of a virtual instructor: (1) with the human instructor's voice and AI-generated image made by the human instructor's photo using lip synthesis method, (2) with an AI-generated voice using modern text-to-speech method and the AI-generated image made by the human instructor's photo using lip synthesis method, and (3) with the AI-generated voice using modern text-to-speech method and an AI-generated likable-image to stand in as the instructor (e.g., a famous singer, a famous host) using lip synthesis method.

We considered the motivational benefits of the on-screen instructor (Mayer et al., 2003) and the equivalence principle (Horovitz & Mayer, 2021) to inform our hypotheses about learning from a virtual instructor. Furthermore, we used empirical evidence regarding voice and appearance effects of virtual instructors (Craig & Schroeder, 2017; Domagk, 2010; Lawson & Mayer, 2021) to inform our hypotheses about learning from a virtual instructor using an AI-generated voice and an AI-generated likable-instructor. We expected that a virtual instructor would have the same impact as the human instructor on students' learning performance and motivation, but would not draw the same level of students' attention. In addition, we hypothesized that learning from a virtual instructor with an AI-generated voice or an AI-generated likable-image of the instructor would facilitate students' learning performance, motivation, and attention relative to the human instructor condition, and that learning from a video lecture using an AI-generated likable-image of the instructor accompanied by the AI-generated voice would be the most effective form of video lecture. More specifically, we posed the following hypotheses:

Hypothesis 1. *Students will show the best learning performance as indicated by retention and transfer after watching the video lecture using the AI-generated voice and the AI-generated likable-image of the instructor, followed by the video lecture that used the AI-generated voice with the AI-generated image of the human instructor visually appearing to speak, then the video lecture that used the human instructor's voice and the AI-generated image of the human instructor, and finally, the video lecture that used a recording of the human instructor.*

Hypothesis 2. *Students will report the greatest level of motivation after watching the video lecture using the AI-generated voice and the AI-generated likable-image of the instructor, followed by the video lecture using the AI-generated voice and the AI-generated image of the human instructor, then the video lecture using the human instructor's voice and the AI-generated image of the human instructor, and finally the video lecture using the human instructor.*

Hypothesis 3. *Students will pay the greatest attention to the instructor when watching the video lecture using the AI-generated voice and the AI-generated likable-image of the instructor, followed by the video lecture using the human instructor, then the video lecture using the human instructor's voice and the AI-generated image of the human instructor, and finally the video lecture using the AI-generated voice and the AI-generated image of the human instructor.*

Hypothesis 4. *Students will pay the least attention to the learning materials when watching the video lecture using the human instructor, followed by the video lecture using the AI-generated likable-image of the instructor and the AI-generated voice, then the video lecture using the human instructor's voice and the AI-generated image of the human instructor, and finally the video lecture using the AI-generated voice and the AI-generated image of the human instructor.*

Hypothesis 5. *Students will switch their attention more between the instructor and the learning materials when watching the video lecture using the AI-generated voice and the AI-generated likable-image of the instructor, followed by the video lecture using the human instructor, then the video lecture using the human instructor's voice and the AI-generated image of the human instructor, and finally the video lecture using the AI-generated voice and the AI-generated image of the human instructor.*

2 | METHOD

2.1 | Participants

We randomly recruited 36 undergraduate and graduate students (17 males and 19 females) from a university in China, aged from 20 to 27 years ($M = 23.94$, $SD = 1.29$). The students were majoring in psychology, educational technology, educational economy and management, religion, ethnic education, curriculum and pedagogy, neurobiology, electronic information, transportation engineering, or mechanical and electronic engineering. There were 26 participants who had passed the National College Computer Level two/three/four Examination in China. All participants had taken courses related to computers for more than 1 year (e.g., courses focusing on C programming language, MATLAB, or data structure). Thirty-three of the participants had also taken courses related to education (e.g., educational psychology, educational principle, or educational research methods). All participants reported that they had their own personal computer and used computers more than 3 h per day. There were 13 participants without formal online study experience, eight participants with only 3 months' formal online study experience, eight participants with only one term of formal online study experience, and seven participants with more than 1 year of formal online study experience. According to their self-reports, all participants had normal or corrected-to-normal vision and hearing when taking part in the current experiment. All participants signed written informed consent and were compensated for their participation.

2.2 | Experimental design

To control additional variables (e.g., learning situation, noise, and multitasking), the study applied a within-subjects experimental design. Each participant had to watch four video lectures which included all the different types of instructor's voice and appearance. Participants watched the four video lectures in a counterbalanced order to learn the 28 English vocabulary words (seven words for each video).

The differences in the instructor's voice and appearance for each of the four video lectures are shown in Table 1. (1) In the human instructor condition (HI), the participant viewed the video lecture using a human female instructor who pre-recorded the lecture using her dynamic image and voice. (2) In the virtual instructor with the human instructor's voice and the AI-generated image of the human instructor condition (HIV + AII), the participant viewed a video lecture which used an embedded a dynamic image of the human instructor whose movements were automatically synthesized to the speech using a motion engine developed by authors (Xu et al., 2021), using text-to-speech and lip synthesis methods. This engine generates videos by only importing text and the same human image as used in the HI condition. When the instructor talked, her mouth moved like a human's. (3) In the virtual instructor with the AI-generated voice and the AI-generated image of the human instructor condition (AIV + AII), the participant viewed a video lecture with the AI-generated voice and the same photo of the real instructor as used in the RI condition, also created using the virtual teacher engine. (4) In the virtual instructor with the AI-generated voice and likable-image condition (AIV + AILI), the participant viewed a video lecture with an AI-generated voice and a likable instructor photo (e.g., a famous singer, a famous host) made to appear to be speaking by using the virtual teacher engine. Participants were asked to provide the name of their chosen "likable" instructor before the day of the formal experiment, and the study team created the video lecture for the AIV + AILI condition using an image of that person in advance of the participant taking part in the formal experiment.

TABLE 1 The differences in the instructor's voice and appearance in each of the four video lecture conditions

Condition	Voice	Appearance
Real human instructor (HI)	Human voice	Human instructor
Virtual instructor using human instructor's voice and the AI-generated image (HIV + AII)	Human voice	Photo of human instructor
Virtual instructor using AI-generated voice and the AI-generated image (AIV + AII)	AI voice	Photo of human instructor
Virtual instructor using AI-generated voice and the AI-generated likable-image (AIV + AILI)	AI voice	Likable photo as instructor

2.3 | Video lectures

Four video lectures were created to teach a total of 28 English vocabulary words (seven words for each video lecture) taken from the Graduate Record Examination. Each video presented the instructor in a different format: (1) HI, (2) HIV + AII, (3) AIV + AII, and (4) AIV + AILI. The duration of the video lectures ranged from 6 min and 32 s to 6 min and 51 s. In each video, an instructor explained the English pronunciation of the word, its part in speech, its corresponding using both English and Chinese explanations, and an example sentence.

To ensure that the English vocabulary words in the four video lectures were not already known to participants, we invited a different 24 undergraduate and graduate students ($M_{age} = 21.92$, $SD_{age} = 1.86$; 21 females) from different study programs to watch each video individually and rate their level of familiarity with the chosen words. Participants rated all words from 1 ("extremely unfamiliar") to 7 ("extremely familiar"). The descriptive results showed that students were not familiar with the words used in the video lectures ($M = 1.78$, $SD = 1.25$). Furthermore, we did not find differences in the reported level of familiarity with the words, $F(3, 501) = 1.96$, $p = 0.119$, $\eta_p^2 = 0.01$, suggesting that viewers' familiarity of the words used across all four video lectures were the same.

2.4 | Measures

2.4.1 | Retention pretests

Four retention pretests were developed for each video lecture to measure participants' prior knowledge of the English vocabulary words. Each test included seven fill-in-the-blanks items. Students were required to write in the Chinese meaning of each word (1 point for a correct answer; otherwise, 0 points), with a total possible score of 7 for each test. Higher scores indicated a higher level of prior knowledge. The reliabilities of the tests were satisfactory: Cronbach's α was 0.88 for the four conditions overall, 0.78 for the HI condition, 0.69 for the HIV + AII condition, 0.76 for the AIV + AII condition, and 0.82 for the AIV + AILI condition.

2.4.2 | Learning performance posttests

Four retention posttests were used, and we developed four additional transfer posttests to measure students' acquisition of the English vocabulary words after viewing the video lectures. Retention posttests were the same as the retention pretests. Each transfer posttest included 21 multiple-choice items, made up of three types of items. The first seven items required students to choose the correct word from eight choices to complete a sentence. Only one choice was correct, while six of the seven incorrect choices comprised words that had also been taught in the video lectures. The final incorrect choice was "I do not know," to avoid participants simply guessing an answer. An example of this question is, "Many families were left by the horrible fire." The eight choices were: A. pique; B. sporadic; C. grandiloquent; D. opaque; E. succumb; F. appendage; G. destitute; H. I do not know. The remaining 14 items on the tests required students to choose a synonym or antonym of the learned word from five choices. Only one of these choices was correct, and the options included "I do not know" as a choice to avoid the participants guessing at an answer. Participants received 1 point if their answer was correct and 0 points if their answer was incorrect. Thus, the total possible score on the test was 21. Higher scores indicated better transfer. The reliabilities of the transfer posttests were satisfactory, with Cronbach's $\alpha = 0.93$ for the four conditions, 0.85 for the HI condition, 0.86 for the HIV + AII condition, 0.78 for the AIV + AII condition, and 0.78 for the AIV + AILI condition.

2.4.3 | Motivation scale

We used the Dimension of Motivation (six items) of the Learning Experience Scale as developed by Stull et al. (2018). The six items measured students' enjoyment, willingness to learn in this way in the future, understanding of the English vocabulary words in the video lectures, desire to learn more about the English vocabulary words in the video lectures, finding the video lecture useful, and motivation to learn the English vocabulary words in the video lectures. Participants rated all items from 1 ("strongly disagree") to 7 ("strongly agree"). The reliability of this scale was satisfactory, and Cronbach's α was 0.93 for the four conditions overall, 0.94 for the HI condition, 0.95 for the HIV + AII condition, 0.94 for the AIV + AII condition, and 0.90 for the AIV + AILI condition.

2.4.4 | Attention

We used an Eyelink 1000 system (SR Research Ltd., Canada) to record participants' eye movements in real-time to measure their attention. We created two areas of interest (AoIs): the instructor area and the learning materials area. Due to the slight differences in the video lecture durations, we used percentage dwell time and saccade counts to analyse participants' attention to the two AoIs, widely used to measure attention allocation (Wang et al., 2019; Zhang et al., 2021). Percentage dwell time refers to the percentage of time that the participant spent focused on an AoI and indicates how much attention participants pay to specific areas on the screen (Zhang et al., 2021). Saccade counts are the total number of times that a fixation transits from one AoI to the other (i.e., from the learning materials area to the instructor area, and vice versa) and indicates how participants shift their attention (Wang et al., 2019).

2.5 | Procedure

The study was conducted in an eye-tracking laboratory. Before starting the experiment, participants completed the demographic information (e.g., age, gender, and major) and the retention pretests. Then, they viewed four video lectures by counterbalance while their eye movements were recorded in real-time. Immediately after viewing each video lecture, participants filled out the learning performance tests and motivation scale. In total, the experiment took about 40 minutes to complete.

2.6 | Data analysis

To test our five hypotheses, a series of repeated measures analyses of variance (ANOVAs) was conducted using the video lectures (i.e., HI, HIV + AII, AIV + AII, AIV + AILI) as the within-subjects factor. The dependent variables included the scores of the retention pretest, the scores of the retention posttest and transfer posttest (H1), the scores of the motivation scale (H2), participants' percentage dwell time on the instructor (H3) and the learning materials (H4), and the saccade counts between the instructor and the learning materials (H5).

3 | FINDINGS

Preliminary analyses of all variables are presented in Table 2. The descriptive results showed that participants had low prior knowledge

TABLE 2 Means and standard deviations of all variables across the four conditions

	HI	HIV + AII	AIV + AII	AIV + AILI
Retention pretest	0.06 (0.23)	0.14 (0.35)	0.11 (0.32)	0.08 (0.28)
Learning performance				
Retention posttest	3.75 (2.21)	4.31 (1.98)	4.53 (2.14)	5.42 (1.99)
Transfer posttest	11.19 (4.92)	11.94 (5.12)	12.92 (4.16)	14.92 (3.97)
Motivation	3.22 (1.37)	2.60 (1.23)	2.22 (1.31)	3.11 (1.47)
Attention				
Percentage dwell time on instructor	15.58 (10.59)	9.96 (8.66)	9.19 (6.68)	12.16 (8.45)
Percentage dwell time on learning materials	83.50 (11.04)	88.61 (9.84)	88.95 (8.01)	87.11 (8.81)
Saccade counts	33.58 (22.56)	26.75 (20.98)	26.17 (20.74)	31.61 (24.33)

Note: The values outside parentheses are means. The values inside parentheses are standard deviations. Percentage dwell time are in decimal form. HI, HIV + AII, AIV + AII, AIV + AILI, respectively, represent the human instructor condition, the human instructor's voice and AI-generated image condition, the AI-generated voice and AI-generated image, and the AI-generated voice and AI-generated likable-image condition.

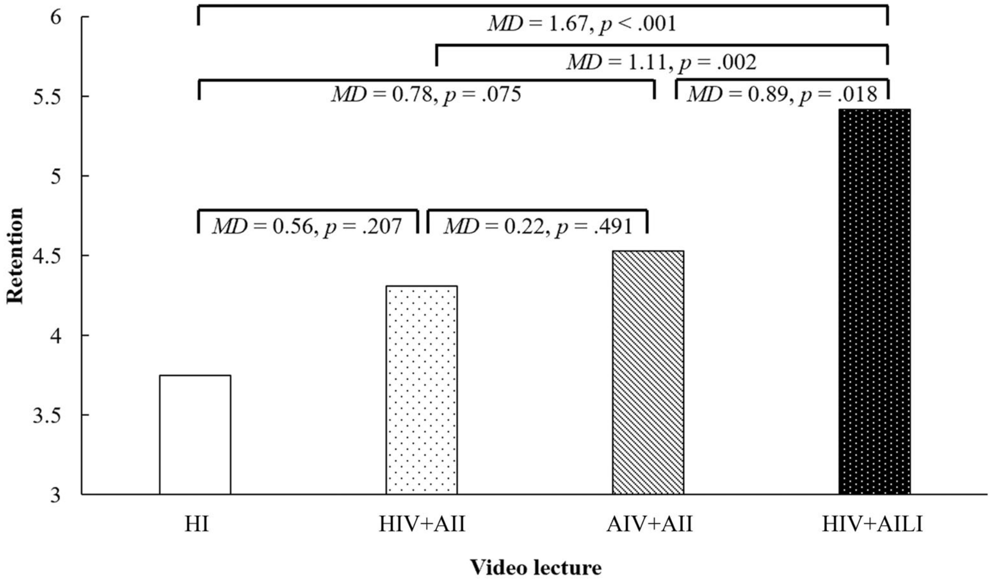

FIGURE 1 Differences in retention posttest across the four video lectures

of the learned words, and the results showed no significant difference in prior knowledge across the four video lectures, $F(3,\ 105) = 0.62$, $p = 0.605$, $\eta_p{}^2 = 0.02$.

3.1 | Learning performance

We hypothesized that the video using an AI-generated voice and an AI-generated likable-image would best facilitate participants' learning performance. Concerning retention, we observed significant differences across the four video lectures: $F(3,\ 105) = 6.88$, $p < 0.001$, $\eta_p{}^2 = 0.16$. Post hoc tests (*LSD*) found that participants showed better retention in the AIV + AILI condition than in the HI, HIV + AII, and AIV + AII conditions, but there was no significant difference in retention across these other conditions (see Figure 1).

Concerning transfer, we observed significant differences across the four video lectures: $F(3,\ 105) = 9.23$, $p < 0.001$, $\eta_p{}^2 = 0.21$. Post hoc test results showed that participants showed better transfer in the AIV + AILI condition than in the HI, HIV + AII, and AIV + AII conditions. Furthermore, participants showed better transfer in the AIV + AII condition than in the HI condition. There was no significant difference in retention between the HI and HIV + AII conditions, or between the HIV + AII and the AIV + AII conditions (see Figure 2).

Taken together, these results partially supported our first hypothesis. Students showed better retention and transfer in the AIV + AILI condition than they did in the other three conditions. Students showed worse transfer in the HI condition than they in the AIV + AII condition. Contrary to our hypothesis, however, students did not show any differences in retention and transfer between the HIV + AII condition and AIV + AII condition.

3.2 | Motivation

We hypothesized that an AI-generated voice and an AI-generated likable-image would enhance participants' motivation. We observed significant differences in motivation across the four video lecture conditions: $F(3,\ 105) = 12.98$, $p < 0.001$, $\eta_p{}^2 = 0.27$. Post hoc test results showed that participants showed the greatest motivation in the HI

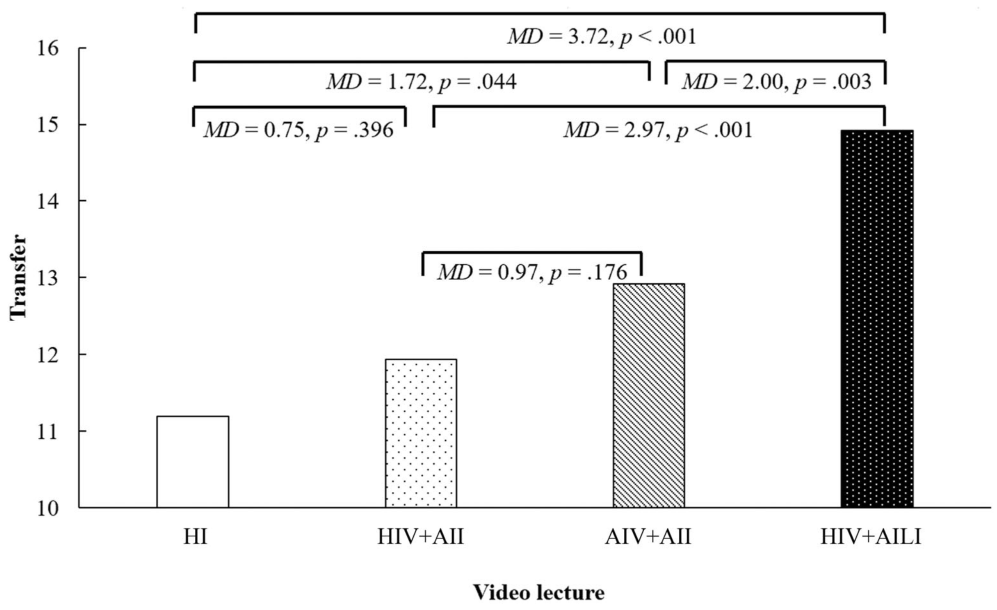

FIGURE 2 Differences in transfer posttest across the four video lectures

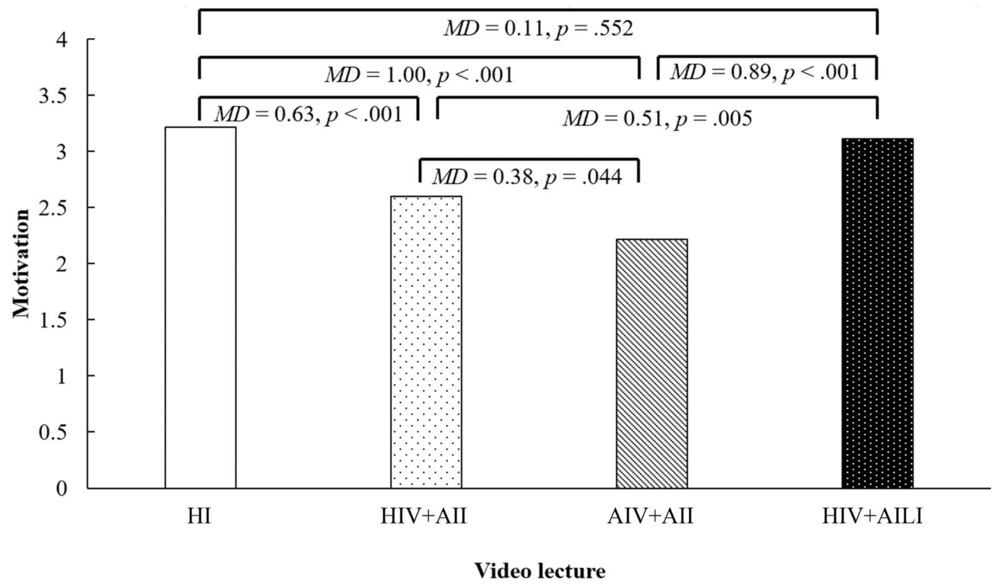

FIGURE 3 Differences in motivation across the four video lectures

and AIV + AILI conditions, followed by the HIV + AII condition, and then the AIV + AII condition. There were no significant differences in motivation noted between the HI and AIV + AILI conditions (see Figure 3). These results partially supported our second hypothesis, that the AIV + AILI condition would enhance students' motivation, although the HI condition showed the same benefit.

3.3 | Attention

3.3.1 | Percentage dwell time on the instructor

We hypothesized that an AI voice and an AI-generated likable-image would enhance participants' attention to the video lecture, resulting in longer measured dwell time on the instructor AoI. We found significant differences across the four video lecture conditions: $F(3, 105) = 10.74$, $p < 0.001$, $\eta_p^2 = 0.24$. Post hoc test results showed that participants had higher dwell time on the instructor in the HI condition than in the HIV + AII, AIV + AII, and AIV + AILI conditions. Furthermore, participants spent more dwell time on the instructor in the AIV + AILI condition than they did in the AIV + AII condition. There was no significant difference in percentage dwell time on the instructor across the other conditions (see Figure 4). These results partially supported our third hypothesis, that participants would spend a longer dwell time on the instructor in the HI condition, followed by the AIV + AILI condition.

3.3.2 | Percentage dwell time on the learning materials

We hypothesized that the use of a human instructor in the video lecture would decrease participants' attention on the learning materials, resulting in shorter dwell time spent on the learning materials. We

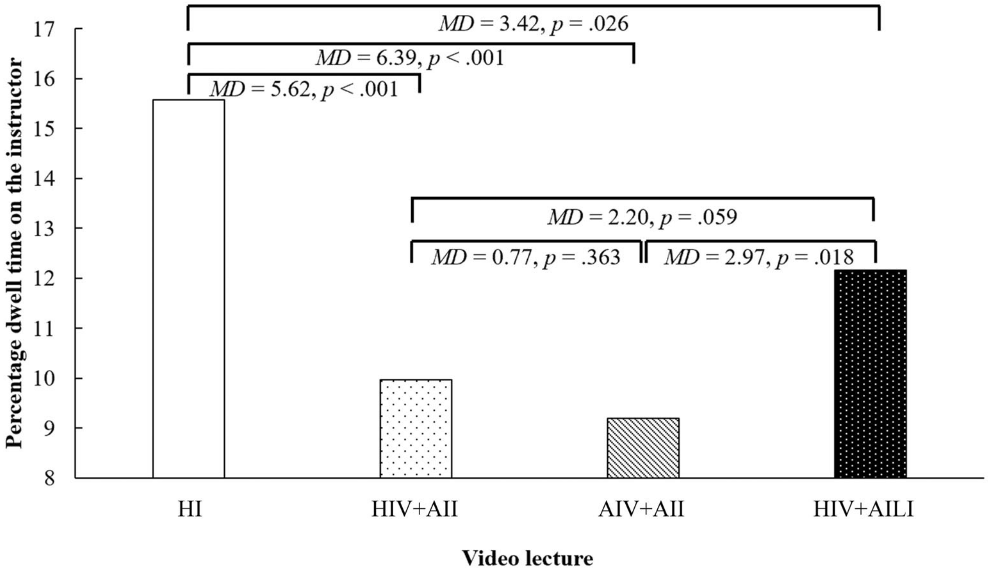

FIGURE 4 Differences in percentage dwell time on the instructor across the four video lectures

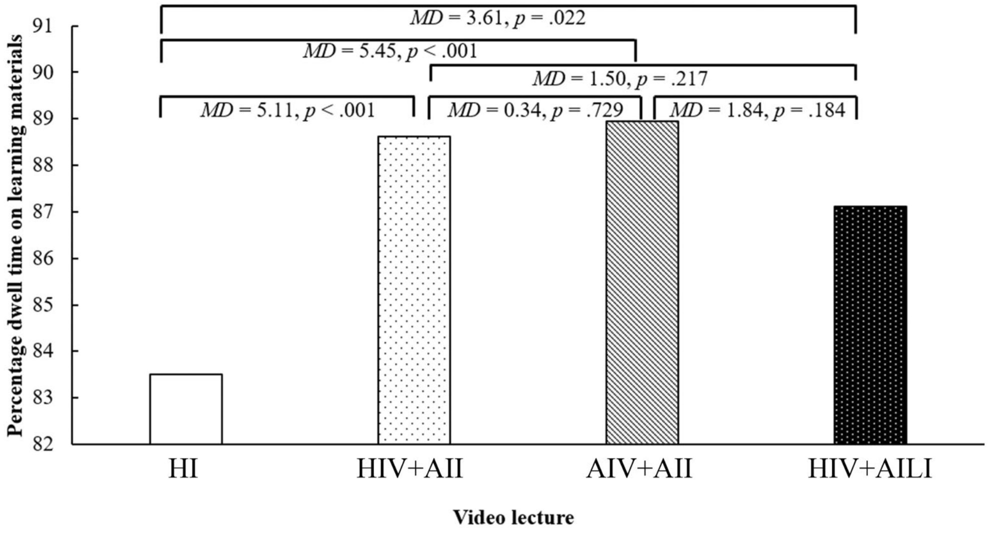

FIGURE 5 Differences in percentage dwell time on learning materials across the four video lectures

found significant differences across the four video lectures: $F(3, 105) = 7.29$, $p < 0.001$, $\eta_p^2 = 0.17$. Post hoc test results showed that participants had less dwell time on the learning materials in the HI condition than in the HIV + AII, AIV + AII, and AIV + AILI conditions. There was no significant difference in percentage dwell time on the learning materials across other conditions (see Figure 5). These results partially supported our fourth hypothesis, that students had shorter dwell time on the learning materials in the HI condition compared to the other three video lectures conditions, although students did not show any differences in dwell time between the AIV + AILI condition and the conditions of HIV + AII and AIV + AII.

3.3.3 | Saccade counts

We hypothesized that an AI-generated voice and an AI-generated likable-image would enhance participants' saccade counts between the learning materials and the instructor areas. We observed significant differences across the four video lectures: $F(3, 105) = 4.60$, $p = 0.005$, $\eta_p^2 = 0.12$. Post hoc tests found that participants had more saccade counts in the HI and AIV + AILI conditions than in the HIV + AII and AIV + AII conditions. There was no significant difference in saccade counts between the learning materials and instructor areas across other conditions (see Figure 6). These results partially supported our hypothesis, that the HI and AIV + AILI conditions both caused participants to shift their attention more between the two areas, more than in the HIV + AII and AIV + AII conditions, although the AIV + AILI condition did not trigger more saccade counts than the HI condition.

4 | DISCUSSION

4.1 | Empirical contributions

This study tested whether a virtual instructor facilitated learning from video lectures as well as a human instructor, and whether voice and

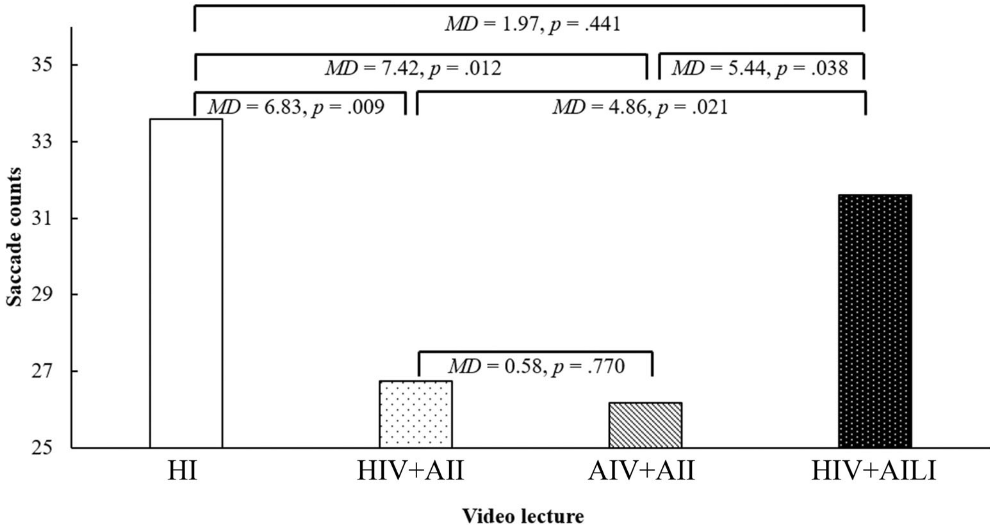

FIGURE 6 Differences in saccade counts between instructor and learning materials areas across the four video lectures

appearance characteristics of the virtual instructor influenced participants' learning performance, motivation, and attention. We found that the use of a likable instructor image had a significant positive effect on participants' learning performance and motivation, without triggering an attention split between the instructor and the learning materials areas. This suggests that a likable instructor image facilitates learning from video lectures more than those that use a human instructor or a virtual instructor which uses the human instructor's photo. This study contributes to the existing research that shows the powerful contribution of using a virtual instructor in the form of an AI-generated likable-image accompanied by an AI-generated voice in video lectures.

Consistent with the equivalence principle (Horovitz & Mayer, 2021), our findings suggest that a virtual instructor and a real instructor can both facilitate learning from video lectures equally, as we found that participants showed the same level of retention whether viewing video lectures using a virtual instructor or with a human instructor. They showed even better transfer when viewing the video lecture that used a virtual instructor with an AI-generated voice with an AI-generated image of the human instructor. These results were consistent with previous findings on AI-generated voices, confirming that modern AI can produce an appealing human-like voice that can facilitate learning performance better than a human instructor's voice (Craig & Schroeder, 2017; Lawson & Mayer, 2021). This suggests that an appealing voice created using modern AI could then facilitate students' transfer performance better than using a human instructor's voice in video lectures. However, we did not include a condition that replaced only the human instructor's voice with an AI-generated voice. Future work should compare a video lecture using a human instructor with a video lecture using a human instructor accompanied by an AI-generated voice to determine whether an appealing voice made using modern AI can in fact better facilitate learning than a human instructor's voice.

As predicted, participants showed better retention and transfer when viewing the video lectures that used a likable instructor image. These results are consistent with previous findings that have shown that a virtual instructor with an appealing appearance and voice facilitate learning better than one with unappealing appearance and voice (Domagk, 2010; Johnson et al., 2013). Interestingly, participants reported a higher level of motivation when viewing both the video lectures with the likable instructor image and the human instructor. However, the motivational benefits of the human instructor might be offset by increased distraction. The eye movement data showed that participants paid greater attention to the instructor when viewing the video that used the human instructor, and therefore less attention to the learning materials. Although participants also paid more attention to the instructor and shifted more between the instructor and learning materials areas when viewing the video with the likable instructor image, they did not appear to pay less attention to the learning materials themselves. Therefore, designing a virtual instructor using a likable image might cater to both motivational benefits and cognitive design principle (Mayer et al., 2003; Sweller et al., 2011). This could explain why the motivational benefits of the human instructor do not contribute to increased learning performance.

4.2 | Theoretical contributions

The current study advances our understanding of the various effects of a virtual instructor using different voice and appearance characteristics in video lectures. Previous studies on virtual instructors have predominantly compared a virtual instructor group with a no-virtual instructor group (Lin et al., 2013; Shiban et al., 2015). In contrast, the study compared the effectiveness of three virtual instructor conditions with a human instructor condition to test whether a virtual instructor can be equally as effective as a human instructor. Our findings confirm the benefits of learning from video lectures using a virtual instructor, with findings showing in both self-report tests or scales as well as from tracking real-time eye movements. This study therefore contributes to broadening our understanding of the impact of voice and appearance characteristics of virtual instructors, and how to optimize them to improve the effectiveness of video lectures.

To our knowledge, this is the first study to test the effects of a virtual instructor on students' motivation and cognition (i.e., how viewers pay attention to the instructor and learning materials, as well as learning performance). Previous studies have noted that when video lectures use an on-screen instructor, students are more motivated to engage in the cognitive process (i.e., selection of incoming information, organization, and integration of the information with the prior knowledge; Mayer, 2014; Mayer et al., 2003). We found that using a virtual instructor had a positive influence on participants' motivation, attentional engagement, and learning performance (i.e., retention and transfer). While a virtual instructor using an AI-generated voice and a likable human's image can lead to increased motivation and greater attentional engagement with the instructor, it does not lead to attentional disengagement in the learning materials. One significance of our findings comes from our recording participants' eye movements in real time via an eye tracking technology, which provides direct and objective evidence on viewers' attention engagement. Although previous studies have suggested that the characteristics of a virtual instructor can influence students' attention (Maner et al., 2007; Sui & Liu, 2009), this study is first to analyse their visual attentional processes directly.

4.3 | Limitations and future directions

There are two limitations to the current study that must be acknowledged. First, it should be noted that the participants' viewing of the video lectures was system-paced rather than self-paced. There are slight differences in the behaviour sequences of these two contexts when viewing video lectures. In the real learning context of video lectures, students would be able to view them at their own pace, with the ability to pause, fast forward, and to go backwards in the video (Pi et al., 2020). Students' attention to an instructor in video lectures may also differ over time (e.g., differences between attention on the first viewing versus the second viewing of a video lecture), although a recent study did not notice any differences in viewing when self- or system-paced (van Wermeskerken et al., 2018). Future research is nonetheless needed to determine the influence of a virtual instructor when viewing is system-paced versus self-paced.

The second limitation in the current study concerns the analysis of eye movement behaviour. We analysed the attention participants paid to the instructor and learning materials areas by percentage dwell time and saccade counts. Percentage dwell time indicates how long viewers paid attention to specific screen areas of the video lectures (van Wermeskerken et al., 2018). Saccade counts indicate viewers' shifting of attention between the different areas (van Wermeskerken & van Gog, 2017; Zhang et al., 2021). These results do not reveal, however, whether the viewers are trying to understand the learning content or whether they simply find the areas interesting when they focus on the areas. Further research should investigate viewers' motivations to look at the various areas through interviews.

4.4 | Practical implications

With the ongoing development of digital technologies as well as schools moving more towards online teaching, and video lectures have become a major teaching format. It is therefore essential to design video lectures to be as effective as possible. Video lectures which use an instructor alongside the information slides are a particularly popular format. However, practically, it is easier and costs less for instructional designers to create video lectures using a virtual instructor rather than having a human instructor present on the screen.

Our findings provide implications for improved video lecture design. First, this study demonstrated that students showed better transfer after viewing video lectures using an AI-generated voice and an AI-generated image made by the human instructor's photo than after those using a recorded version of the human instructor giving the lecture. We therefore encourage instructional designers to use AI-generated voices and AI-generated images connected to a virtual instructor instead of a human instructor in video lectures. Not only will this reduce both financial and time costs but our findings show that a virtual instructor with an AI-generated voice does not trigger shifts in students' attention between the instructor and learning materials, while furthermore improving students' learning performance. Furthermore, the current study identified the benefits of using a virtual instructor made up of an AI-generated voice and a likable human's image in video lectures. We therefore encourage instructional designers to use likable human images (as defined by the students themselves) as a way to motivate students more and enhance their learning performance.

AUTHOR CONTRIBUTIONS

Zhongling Pi, Tao Xu and Yun Zhou contributed to the study conception and design. Material preparation, data collection and analysis were performed by Lixia Deng, Xu Wang and Peirong Guo. The first draft of the manuscript was written by Zhongling Pi. All authors read and approved the final manuscript.

FUNDING INFORMATION

This work was supported by the Research Projects of the National Natural Science Foundation of China under Grant (62007023; 62077036), the Research Program Funds of the Collaborative Innovation Center of Assessment towards Basic Education Quality at Beijing Normal University (2021-05-029-BZPK01), and the National Key Research and Development Program of China (2018AAA0100501).

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12704.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are available from the corresponding author upon request.

ORCID

Zhongling Pi https://orcid.org/0000-0002-8776-6177
Lixia Deng https://orcid.org/0000-0002-9028-9439
Xu Wang https://orcid.org/0000-0003-0510-5242
Peirong Guo https://orcid.org/0000-0003-3434-6037
Tao Xu https://orcid.org/0000-0002-1721-561X
Yun Zhou https://orcid.org/0000-0002-2306-8986

REFERENCES

Alyahya, S. M. (2021). Social cues in animated pedagogical agents for second language learners: The application of the embodiment principle in video design. Dissertations Publishing, University of South Florida.

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, *30*(1), 117–139.

Chiou, E. K., Schroeder, N. L., & Craig, S. D. (2020). How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education*, *146*, 103756.

Choi, S., & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research*, *34*(4), 441–466.

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers & Education*, *114*, 193–205.

Domagk, S. (2010). Do pedagogical agents facilitate learner motivation and learning outcomes? *Journal of Media Psychology*, *22*(2), 82–95.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social identity theory in human-robot interactions. *Computers in Human Behavior*, *90*, 357–362.

Heidig, S., & Clarebout, G. (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27–54.

Horovitz, T., & Mayer, R. E. (2021). Learning with human and virtual instructors who display happy or bored emotions in video lectures. *Computers in Human Behavior*, *119*, 106724.

Hsieh, R., & Sato, H. (2021). Evaluation of avatar and voice transform in programming e-learning lectures. *Journal on Multimodal User Interfaces*, *15*, 121–129.

Johnson, A. M., Didonato, M. D., & Reisslein, M. (2013). Animated agents in k-12 engineering outreach: Preferred agent characteristics across age levels. *Computers in Human Behavior*, *29*(4), 1807–1815.

Lawson, A. P., & Mayer, R. E. (2021). The power of voice to convey emotion in multimedia instructional messages. *International Journal of Artificial Intelligence in Education*, *31*, 1–20. https://doi.org/10.1007/s40593-021-00282-y

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021). Recognizing the emotional state of human and virtual instructors. *Computers in Human Behavior*, *114*, 106554.

Li, J., Kizilcec, R., Bailenson, J., & Ju, W. (2016). Social robots and virtual agents as lecturers for video instruction. *Computers in Human Behavior*, *55*, 1222–1230.

Li, J., Oksama, L., & Hyönä, J. (2016). How facial attractiveness affects sustained attention? *Scandinavian Journal of Psychology*, *57*, 383–392.

Lin, L., Atkinson, R. K., Christopherson, R. M., Joseph, S. S., & Harrison, C. J. (2013). Animated agents and learning: Does the type of verbal feedback they provide matter? *Computers & Education*, *67*, 239–249.

Liu, C. H., & Chen, W. (2012). Beauty is better pursued: Effects of attractiveness in multiple-face tracking. *Quarterly Journal of Experimental Psychology*, *65*, 553–564.

Maner, J. K., Gailliot, M. T., & DeWall, C. N. (2007). Adaptive attentional attunement: Evidence for mating-related perceptual bias. *Evolution and Human Behavior*, *28*, 28–36.

Mayer, R. E. (2014). Principles based on social cues in multimedia learning: Personalization, voice, embodiment, and image principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 345–368). Cambridge University Press.

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied*, *18*(3), 239–252.

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*, 419–425.

Pi, Z., & Hong, J. (2016). Learning process and learning outcomes of video podcasts including the instructor and PPT slides: A Chinese case. *Innovations in Education and Teaching International*, *53*(2), 135–144.

Pi, Z., Tang, M., & Yang, J. (2020). Seeing others' messages on the screen during video lectures hinders transfer of learning. *Interactive Learning Environments*, *28*, 1–14. https://doi.org/10.1080/10494820.2020.1749671

Shiban, Y., Schelhorn, I., Jobst, V., Hörnlein, A., Puppe, F., Pauli, P., & Mühlberger, A. (2015). The appearance effect: Influences of virtual agent features on performance and motivation. *Computers in Human Behavior*, *49*, 5–11.

Stull, A. T., Fiorella, L., Gainer, M. J., & Mayer, R. E. (2018). Using transparent whiteboards to boost learning from online STEM lectures. *Computers & Education*, *120*, 146–159.

Sui, J., & Liu, C. H. (2009). Can beauty be ignored? Effects of facial attractiveness on covert attention. *Psychonomic Bulletin & Review*, *16*, 276–281.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

van Wermeskerken, M., Ravensbergen, S., & van Gog, T. (2018). Effects of instructor presence in video modeling examples on attention and learning. *Computers in Human Behavior*, *89*, 430–438.

van Wermeskerken, M., & van Gog, T. (2017). Seeing the instructor's face and gaze in demonstration video examples affects attention allocation but not learning. *Computers & Education*, *113*, 98–107.

Wang, H., Pi, Z., & Hu, W. (2019). The instructor's gaze guidance in video lectures improves learning. *Journal of Computer Assisted Learning*, *35*, 42–50.

Xu, T., Wang, X., Wang, J., & Zhou, Y. (2021). From textbook to teacher: An adaptive intelligent tutoring system based on BCI. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 7621–7624). IEEE.

Zhang, Y., Xu, K., Pi, Z., & Yang, J. (2021). Instructor's position affects learning from video lectures in Chinese context: An eye-tracking study. *Behaviour & Information Technology*, *40*, 1–10. https://doi.org/10.1080/0144929X.2021.1910731

How to cite this article: Pi, Z., Deng, L., Wang, X., Guo, P., Xu, T., & Zhou, Y. (2022). The influences of a virtual instructor's voice and appearance on learning from video lectures. *Journal of Computer Assisted Learning*, 1–11. https://doi.org/10.1111/jcal.12704