International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-021-00282-y

ARTICLE

The Power of Voice to Convey Emotion in Multimedia Instructional Messages

Alyssa P. Lawson[1] · **Richard E. Mayer**[1]

Accepted: 5 September 2021

Abstract

This study examines an aspect of the role of emotion in multimedia learning, i.e., whether participants can recognize the instructor's positive or negative emotion based on hearing short clips involving only the instructor's voice just as well as also seeing an embodied onscreen agent. Participants viewed 16 short video clips from a statistics lecture in which an animated instructor, conveying a happy, content, frustrated, or bored emotion, stands next to a slide as she lectures (agent present) or uses only her voice (agent absent). For each clip, participants rated the instructor on five-point scales for how happy, content, frustrated, and bored the instructor seemed. First, for happy, content, and bored instructors, participants were just as accurate in rating emotional tone based on voice only as with voice plus onscreen agent. This supports the *voice hypothesis*, which posits that voice is a powerful source of social-emotional information. Second, participants rated happy and content instructors higher on happy and content scales and rated frustrated and bored instructors higher on frustrated and bored scales. This supports the *positivity hypothesis*, which posits that people are particularly sensitive to the positive or negative tone of multimedia instructional messages.

Keywords Animated pedagogical agent · Emotion · Instructional video · Multimedia learning · Voice

Introduction

Objective and Rationale

Consider an instructional video such as exemplified in Fig. 1 in which an instructor (e.g., in this case, an animated pedagogical agent) stands next to a slide as

✉ Richard E. Mayer
mayer@psych.ucsb.edu

[1] Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106, USA

Published online: 01 October 2021

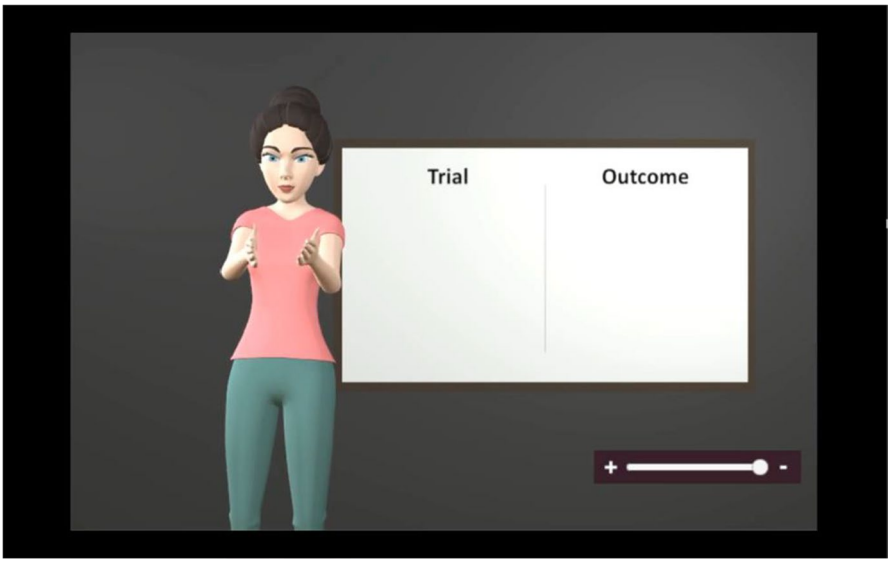

Fig. 1 Image of animated instructor presenting lecture on binomial probability

she lectures. This is an example of an instructional multimedia message (Mayer, 2020) because the video contains both words (i.e., spoken by the instructor and printed on the slides) and graphics (i.e., embodied by the image of the instructor and rendered on the slides), which are intended to foster learning. According to the cognitive theory of multimedia learning (Mayer, 2014, 2020), this may appear to be a purely cognitive event, in which the learner engages in cognitive processes such as attending to the relevant words and images (i.e., selecting), mentally organizing them into coherent verbal and visual representations (i.e., organizing), and relating the representations with each other and with relevant prior knowledge activated from long-term memory (i.e., integrating).

However, in the present study, we explore the idea that online multimedia learning may involve affective processing as well as cognitive processing, consistent with research on affective processes during learning with technology including animated pedagogical agents (Baylor et al., 2003; D'Mello, 2013; D'Mello & Graesser, 2012; Johnson & Lester, 2016; Johnson et al., 2000; Lester et al., 1997) within the larger field of affective computing (Calvo et al., 2015; Picard, 2000; Wu et al., 2016). More specifically, we examine whether adding the presence of an onscreen instructor (including eye gaze, gestures, facial expression, and body stance) had a stronger impact on how well learners recognized the emotional tone of an instructor compared to only having auditory cues (i.e., the emotional tone of the voice). The impact of this work is intended to inform researchers and educators on how an instructor's dynamic image (including gesture, eye gaze, facial expression, and body stance) and voice convey affective cues that the learner recognizes and reacts to during learning, thereby priming affective processing during learning (Mayer, 2020).

In particular, we test the *voice hypothesis*, which posits that the instructor's voice conveys affective cues (for affective processing during learning) in addition to verbal content (for cognitive processing during learning) (Nass & Brave, 2005). As the first step in investigating the voice hypothesis, we examine whether people can recognize the emotional tone of an instructor (e.g., happy, content, frustrated, or bored) in short video clips that contain only the instructor's voice along with informational slides just as well as in video clips that contain the instructor's image (including gesture, eye gaze, facial expression, and body movement) and voice along with slides. If people are just as accurate in recognizing the emotional tone of the instructor solely from voice as from voice and embodied image, this would be evidence for the power of voice to convey emotion in multimedia messages and support for the voice hypothesis.

In conjunction, we also test the *positivity hypothesis*, which posits that people can recognize whether an instructor is displaying positive or negative emotion (Horovitz & Mayer, 2021; Lawson et al., 2021a, 2021b). If people give higher ratings on the happy and content scales for happy and content instructors, and higher ratings on the frustrated and bored ratings for frustrated and bored instructors, this would be evidence for the idea that people are sensitive to the emotional tone of instructors, particularly whether the emotional tone is positive or negative.

Understanding the role of disembodied voice in conveying emotion in instructional video has theoretical and practical implications. On the theoretical side, if students can recognize the emotional tone of instructors simply from a disembodied voice, this provides evidence that the first step in a model of cognitive-affective learning can be achieved without the need for an onscreen agent. On the practical side, it might be easier and most cost-efficient for instructional designers to create instructional videos with voice over rather than with an instructor present on the screen. Additionally, many instructional videos online have already been employing the use of the disembodied voice, such as Khan Academy. Overall, the primary innovative contribution of the present study is to determine whether voice cues are sufficient to convey the emotional tone of the instructor, in the absence of embodiment cues conveyed by having a gesturing agent visually represented on the screen.

Literature Review

The present study is situated within the larger field of affective agents, which includes long-standing efforts to create onscreen agents that are perceived by users as having a distinct personality or persona (Cassels et al., 2000; Johnson & Rickel, 1997; Lester et al., 1997; Schroeder & Craig, 2021). Within this field, a more specific subfield concerns affective agents in education, which includes the long-standing study of how features of a pedagogical agent affect how learners perceive the agent's persona, respond to it, and ultimately, how the agent affects their learning outcomes (Baylor et al., 2003; Clarebout et al., 2002; Craig & Schroeder, 2018; Craig et al., 2002; Heidig & Clarebout, 2011; Schroeder et al., 2013). Finally, within the field of affective agents in education, an even more specific subfield concerns how the agent's voice affects the learner's perceptions

of the agent and their learning outcome (e.g., Horovitz & Mayer, 2021; Lawson et al., 2021a, 2021b; Ryu & Ke, 2018). In this study, we focus on the role of the agent's voice and conversational style in online multimedia instructional presentations, as preliminary step towards understanding the role of voice in interactive conversational agents who can engage in a natural language conversation with learner.

Previous research has demonstrated the role of the onscreen agent's voice and embodiment in promoting rapport and learning, consistent with idea that affective virtual agents can convey emotion that affects learning outcomes (Fiorella, 2022; Fiorella & Mayer, 2022; Mayer, 2021). Similarly, in a meta-analysis of affect in embodied pedagogical agents, Guo and Goh (2015) reported a significant impact of virtual agent affect on knowledge retention and transfer. Furthermore, preliminary work has provided encouraging evidence for positivity principle, which states that students learn more from an onscreen agent who displays positive emotional cues through voice, gesture, eye gaze, facial expression, and body stance (Horovitz & Mayer, 2021; Lawson et al., 2021a). In a recent study, Horovitz and Mayer (2021) asked college students to view a video lecture on a statistical concept displaying a happy human instructor, a bored human instructor, a happy virtual instructor, or a bored virtual instructor. For both human and virtual instructors, students who received the positive instructor rated the emotional state of positive instructor as more positive, rated their own emotional state as more positive, rated their motivation to learn as stronger, but did not perform better on a transfer posttest. Lawson et al. (2021a) reported similar findings in a study involving only animated instructors who displayed positive (i.e., happy or content) or negative (i.e., frustrated or bored) emotion. This work is consistent with related work showing that students were able to perceive the emotional stance of an onscreen human or virtual instructor after watching a short video clip from a statistics lesson, particularly whether the instructor was displaying positive (i.e., happy or content) or negative (i.e., frustrated or bored) emotion (Lawson et al., 2021b). Overall, this preliminary work shows that providing both voice cues and embodiment cues can help learners perceive and react to the emotional tone of onscreen agents. The present study extends this work by examining whether voice cues are sufficient to create the same effects in conveying emotion to learners.

Previous research has demonstrated the role of an intelligent tutor's conversational style in building rapport with learners, in which the social connection is intended to improve student learning. For example, Finkelstein et al. (2013) reported that elementary school students who were native speakers of African American Vernacular English (AAVE) learned better from an online peer who used features of in AAVE throughout the session. In a parallel line of research, Makransky et al. (2019) found a gender matching effect for high school students learning from a tutor in immersive virtual reality, in which girls learned better when the tutor was a young woman in white lab coat who spoke with a friendly female voice whereas boys learned better from a superhero drone who spoke in a tough male voice. In a related study, Ogan et al. (2012) found that learning in a high school peer tutoring environment was related to the conversational style used by the participants, suggesting that

intelligent tutors should shift gradually from direct style to more playful face-threat style over time.

Parallel results have been obtained in determining the conditions under which polite wording makes an intelligent more effective. For example, high school students learned better from a polite intelligent tutor only if they made the many errors during the intervention, indicating that polite wording may be important for the most needy students (McLaren et al., 2011). Similarly, college students learned better from a polite intelligent tutor only if they were highly inexperienced with the content of the lesson, again indicating that polite wording may be important for struggling students (Wang et al., 2008). These studies help confirm the importance of the tutor's conversational style in promoting learning, which complements research on the emotional tone of the tutor's voice.

In their classic book, *Wired for Speech*, Nass and Brave (2005) offer empirical evidence and logical argument for the proposal that voice conveys social and affective information in technology-mediated communications. For example, Nass et al. (2001) found that people preferred a happy voice to a sad voice. Additionally, voice affects the emotional impact of a narrated story, in which happy voices make happy stories seem happier and sad voices make sad stories seem sadder (Nass et al., 2001). Brave et al. (2005) reported that onscreen agents who exhibited empathic emotions were rated as more likeable and trustworthy. Edwards and Kortum (2012) reported that voice characteristics affected ratings of the perceived usability of an interactive voice response system. Edwards et al. (2019) reported that voice characteristics affected ratings of credibility, social presence, and motivation to learn. Baylor et al. (2003) found that people rated an onscreen agent as more engaging when it had a human voice rather than a machine-generated voice. Mayer and DaPra (2012) and Mayer et al. (2003) reported that students learned better and reported better social rapport when an onscreen agent used an appealing human voice rather than a machine-generated voice. Liew et al. (2020) reported that learners reported higher ratings of social rapport and performed better on transfer posttests when the instructor had an enthusiastic voice rather than a bland voice.

Overall, in reviews of research, Mayer (2014, 2020) concluded that students feel a better social-emotional connection with the instructor and learn better from lessons in which the instructor uses an appealing human voice rather than a dull machine voice, although there is evidence that modern text-to-speech engines can produce appealing human-like voice (Craig & Schroeder, 2017). Thus, the research literature highlights examples in which affective information is carried by the voice of an instructor. We conclude that there is justification to further explore the role of voice for conveying emotion in multimedia instructional messages.

Theory and Predictions

As shown in Fig. 2, Russell's (1980, 2003) model of core affect represents human emotion along two orthogonal dimensions: a positive–negative dimension and an active–passive dimension. In the present study, we explore animated instructors who are intended to convey the emotion corresponding to each of the four quadrants in

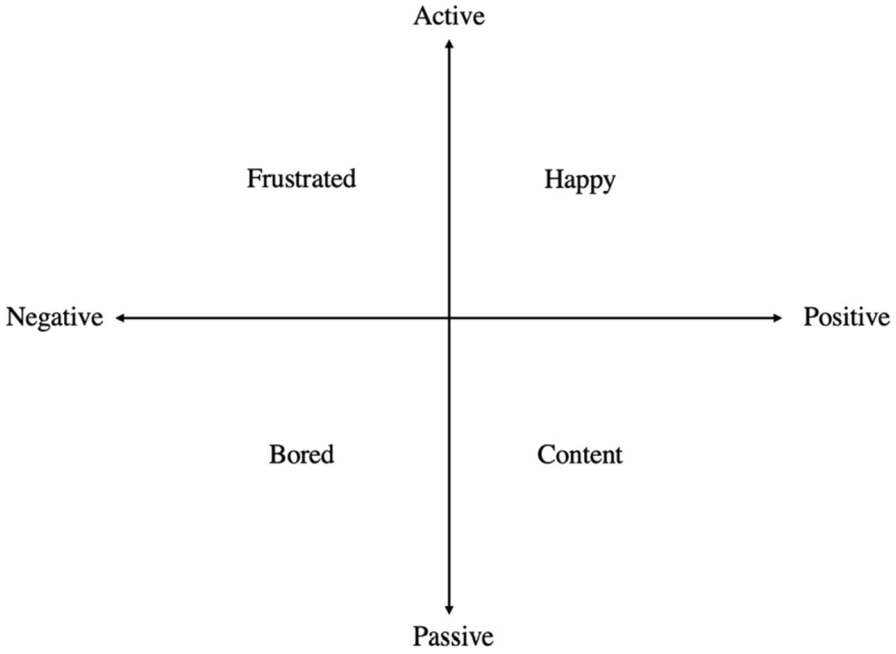

Fig. 2 Adapted version of Russell's (1980, 2003) model of core affect

Russell's model (Fig. 2), which we have labeled as happy (positive/active), content (positive/passive), frustrated (negative/active), and bored (negative/passive). This way of classifying emotions is similar to Pekrun and Perry's (2014; Loderer et al., 2019) taxonomy of achievement motivation, which highlights the long-standing body of research on the role of emotion in academic learning.

The instructor in the present study is an animated young woman who conveys these four emotions through image (including animated gesture, eye gaze, facial expression, and body stance) and voice (i.e., a recorded human voice) as she gives a lecture on the statistical topic of binomial probability. Her voice is a recorded human voice, produced by a young adult woman actor completing her studies in Theater. Her gestures, body movement, facial expression, and eye movements were created to mimic a video of the human actor as she delivered the lecture with each of the four emotions (as specified in Lawson et al., 2021a, 2021b).

The theoretical framework we use to guide our work is an adaptation of the cognitive-affective model of e-learning, represented in Fig. 3, which we have been developing to understand the role of emotional elements in online lessons (Horovitz & Mayer, 2021; Lawson et al., 2021a, 2021b; Mayer, 2021). The cognitive-affective model of e-learning is a more specific framework that we derived from the Cognitive Affective Theory of Learning with Media (CATLM, Moreno & Mayer, 2007), and is consistent with aspects of the Integrated Cognitive Affective Model of Learning with Multimedia (ICAMLM, Plass & Kaplan, 2016). In particular, the cognitive-affective model of e-learning attempts to capture the steps in how the emotional

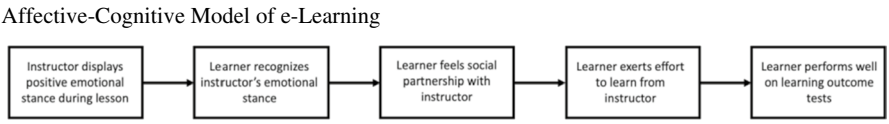

Fig. 3 Affective-cognitive model of e-learning

elements of a multimedia instructional message affect the learner's affective and cognitive processing during learning. As summarized in Fig. 3, the model posits that learners recognize the emotional tone displayed by the instructor (i.e., first link in Fig. 3), which triggers an emotional response within the learner such as feeling the same emotion as the instructor (i.e., second link in Fig. 3), which primes a cognitive response in the learner such as exerting more effort to make sense of the instructional message (i.e., the third link), resulting in a better learning outcome (i.e., fourth link). Our focus in the present study is on the first link in e-learning, which involves the students' recognition of the emotional tone of the instructor. In particular, we examine the link between the emotional tone of the instructor's voice and the learner's recognition of the instructor's emotional tone. Our rationale for focusing on this first step is that it is crucial for initiating the chain of affective and cognitive processes that lead to improvements in learning outcomes. Therefore, it is useful for both theory and practice to calibrate the emotional cues in the instructional message, in order to determine, for example, whether the instructor's voice is enough to initiate the learning processes summarized in Fig. 3.

In the present study, participants view short video clips from an instructional video on statistics and they rate each clip on 5-point scales for happy, content, frustrated, and bored emotional tone. Each clip contains slides about a statistical topic and an animated instructor who conveys a happy, content, frustrated, or bored emotional tone as she lectures based on her animated gesture, eye-gaze, facial expression, body stance, and her recorded human voice (agent present) or based solely on her recorded human voice (agent absent condition). We are interested in (1) whether people can recognize the emotional tone of the instructor just as well from the instructor's voice alone (i.e., the agent absent condition) as from the instructor's image and voice (i.e., the agent present condition) in line with the voice hypothesis, and (2) whether people can correctly recognize the emotional tone of the instructor, particularly whether the instructor's emotional stance is positive or negative, in line with the first step in the positivity hypothesis.

First, based on the voice hypothesis, we predict that the recorded human voice of the instructor should carry enough affective cues that participants should be able to recognize the emotional tone of the lesson, regardless of whether there is an animated agent on the screen or not (hypothesis 1). The pattern of ratings (i.e., for the happy, content, frustrated, and bored scales) should be the same for the agent absent condition as for the agent present condition when the instructor is happy (hypothesis 1a), content (hypothesis 1b), frustrated (hypothesis 1c), and bored (hypothesis 1d).

Second, based on the affective-cognitive model of e-learning, participants should be able to the distinguish between different emotional tones (hypothesis

2). Happy instructors should be rated higher on the happy scale than on the scales for all other emotions (hypothesis 2a), content instructors should be rated as more content than all other emotions (hypothesis 2b), frustrated instructors should be rated as more frustrated than all other emotions (hypothesis 2c), and bored instructors should be rated as more bored than all other emotions (hypothesis 2d). More specifically, based on research showing that people are most sensitive to the positive–negative dimension (Loderer, et al., 2020), we propose a positivity hypothesis, which posits that people can distinguish between whether an instructor is displaying a positive or negative emotion. Happy and content instructors should be rated higher on the happy and content scales (hypothesis 2e) whereas frustrated and bored instructors should be rated higher on the frustrated and bored scales (hypothesis 2f).

Method

Participants

Participants were recruited from Amazon Mechanical Turk (Mturk) and had to currently reside in the Unites States to be allowed to participate. There were 100 participants in total, with 34 female participants, 65 male participants, and 1 person identifying as a different gender. The mean age of participants was 36.86 years old ($SD = 11.48$), ranging from 21 to 69 years old. Of the 100 participants, 96 indicated that they were born in the United States. Additionally, 6 participants indicated they were "Asian/Asian-American," 20 participants indicated they were "Black/African/African-American," 5 participants indicated they were "Hispanic/Latinx," 1 participant indicated they were "Native American," 66 participants indicated they were "White/Caucasian," and 2 participants reported they were a mix of different ethnicities.

Design

The experiment used a 2 (instructor presence: instructor present and instructor absent) × 4 (emotion of instructor: happy, content, frustrated, and bored) within-subjects design. All participants saw 2 sets of videos clips (each covering a portion of a statistics lesson) displaying each of the 4 emotions for each of the 2 instructor presence conditions, yielding a total of 16 clips.

Materials

The materials were all computer-based and included 16 video clips and rating scales associated with each video clip, as well as a postquestionnaire.

Video Clips

This study included 16 video clips taken from a lesson on binomial probability. Eight of the video clips included an animated pedagogical agent standing next to a screen displaying different slides. Of these eight video clips, 4 of these clips showed the same information on an introduction to binomial probability in different emotional tones: happy (example clip: https://youtu.be/JXiPpsm7lPA, content (example clip: (example clip: https://youtu.be/1gTXeu6UuP8), frustrated, (example clip: https://youtu.be/xfQ1n5lVcBg), and bored (example clip: https://youtu.be/NYQc9zsyVEk). The other 4 of these video clips consisted of a clip discussing joint probability from the lesson, and once again displayed this same information in each of four different emotional tones.

The animated pedagogical agents were created based on basic principles of emotional design with gesture, eye-contact, body movement, and facial expression, in which positive agents displayed outward gesture (e.g., trusting one's arms outward), leaning forward, made frequent eye-contact, and showed a smile, whereas negative agents displayed inward gesture (e.g., folding one's arms next to one's body), leaned away from the audience, made infrequent eye contact, and showed inattentive facial expression (as described in Adamo et al., 2021). In addition, animated characters were created to parallel the gesture, body stance, facial expression, and eye movements in a video of an actor who was asked to present the lesson in each of four emotional states (i.e., happy, content, frustrated, or bored) and was monitored to correct any deviations from the basic principles of emotional design (as described in Adamo et al., 2021).

The other 8 video clips were organized in the same way (2 sets of clips, each explaining the material using four different emotional tones), but these video clips did not include an animated pedagogical agent on screen. The agent was missing, but the screen displaying the different slides was still visible along with the same recorded human voice with each of the four emotional tones: happy (example clip: https://youtu.be/qH0G4WXW6SM), content (example clip: https://youtu.be/wgLpDz9zcgI), frustrated (example clip: https://youtu.be/QR7-2Nbk_Mw), and bored (example clip: https://youtu.be/5aIry9fBiA0).

The eight video clips displaying the first part of the lesson, explaining an introduction to binomial probability, were displayed in a randomized order. Once the participant saw all eight of these video clips, the second set of eight video clips displaying the second part of the lesson, explaining joint probability, were displayed in a randomized order.

Video Clip Ratings

After each video clip, participants were asked to rate the emotional tone of the video. There were six questions displayed after each video. The first four questions asked the participants, "Please slide the bar to the number associated with the level at which you think the video portrayed these emotions:" with a 5-point slide bar for "Happy," "Content," Frustrated," and "Bored." Each slide bar went from "1—Not at All" to "5—Very." The next question asked participants, "Please slide the

bar to the number associated with the level at which you think the video was active/passive." Participants responded on a 5-point slide bar from "1—Passive" to "5—Active." The last question participants responded to asked, "Please slide the bar to the number associated with the level at which you think the video was pleasant/unpleasant." Participants responded on a 5-point slide bar from "1—Unpleasant" to "5—Pleasant."

Postquestionnaire

After watching all the videos and rating them, participants completed a short postquestionnaire. Participants were first asked, "How interesting was the presented material?" and asked to rate on a 5-point slide bar from "1—Not at all interesting" to "5—Very interesting." They were also asked, "How much knowledge did you have about binomial probability prior to this study?" and asked to rate on a 4-point slide bar from "1—None" to "4—Extensive." Then, participants were able to write comments about the videos and the emotional tone of the videos. Lastly, participants were asked to provide demographic information, including age and gender.

Procedure

Participants volunteered for this study on the crowdsourced platform, Amazon Mechanical Turk (Mturk). In order to participate, they clicked a Qualtrics link. Participants were first shown a consent page. Once they agreed to participate, they moved forward in the survey. They were asked to prove they were human by using reCAPTCHA, followed by being given onscreen printed instructions on how to participate in the study. Then, they watched each of the 16 videos and rated the emotional tone after each video. Once done with that, they completed the postquestionnaire and were compensated $3 for their time.

Results

Statistical Analyses

For each type of instructor (i.e., happy, content, frustrated, and bored), we averaged the ratings across two clips on the happy, content, frustrated, and bored scales for the agent present and agent absent conditions. Then, we conducted 2 (agent condition: agent present vs. agent absent) $\times$ 4 (emotion rating: for happy, content, frustrated, and bored scales) repeated measures ANOVAs, followed up by pairwise tests. The voice hypothesis predicts no significant interaction between the two agent conditions and the four emotion ratings. The cognitive-affective model of e-learning predicts a main effect of emotion, in which the target emotion is rated higher than each of the others. More specifically, to test the positivity hypothesis, we conducted a 2 (agent condition) $\times$ 2 (valence: positive vs. negative emotion scales) $\times$ 2 (activity: active vs. passive emotion scales) repeated measures ANOVAs. The positivity hypothesis

predicts a main effect of valence in which happy and content instructors are rated higher on positive scales and frustrated and bored instructors are rated higher on negative scales.

Happy Videos

Voice Hypothesis

Hypothesis 1a states that participants should be equivalent in the pattern of ratings of the emotion tone of the happy instructor in matching clips with voice alone as with voice and onscreen agent. Means and standard deviations of ratings for the happy instructor are displayed in Table 1 by agent presence condition (agent absent and agent present) and emotion scale (happy, content, frustrated, and bored). There was not a significant main effect of agent presence, $F(1, 99) = 0.59$, $p = 0.443$, nor an interaction, $F(3, 297) = 0.39$, $p = 0.759$. This pattern is consistent with the voice hypothesis, as there is not any evidence that the pattern of ratings was different for the agent absent and agent present conditions.

Positivity Hypothesis

Hypothesis 2a states that participants will rate the happy instructor higher on the happy scale than on any of the other scales. A 2×4 ANOVA showed that there was a significant main effect of emotion, $F(3, 297) = 151.04$, $p < 0.001$. To further understand this main effect, multiple *t*-tests were run on the emotion ratings using a Bonferroni correction ($\alpha = 0.017$). Participants gave a higher rating for the happy instructor on the happy scale than the frustrated scale ($p < 0.001$, $d = 1.40$) and the bored scale ($p < 0.001$, $d = 1.15$). There was no significant difference between the happy and content ratings ($p = 0.136$), suggesting that participants did not distinguish between positive emotions.

To focus specifically on the positivity hypothesis, we conducted a 2 (agent present vs. agent absent) × 2 (positive vs. negative emotion) × 2 (active vs. passive emotion) ANOVA. There was a main effect for positive vs. negative emotion in which participants rated the happy instructor significantly higher on positive scales ($M = 3.75$, $SD = 0.71$) than on negative scales ($M = 1.96$, $SD = 1.20$), $F(1, 99) = 182.99$, $p < 0.001$, $d = 1.35$, in line with the positivity principle, and there was no significant interaction with presence of agent, $F(1, 99) = 0.64$, $p = 0.427$.

Table 1 Means and standard deviations on four emotional tone ratings for the happy instructor by agent condition

Agent condition	Happy		Content		Frustrated		Bored	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Agent present	**3.80**	**0.81**	3.72	0.78	1.89	1.17	1.99	1.31
Agent absent	**3.80**	**0.89**	3.70	0.87	1.94	1.23	2.03	1.34

Bolded numbers represent the means and standard deviations of the target emotion. Scale runs from 1 (Not at all) to 5 (Very)

These findings are generally consistent with hypothesis 2a and supportive of the positivity hypothesis, showing that participants were able to differentiate between positive and negative instructors, regardless of whether there was an instructor on the screen or not.

Content Videos

Voice Hypothesis

Hypothesis 1b states that participants should be equivalent in rating the emotional tone of the content instructor (on the happy, content, frustrated, and bored scales) for the condition with voice alone as for the condition with voice and onscreen agent. Means and standard deviations of ratings for the content instructor are displayed Table 2 by agent presence condition (agent absent and agent present) and emotion scale (happy, content, frustrated, and bored). There was no main effect of agent presence, $F(1, 99) = 1.70$, $p = 0.195$, nor was there an interaction, $F(3, 297) = 0.86$, $p = 0.464$. These results are consistent with hypothesis 2b and the voice hypothesis from which it is derived.

Positivity Hypothesis

Hypothesis 2b states that participants will rate the content instructor higher on the content scale than on any of the other scales. In support, the 2×4 ANOVA showed there was a significant main effect of emotion, $F(3, 297) = 104.30$, $p < 0.001$, and Bonferroni corrected t-tests ($\alpha = 0.017$) showed that participants rated the content video as more content than happy ($p < 0.001$, $d = 0.40$), frustrated ($p < 0.001$, $d = 1.46$), and bored ($p < 0.001$, $d = 1.03$). In line with the positivity hypothesis, a 2 (agent present vs. agent absent) $\times$ 2 (positive vs. negative emotion) $\times$ 2 (active vs. passive emotion) ANOVA showed a main effect for positive vs. negative emotions in which participants rated the content instructor significantly higher on positive scales ($M = 3.42$, $SD = 0.67$) than on negative scales ($M = 2.15$, $SD = 1.14$), $F(1, 99) = 142.25$, $p < 0.001$, $d = 1.18$, and there was no significant interaction with presence of agent, $F(1, 99) = 1.70$, $p = 0.195$. These findings are consistent with hypothesis 2b and the positivity hypothesis.

Table 2 Means and standard deviations on four emotional tone ratings for the content instructor by agent condition

Agent condition	Happy		Content		Frustrated		Bored	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Agent present	3.26	0.89	**3.58**	**0.75**	1.91	1.19	2.31	1.27
Agent absent	3.32	0.85	**3.54**	**0.75**	2.00	1.27	2.37	1.26

Bolded numbers represent the means and standard deviations of the target emotion. Scale runs from 1 (Not at all) to 5 (Very)

Frustrated Videos

Voice Hypothesis

Hypothesis 1c states that participants should be equivalent in rating the emotional tone of the frustrated instructor (on the happy, content, frustrated, and bored scales) for the condition with voice alone and the condition with voice and onscreen agent. Means and standard deviations of ratings for the frustrated instructor are displayed Table 3 by agent presence condition (agent absent and agent present) and emotion scale (happy, content, frustrated, and bored). There was no main effect of agent presence, $F(1, 98) = 0.11$, $p = 0.741$, but there was an interaction between agent presence and emotion, $F(3, 294) = 11.19$, $p < 0.001$, in which the agent present condition appears to be better than the agent absent condition in recognizing the instructor's frustrated emotional tone. To further investigate this interaction, multiple *t*-tests were run on the emotion ratings using a Bonferroni correction ($\alpha = 0.008$). For the video with an instructor present, the frustrated instructor was rated as more frustrated than happy ($p < 0.001$, $d = 0.42$) and content ($p < 0.001$, $d = 0.40$), but there was no significant difference between frustrated and bored ratings ($p = 0.185$). For the video without an instructor, the frustrated instructor was rated similarly across all emotions ($ps > 0.210$). In contrast to all other videos, these findings are not consistent with hypothesis 1c nor the voice principle from which it is derived because participants were better able to recognize the frustrated emotion as different from other emotions when there was an instructor on screen, but not when there was only voice.

Positivity Hypothesis

Hypothesis 2c states that participants will rate the frustrated instructor higher on the frustrated scale than on any of the other scales. In support, the 2×4 ANOVA showed there was a significant main effect of emotion, $F(3, 294) = 7.08$, $p < 0.001$, and Bonferroni-corrected *t*-tests ($\alpha = 0.017$) showed that participants rated the frustrated video as more frustrated than happy ($p = 0.004$, $d = 0.30$) and content ($p = 0.007$, $d = 0.29$), but there was no significant difference between the frustrated and bored ratings ($p = 0.430$). In line with the positivity hypothesis, a 2 (agent present vs. agent absent) $\times$ 2 (positive vs. negative emotion) $\times$ 2 (active vs. passive emotion) ANOVA, showed a main effect for positive vs. negative emotion in which participants rated

Table 3 Means and standard deviations on four emotional tone ratings for the frustrated instructor by agent condition

Agent condition	Happy		Content		Frustrated		Bored	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Agent present	2.44	1.32	2.48	1.27	**3.26**	**1.21**	3.11	1.18
Agent absent	2.68	1.22	2.74	1.20	**2.91**	**1.23**	2.91	1.21

Bolded numbers represent the means and standard deviations of the target emotion. Scale runs from 1 (Not at all) to 5 (Very)

the frustrated instructor significantly higher on negative scales ($M = 3.04$, $SD = 0.96$) than on positive scales ($M = 2.58$, $SD = 1.17$), $F(1, 99) = 9.07$, $p < 0.003$, $d = 0.31$; however, this main effect was mediated by a significant interaction between agent type and positive–negative valence, $F(1, 99) = 15.18$, $p < 0.001$, in which the effect was stronger for the agent present condition that the agent absent condition. These findings are generally consistent with hypothesis 2c and the positivity hypothesis, but only for the agent present condition.

Bored Videos

Voice Hypothesis

Hypothesis 1d states that participants should be equivalent in rating the emotional tone of the bored instructor (on the happy, content, frustrated, and bored scales) for the condition with voice alone as for the condition with voice and onscreen agent. Means and standard deviations of ratings for the bored instructor are displayed in Table 4 by agent presence condition (agent absent and agent present) and emotion scale (happy, content, frustrated, and bored). There was no main effect of agent presence, $F(1, 99) = 0.40$, $p = 0.527$, nor an interaction, $F(3, 297) = 1.83$, $p = 0.143$. These results are consistent with hypothesis 2d and the voice hypothesis from which it is derived.

Positivity Hypothesis

Hypothesis 2d is that participants will rate the bored instructor higher on the bored scale than on any of the other scales. The 2×4 ANOVA showed there was a significant main effect of emotion, $F(3, 297) = 84.02$, $p < 0.001$. To further investigate this interaction, multiple t-tests were run on the emotion ratings using a Bonferroni correction ($\alpha = 0.017$). In line with hypothesis 2d, participants rated the bored instructor as more bored than happy ($p < 0.001$, $d = 1.02$), content ($p < 0.001$, $d = 1.03$), and frustrated ($p < 0.001$, $d = 0.99$). To focus specifically on the positivity hypothesis, we conducted a 2 (agent present vs. agent absent)×2 (positive vs. negative emotion)×2 (active vs. passive emotion) ANOVA. In line with the positivity hypothesis, there was a main effect for positive vs. negative emotion in which participants rated the bored instructor significantly higher on negative scales ($M = 3.51$, $SD = 0.79$) than on positive scales ($M = 2.13$, $SD = 1.25$), $F(1, 99) = 83.55$, $p < 0.001$, $d = 0.91$. These findings are

Table 4 Means and standard deviations on four emotional tone ratings for the bored instructor by agent condition

Agent condition	Happy		Content		Frustrated		Bored	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Agent present	2.10	1.35	2.18	1.33	2.89	1.08	**4.17**	**0.99**
Agent absent	2.10	1.28	2.16	1.28	2.99	1.11	**4.02**	**1.10**

Bolded numbers represent the means and standard deviations of the target emotion. Scale runs from 1 (Not at all) to 5 (Very)

consistent with the positivity hypothesis, showing that participants were able to differentiate between positive and negative instructors, regardless of whether there was an instructor on the screen or not.

In sum, the major new contribution of this study is to show that people are able to detect whether an instructor is displaying positive or negative emotion from a short clip of a narrated slideshow, without the need to see an embodied onscreen agent. Thus, we conclude that the instructor's voice is sufficient to accomplish the first step in the cognitive-affective model of e-learning–conveying positive or negative emotion that is detected by the learner.

Discussion

Voice Hypothesis

Consistent with the voice hypothesis, voice alone (i.e., agent absent condition) was sufficient to convey the emotional tone for three of the four instructors, with no additional benefit created by adding an embodied onscreen agent (i.e., agent present condition). Specifically, participants were just as accurate in rating the happy, content, and bored instructors on the four emotion scales based solely on voice (agent absent condition) as based on voice with onscreen embodied agent (agent present condition). However, the participants needed both to see and hear the frustrated instructor rather than simply hear her in order give an accurate pattern of ratings. This suggests that voice generally is a powerful source of affective information, but in certain situations (e.g., a frustrated instructor) it needs to be supplemented with a visual image of an embodied onscreen agent. This work contributes to the research base showing the powerful contribution of human voice to convey emotion in instructional messages.

On the theoretical side, we conclude that there is partial support for the voice hypothesis, particularly for positive emotions as shown by findings in which voice was sufficient to convey happy and content emotion (and one of the two negative emotions). Overall, a major theoretical implication of this study is that voice is a powerful source of affective information in multimedia instructional messages as represented by the first step in the cognitive-affective model of e-learning represented in Fig. 3.

On the practical side, we conclude that in many situations the same emotional impact can be achieved by an instructor's voice as by an instructor's voice and embodied image on the screen. This means that when the goal is to convey positive emotion, instructional designers may be able to present narrated slides with voice-over rather than to go to the expense of also adding onscreen embodied agents. However, when the goal is to convey more nuanced or unexpected emotions, such as frustration, it may be necessary to present both voice and embodiment cues.

Positivity Hypothesis

Consistent with the cognitive-affective model of e-learning, participants generally rated the instructor higher on the target emotion conveyed by the instructor than

on other emotions. However, for two of the four types of instructors, participants were not able to distinguish between the two positive emotions (i.e., rating the happy instructor high on both the happy scale and the content scale) or the two negative emotions (i.e., rating the frustrated instructor high on both the frustrated scale and bored scale). Taking a more focused approach, there was consistent support for the positivity hypothesis across all four types of instructors, in which happy and content instructors were rated higher on happy and content scales whereas frustrated and bored instructors were rated higher on frustrated and bored scales.

On the theoretical side, we conclude that there is evidence to support the positivity hypothesis in that the results show that people are sensitive as to whether an instructor is conveying positive or negative emotional tone. Usually they can make that determination based solely on voice, but in some situations (i.e., for a frustrated instructor) they also need embodiment cues to be able to make the distinction. Overall, a major theoretical implication of this study is that positive–negative emotional tone is a salient dimension for people as they process multimedia instructional messages.

On the practical side, this work has implications for the design of socially-sensitive intelligent tutoring systems based on artificial intelligence. We conclude that instructional designers should carefully consider the degree to which instructional messages convey positive or negative emotional tone. To the extent that positive emotion is related to improved learning outcomes, as hypothesized in the cognitive-affective model of e-learning, instructional designers should ensure that instructional messages convey positive emotional tone through voice, and if necessary, through an onscreen agent's embodiment.

Limitations and Future Directions

There are several limitations to this study that should be noted. First, the clips shown to participants were very short, i.e., all were under a minute. Because of this, it is hard to generalize how well participants may be able to recognize the emotional tones of a longer lesson. Additionally, the affect of an instructor, and thus the emotional tone of the voice, may change over time in a longer lecture. Because of this, it is more difficult to determine how learners may react to changes in emotional tone. Future research should investigate how learners identify the emotional tones of an instructor in a natural setting and over time. However, short clips may be characteristic of some intelligent learning environments.

Additionally, the results of this study may be difficult to generalize to all animated instructors when compared to no instructor. The way in which these animated instructors were created is not necessarily how all animated pedagogical agents are created, and thus suggests that different results may be gleaned from the use of different animated instructors. For example, other programs may be better able to create embodied onscreen agents that convey positive emotions, which could create differences between instructor present and instructor absence conditions, not seen in this study. Future research should investigate the voice hypothesis using different animated agents.

This is a preliminary study that focuses just on the first step in the cognitive-affective model of e-learning. We did not examine effects on learning processes and outcomes. It should be noted that just because people can detect the emotion being displayed by the voice in a video clip, this may not necessarily mean they will feel the emotion in a way that influences their learning processes and outcomes. Although, other work shows that people are affected by the emotion they detect in embodied pedagogical agents (Horovitz & Mayer, 2021; Lawson et al., 2021a, 2021b), a main new contribution of this study is that people are able to detect the whether the instructor is exhibiting positive or negative emotion from voice alone. Further research should be done investigating how including an instructor's visual presence impacts how well learners engage with and learn from the presented material.

Future research should also investigate how the gender of an instructor could influence the voice hypothesis. The instructor in this experiment was a young woman with a feminine voice. There could be differences in how learners recognize the emotions of masculine voices compared for feminine voices.

Conclusion

In conclusion, first, this study provides evidence for the voice hypothesis, which posits that voice is a powerful vehicle for conveying emotion in multimedia instructional messages. Second, this study provides evidence for the positivity hypothesis, which posits that people are particularly sensitive to whether a multimedia instructional message conveys a positive or negative emotional tone.

Acknowledgements The onscreen agents were produced by Nicoletta Adomo-Villani, Bedrich Benes Xingyu Lei, and Justin Cheng at Purdue University.

Author contributions APL worked on preparing materials, collecting data, analyzing data, and assisting in report writing. REM worked on research design, project management, and report writing.

Funding This project was supported by Grant 1821833 from the National Science Foundation.

Data Availability The data set and materials may be obtained from the corresponding author upon request.

Declarations

Conflict of interest The authors report no conflict of interest.

Consent to Participate Participants provided written consent to participate in the study.

Consent for Publication Participants were informed that their individual data would not be published.

Ethical approval We followed standards for ethical treatment of human subjects and received IRB approval at the University of California, Santa Barbara. This project was conducted in line with guidelines for research with human subjects and had IRB approval from UCSB's Human Subjects Committee.

References

Adamo, N., Benes, B., Mayer, R. E., Lei, X., Wang, Z., Meyer, Z., & Lawson, A. (2021). Multimodal affective pedagogical agents for different types of learners. In D. Russo, T. Ahram, W. Karwowski, G. Di Bucchianico, & R. Taiar (Eds.), *Advances in intelligent systems and computing: Lecture notes in computer science* (Vol. 1322, pp. 218–224). Springer.

Baylor, A., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. *EdMedia+ Innovate Learning* (pp. 452–458). Association for the Advancement of Computing in Education (AACE).

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161–178.

Calvo, R. A., D'Mello, S., Gratch, J. M., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. Oxford University Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. F. (Eds.). (2000). *Embodied conversational agents*. MIT Press.

Clarebout, G., Elen, J., Johnson, W. L., & Shaw, E. (2002). Animated pedagogical agents: An opportunity to be grasped? *Journal of Educational Multimedia and Hypermedia, 11*(3), 267–286.

Craig, S. D., Gholson, B., & Driscoll, D. M. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features and redundancy. *Journal of Educational Psychology, 94*(2), 428–434.

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

Craig, S. D., & Schroeder, N. L. (2018). Design principles for virtual humans in educational technology environments. *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp. 128–139). Taylor and Francis.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105*(4), 1082–1099.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.

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.

Edwards, R., & Kortum, P. (2012). He says, She says: Does voice affect usability? *Proceedings of the human factors and ergonomics society annual meeting* (pp. 1486–1490). SAGE Publications.

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013). The effects of culturally congruent educational technologies on student achievement. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education: Lecture notes in computer science* (Vol. 7926, pp. 493–502). Berlin: Springer.

Fiorella, L. (2022). The embodiment principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 286–295). Cambridge University Press.

Fiorella, L., & Mayer, R. E. (2022). Principles based on social cues in myltimedia learning: Personalization, voice, embodiment, and image principles. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 277–285). Cambridge University Press.

Guo, Y. R., & Goh, D. H. L. (2015). Affect in embodied pedagogical agents: Meta-analytic review. *Journal of Educational Computing Research, 53*(1), 124–149.

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*. https://doi.org/10.1016/j.chb.2021.106724

Johnson, W. L., & Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial Intelligence in Education, 26*(1), 25–36.

Johnson, W. L., & Rickel, J. W. (1997). Steve: An animated pedagogical agent for procedural training in virtual environments. *ACM SIGART Bulletin, 8*(1–4), 16–21.

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 11*(1), 47–78.

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021a). Do learners recognize and relate to the emotions displayed by virtual instructors? *International Journal of Artificial Intelligence in Education, 31*(1), 134. https://doi.org/10.1007/s40593-021-00238-2

Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021b). Recognizing the emotional state of human and virtual instructors. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2020.106554

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997, March). The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 359–366).

Liew, T. W., Tan, S. M., Tan, T. M., & Kew, S. N. (2020). Does speaker's voice enthusiasm affect social cue, cognitive load and transfer in multimedia learning? *Information and Learning Sciences, 121*(3/4), 117–135. https://doi.org/10.1108/ILS-11-2019-0124

Loderer, K., Pekrun, R., & Lester, J. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2018.08.002

Loderer, K., Pekrun, R., & Plass, J. L. (2019). Emotional foundations of game-based learning. In J. L. Plass, R. E. Mayer, & B. D. Homer (Eds.), *Handbook of game-based learning* (pp. 111–152). MIT Press.

Makransky, G., Wismer, P., & Mayer, R. E. (2019). A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning, 35*(3), 349–358.

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. (2020). Searching for the role of emotions in e-learning. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2019.05.010

Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). 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.

McLaren, B. M., DeLeeuw, K. E., & Mayer, R. E. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education, 53*, 574–584.

Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309–326.

Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Nass, C., Foehr, U., Brave, S., & Somoza, M. (2001). The effects of emotion of voice in synthesized and recorded speech. *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition.* AAAI.

Ogan, A., Finkelstein, S., Walker, E., Carlson, R., & Cassell, J. (2012). Rudeness and rapport: Insults and learning gains in peer tutoring. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Pamourgia (Eds.), *Intelligent tutoring systems: Lecture notes in computer science* (Vol. 7315, pp. 11–21). Berlin: Springer.

Pekrun, R., & Perry, R. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 120–141). Taylor and Francis.

Picard, R. W. (2000). *Affective computing*. MIT Press.

Plass, J. L., & Kaplan, U. (2016). Emotional design in digital media for learning. In S. Y. Tettegah & M. Gartmeier (Eds.), *Emotion, technology, design, and learning* (pp. 131–161). Elsevier Academic Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Ryu, J., & Ke, F. (2018). Increasing persona effects: Does it matter the voice and appearance of animated pedagogical agent. *Educational Technology International, 19*(1), 61–91.

Schroeder, N. L., Adesope, O. O., & Gilbert, R. B. (2013). How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research, 49*(1), 1–39.

Schroeder, N. L., & Craig, S. D. (2021). Learning with virtual humans: Introduction to the special issue. *Journal of Research on Technology in Education, 53*(1), 1–7. https://doi.org/10.1080/15391523.2020.1863114

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human Computer Studies, 66*, 96–112.

Wu, C. H., Huang, Y. M., & Hwang, J. P. (2016). Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology, 47*(6), 1304–1323.