Does speaker's voice enthusiasm affect social cue, cognitive load, and transfer in multimedia learning?

Tze Wei Liew
Human-Centric Technology Interaction SIG, Multimedia University,
Melaka, Malaysia

Su-Mae Tan
Department of Information Science and Technology, Multimedia University,
Melaka, Malaysia

Teck Ming Tan
Oulu Business School, University of Oulu, Oulu, Finland, and

Si Na Kew
Language Academy, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Abstract

Purpose — The present study examined the effects of voice enthusiasm (enthusiastic voice versus calm voice) on social ratings of speaker, cognitive load, and transfer performance in multimedia learning.

Design/methodology/approach — Two laboratory experiments were conducted in which learners learned from a multimedia presentation about computer algorithm that was narrated by either an enthusiastic human voice or calm human voice.

Findings — Results from experiment 1 revealed that enthusiastic voice narration led to higher social ratings of speaker and transfer performance when compared to the calm voice narration. Experiment 2 demonstrated that enthusiastic voice led to higher affective social ratings (human-like and engaging) and transfer performance as compared to the calm voice. Moreover, it was shown that calm voice prompted higher germane load than enthusiastic voice, which conformed to the argument that prosodic cues in voice can influence processing in multimedia learning among non-native speakers.

Originality/value — Prior studies have examined voice effects related to mechanization, accent, dialect, and slang in multimedia learning. This study extends to examining the effects of voice enthusiasm in multimedia learning.

Keywords: voice, enthusiasm, social agency theory, immediacy, multimedia learning, cognitive load

1. Introduction

E-Learning is a form of multimedia instruction in which information are represented by both visual (e.g., diagrams, maps, animations, and illustrations) and verbal elements (spoken narrations and on-screen texts) (Mayer, 2017). Mayer and Moreno (1998) proposed the cognitive theory of multimedia learning to describe the mechanism in which visual and verbal information from multimedia presentation are processed cognitively. Based on this framework, multimedia learning involves three crucial cognitive processes, which are selecting, organizing, and integrating. Selecting refers to the process engaged by learners to selectively focus on relevant visual and verbal information. After the selection process, learners engaged in the organizing process when visual and verbal information are formed into a meaningful and coherent representation. Lastly, the integration is when prior knowledge is activated and used to build connection between the newly presented information and pre-existing knowledge schema.

In accordance with the framework of cognitive theory of multimedia learning, Mayer and his colleagues have established a set of *multimedia principles*, which are evidenced-based recommendations for instructional design that are aimed to effectively produce deep meaningful learning (Mayer and Pilegard, 2005; Clark and Mayer, 2016; Mayer, 2017). One of the multimedia principles is the *voice principle*, which posit that people learn more deeply when the words in a multimedia message are spoken in a human voice with standard accent rather than in a machine-synthesized voice or a human voice with foreign accent (Mayer, 2005; Mayer *et al.*, 2003; Atkinson *et al.*, 2005; Mayer and DaPra, 2012). This effect is attributed to the social agency theory, which states that the multimedia instruction should be designed to trigger social interaction schema in learner's mind, which then lead learners to assume the computer source as a social partner (Mayer *et al.*, 2003; Mayer, 2005). Assuming a social interaction stance will encourage learners to deeply engage in the process of selecting, organizing, and integrating instructional message. Therefore, in accordance with social agency theory, using human voice with standard accent fosters higher social agency than human voice with foreign accent or computer-synthesized voice; and thus leads to superior learning outcome.

However, beyond the voice characteristics of accents and mechanization (i.e., human or machine-synthesized voice), there is a paucity of studies that investigate the voice enthusiasm in the multimedia learning environment. Mayer *et al.* (2003) noted that "additional work is needed to pinpoint which aspects of voice are most important in promoting deep learning" (p. 424). Furthermore, many multimedia presentations feature "invisible" narrators (disembodied source speaker that have no visual features e.g., face and body) where voices are the only source social cues (e.g., Khan videos), with narrations delivered via a pleasant but calm voice that does not convey high enthusiasm. While the current literature has shown that the instructor's high enthusiasm behaviors which involve visual nonverbal cues such as body gesture and facial expression can benefit engagement and learning (Wang *et al.*, 2019; Guo *et al.*, 2014; Liew *et al.*, 2017); however, it is not known whether the positive effect of high enthusiasm can also manifest in a multimedia environment presented by a voice-only virtual speaker. Thus, grounded on the social agency theory, this paper aims to examine if an enthusiastic voice as compared to the calm voice (low enthusiasm) will differently affect the perceived social cues, cognitive load, and learning outcome of learners.

2. Literature review

2.1. Social agency theory

According to the social agency theory, imbuing multimedia presentation with verbal and non-verbal stimuli that convey social cues, can lead learners to interpret the multimedia message as a social communication process which in turn encourage learners to put more effort in understanding the materials (Mayer *et al.*, 2003; Atkinson *et al.*, 2005). Specifically, social agency theory defines five steps that explain the effects of social cues in a multimedia presentation (Linek *et al.*, 2010). First, it is postulated that stimuli such as voice and image of speakers embedded in multimedia presentation can act as social cues. Notably, not all stimuli express the same degree of social cues; for instance, human voice has been shown to convey stronger social cues than a machine-synthesized voice (Mayer *et al.*, 2003; Atkinson *et al.*, 2005). Second, these social cues prompt learners to regard the process of learning with multimedia presentation as a social communication, rather than pure information delivery. Third, following the social communication assumption in learners' mind, learners will apply human-to-human social communication rules to their interaction with computers. This notion is derived from the media equation theory that posits that people tend to respond socially to media as they would to another person based on the cues conveyed by the media (Nass and Brave, 2005). Fourth, based on the social rule of Grice *et al.*'s cooperation principle (1975); learners will assume that the speaker is trying to convey a clear and meaningful message, and thus reciprocate by putting in more effort to understand the message. Within the framework of multimedia learning, this means that learners will increase their levels of effort in selecting, organizing, and integrating the learning contents from the multimedia presentation. Fifth, the higher commitment of efforts during the multimedia learning process will produce better meaningful learning which in turn leads to better transfer performance.

Based on the social agency framework, a number of multimedia design principles that facilitate multimedia learning through social cues have emerged, namely the personalization, voice, and embodiment principles (Mayer, 2005). The personalization principle involves the strategy of converting words in multimedia message from formal to conversational style. Instructional designers can create conversational format in multimedia presentation by following two rules — 1) using first and second person pronouns to address learners (e.g., "I", "we", "you", "our"), and 2) adding sentences in which the speaker makes direct self-revealing comments. According to the voice principle, multimedia narrations using human voice with standard accent should convey stronger social cues than multimedia narrations that use machine-synthesized voice or a human voice with foreign accent. Finally, the embodiment principle postulates that an on-screen animated agent that displays higher levels of social expressions (e.g., eye gaze, facial expression, human-like gestures) should prime a higher sense of social stance in learners, which in turn encourage learners to commit more efforts when processing the learning contents (Mayer and DaPra, 2012). Contrariwise, an on-screen animated agent that displays low levels of expressions (e.g., static facial expressions and gestures) conveys less of a sense of social presence and therefore, induces weak activation of social responses in learners' mind. As a result, learners do not increase their level of commitment when processing the multimedia message.

2.2. Voice characteristics and social cues in multimedia learning

When assessing a speaker's appropriateness as a social partner, learners rely on their perceptions regarding the speaker's social qualities (Mayer *et al.*, 2003). One of the factor that influences the perception of social qualities is the speaker's voice. Besides transferring words from a speaker to a listener; a speaker's voice conveys a wide range of socially relevant cues that automatically trigger social responses from people (Nass and Brave, 2005) (Bechtold, 2017). Therefore, in the context of multimedia learning, the characteristics of a speaker's voice influence a learner's perceived social rating of the speaker. The activation of social schema in the learner's mind during the process of multimedia learning is also influenced by the speaker's social cues. Based on the literature, some of the characteristics of a speaker's voice that can affect multimedia learning outcome and social perception are mechanization (human vs. machine-synthesized voice), accent (native vs. foreign accent), gender (male vs. female voice), dialect (regional dialect vs. standard speech), and slang (youth slang vs. standard speech).

With regards to voice mechanization, previous studies have demonstrated that learners who listened to the multimedia presentation narrated with a human voice had higher transfer scores and assigned higher social ratings for the speaker than did learners who listened to the multimedia presentation narrated with a machine-synthesized voice (Mayer *et al.*, 2003; Atkinson *et al.*, 2005). However, more recently, Craig and Schroeder (2017) studied the effects of modern machine-generated voice (i.e, Neospeech speech synthesizer) as narrations in a multimedia presentation by comparing it with classic machine generated voice (i.e., Microsoft speech synthesizer) and pre-recorded human voice. Their analysis indicated that the modern machine-synthesized voice produced the best transfer outcome among learners, as compared to classic machine-synthesized voice and to human voice. It can be inferred that technological advancement of modern speech synthesizer may have rendered the voice effect pertaining to the superiority of human-recorded voice over machine-generated voices as less relevant to multimedia learning today as compared to prior times.

Concerning voice accent, Mayer *et al.* (2003) found that learners who listened to the multimedia presentation with standard-accented human voice (American English) had better transfer performances and gave higher social ratings for the speaker than did learners who listened to the multimedia presentation with foreign-accented human voice (Russian). Ahn (2010) compared four voices ranging from low to heavy accents and found that they did not differently affect learning outcomes. However, it was noted that the learning outcomes were affected only when the learners had indicated beforehand that they did not like a certain accent. This observation indicates that the perceptions of learners regarding the appeal of social cues can actually affect the learning process.

The effects of voice gender of a speaker in the context of multimedia learning was investigated by Linek *et al.* (2010). Their findings indicated that learners who listened to the narration by a female speaker outperformed learners who listened to the narration by a male speaker in terms of problem-solving test scores. The female speaker was also rated as socially more favorable (i.e., attractiveness) than the male speaker.

More recently, the effects of speaker's voice dialect on multimedia learning was examined by Rey and Steib (2013). The researchers hypothesized that the familiarity that comes from listening to a speaker's voice with similar dialect that is relevant to the learners' own social attributes may increase learner's interest and learning outcome. When comparing the effects between the Austrian dialect that was characteristically familiar to the participants (from Austrian lower secondary school) and the standard German speech, their data indicated that the familiar-dialect voice speech had a positive effect on retention, but not for transfer scores. Their results suggest that the feeling of familiarity afforded by stimuli that closely resemble a learner's social characteristics can convey positive social cues to learners.

Based on the familiarity cue hypothesis and extending from the study on voice dialect, Schneider *et al.* (2015) investigated the use of young slang (e.g., "cool" or "absofuckinglutely") in an audio text presentation. The researchers argued that youth slang can convey familiarity cues to learners, which then triggers a social cue that leads to activation of social response in learners' mind. As per the social agency theory, the activation of social schema will encourage learners to invest more cognitive effort during the learning process and subsequently result in better transfer performances. Consistent with their prediction, it was revealed that learners who listened to the audio text that used youth slang had higher transfer scores that did learners who listened to the standard audio text.

2.3. The effects of voice characteristics on cognitive load

Cognitive load theory states that the learning process involves three types of cognitive load i.e., intrinsic, extraneous, and germane loads (Sweller *et al.*, 1998). Intrinsic load refers to the processing effort needed to sufficiently process the inherent difficulty of a particular learning subject. For example, the intrinsic load for addition and subtraction is lower (less difficult) than the intrinsic load for algebra (more difficult). Moreover, leaner's prior knowledge influences intrinsic load; for instance, an expert learner who has high prior knowledge about algebra will experience lower intrinsic load than a novice learner who has low prior knowledge about the subject. Extraneous load refers to the additional mental resources required to process non-essential materials that are caused by poorly-designed instructional format. Extraneous load is undesired as it competes with limited mental resources that are crucial to process information for meaningful learning. Germane load is associated with the mental resources required to create and automate knowledge schema in long-term memory. It has been suggested the when extraneous load is reduced, the resulting available mental resources will be used to increase germane load (Tabbers *et al.*, 2000; Cierniak *et al.*, 2009).

Through the lens of cognitive load theory, it has been argued that narrations that use human voice would be easier to be processed due to its familiarity and consistency with pre-existing conversational schema, and thus impose lesser cognitive demand than narrations that use machine-synthesized voice. However, only a few studies have included the measures of cognitive load when examining the effects of voice characteristics. Further, most of the studies that assessed cognitive load impacts of voice focused on voice mechanization and voice gender (Atkinson *et al.*, 2005; Mayer *et al.*, 2003; Craig and Schroeder, 2017; Linek *et al.*, 2010). One study showed that the learners who listened to the human-voice narration assigned lower perceived difficulty when learning about the subject (cognitive load) than did the learners who listened to the machine-synthesized narration (Mayer *et al.*, 2003). However, other studies revealed no differences of cognitive load between human-voice and machine-synthesized voice (Atkinson *et al.*, 2005; Craig and Schroeder, 2017). Concerning voice gender (i.e., female vs. male speakers), Linek *et al.* (2010) revealed no differences in terms of intrinsic, extraneous, and germane load between the female and male voice narration conditions.

While most studies indicated no differences of cognitive load between human and machine-synthesized voices; however, research in this vein should also be extended to examine other voice characteristics. This extension is crucial to examine whether or not social cues derived from other voice characteristics can impose cognitive load during multimedia learning; particularly when the voice cues are considered to be undesirable, distracting, and frustrating (Davis *et al.*, 2019; Veletsianos, 2012; Wouters *et al.*, 2008). Moreover, most prior studies assessed cognitive load as *perceived difficulty* which did not provide the distinction between intrinsic and extraneous cognitive load. Related to this matter, Davis *et al.* (2019) advocated the use of cognitive load measures that distinguish intrinsic, extraneous, and germane load when assessing the voice effect in multimedia learning, so that researchers can tease apart the effects of voice cues on different types of cognitive load.

As compared to other types of cognitive load, germane load has received the least attention in research on voice effects in multimedia learning. In a recent study, Davis et al. (2019) argued that for non-native English speakers learning from a multimedia environment presented in English, a weak-prosodic human voice narration will prompt higher germane load than a strong-prosodic human voice narration and modern computer voice. This argument was predicated on the notion that as compared to native speakers, non-native speakers are generally less efficient in processing prosodic cues such as pitch, tempo, stress, intonation, melody, loudness, accent and pause (Akker and Cutler, 2003; Goh, 2000). The results of their experiment conformed to this prediction — it was shown that non-native speakers who listened to the multimedia presentation with weak-prosodic human voice reported higher germane load than did non-native speakers who listened to the multimedia presentation with strong-prosodic human voice and modern computer voice.

2.4. Display of enthusiasm in teaching and learning

Within the context of education, displayed enthusiasm describes teaching delivery of "stimulating, energetic and motivating" (Keller *et al.*, 2016). Indicators of enthusiasm in teaching include vocal delivery that have great and sudden changes, uplifting intonations, and many changes in tone and pitch; eyes that light up, eyebrows raised, and constant eye contact with listeners; gestures like clapping, head nods, and frequent body movements; and facial expression that appear vibrant, joyful, and demonstrative (Collins, 1978). The literature has shown some evidence that displayed verbal and nonverbal enthusiasm by instructors can benefit learning. For instance, children who listened to a reader who displayed enthusiasm through vocal and body languages had higher recall of the contents than did children listened to a reader who expressed neutral vocal and body languages (Moè, 2016). In a similar vein, Towler and Dipboye (2001) showed that a trainer who displayed high enthusiasm and expressiveness produced higher recall test scores among trainees more than a trainer who displayed low enthusiasm and expressiveness. In the context of learning videos, an analysis of MOOC videos showed that learners showed more engagement by watching the learning presentation longer and attempting to answer more post-video assessment problems with videos where instructors spoke with higher expression of enthusiasm (Guo *et al.*, 2014). Additionally, a recent experiment demonstrated that learners who learned from the multimedia with an animated pedagogical agent that expressed enthusiasm through facial expression, gestures, and voice performed better in transfer test than learners who learned from the multimedia with an animated pedagogical agent that expressed calm facial expression, gestures, and voice (Liew *et al.*, 2017). Other studies concerning the use of "expressive" voice by robot and agent for education indicated the superiority of "expressive" voice over "flat" voice in terms of cognitive and affective outcomes (Westlund *et al.*, 2017; Fountoukidou *etal.*, 2019).

An underlying concept that explains the facilitating effects of displayed enthusiasm on learning is the *immediacy* principle, which refers to ability of instructors to foster teacher-student psychological closeness (Thomas *et al.*, 1994) Richmond *et al.* (2003). Richmond *et al.* (2003) argued that instructor's nonverbal cues such as gestures, facial expressions, and vocal tone can influence the sense of immediacy. The concept of immediacy is closely related to the enthusiasm, particularly with respect to nonverbal cues of displayed enthusiasm by instructors (Keller *et al.*, 2016). For instance, an instructor that conveys voice, facial expression, and gestures that express enthusiasm may be regarded as friendly, exciting, energetic, and warm — qualities that can induce the sense of psychological closeness in learners. In contrast, an instructor that displays a calm voice, neutral facial expression, and minimal gestures, may be regarded as uncaring, disinterested, and bored. Thus, the feeling of psychological closeness will be impeded. This notion is also related to the "Dr. Fox Effect" which suggests that when instructors express enthusiasm cues, learners will tend to assign favorable ratings toward the instructors, regardless of their teaching qualities (Kunter *et al.*, 2008; Marsh and Ware, 1982).

2.5. Social cues of voice enthusiasm in multimedia learning

According to Collins (1978), instructor's enthusiastic voice possesses vocal qualities that are rapid, varied, emphatic vocal delivery; excited speech, with sudden and considerable changes in tone. In contrast, the neutral human voice is defined as calm, unvaried in terms of pitch, and without enthusiasm (Moè, 2016). It is common for learning videos to use a neutral and calm "documentary-narrative" voice that is devoid of strong emotional expression to deliver lessons.

Relating to the social agency theory, the immediacy cues afforded by a voice conveying enthusiasm or a calm voice can influence the learner's perceived *valence* of these cues. Notably, valence of a social cue can be either positive, which holds desirable qualities such as "appealing", "friendly", and "interesting", or negative, which holds undesirable qualities such as "dislikable" and "distracting", and frustrating" (Domagk, 2010; Johnson *et al.*, 2000; Atkinson *et al.*, 2005; Davis *et al.*, 2019; Veletsianos, 2012; Wouters *et al.*, 2008). While all social cues can prompt social responses; however, the valence of social cues will influence the level of cognitive engagement by learners during the multimedia learning process (Domagk, 2010). In other words, the valence of social cues will affect learning outcome. For instance, Domagk (2010) demonstrated that the pedagogical agent with likeable appearance led to higher transfer performance when compared with the pedagogical agent with dislikeable appearance and with the pedagogical agent with neutral appearance (Exp. 1). Whereas, the pedagogical agent with dislikeable appearance and unappealing voice impeded transfer performance (Exp. 2).

Based on the preceding, it is suggested that an enthusiastic voice can increase immediacy through appealing social cues such as "friendly", "exciting", "energetic", and "warm" to learners, which then results in higher cognitive engagement in users. On the other hand, when a voice is calm and devoid of expressive cues such as enthusiasm, learners may be less motivated to invest cognitive effort given the lack of socially appealing cues. This notion hinges on previous findings regarding the embodiment effect, which demonstrated that the pedagogical agents that expressed minimal gestures and facial expressions led to lower social agency and weaker transfer performance than the pedagogical agents that expressed full natural gestures and facial expressions (Mayer and DaPra, 2012). It can also be interpreted that the non-gesturing agents conveyed social cue that hold *negative* valence, e.g., "less human-like" (Mayer, 2005). Thus, it is also plausible that the calm voice without enthusiasm cues may be attributed with negative social cues such as "uncaring", "disinterested", and "bored". Consequently, due to the perceived weak immediacy factor, a learner may be less willing to invest cognitive effort during the process of multimedia learning, resulting in poor transfer performance. Consistent with the Grice *et al.*'s cooperation principle (1975), it has been conjectured that learners will be committed to make sense of the learning information when they assume that the source speaker is a conversational partner. Thus, it can be argued that the level of cognitive engagement during the multimedia learning process will be also be influenced by the perceived social rating of a speaker that is derived from enthusiasm cues.

This study also notes the possible effects of voice enthusiasm (vs calm voice) on cognitive load. Concerning discernibility, there is no reason to expect any differences of perceived difficulty between the calm and enthusiastic voice, given the fact that both voices will be recorded with a human voice. However, as a calm voice without enthusiasm cues may convey undesirable social cues, it is plausible that these negative cues may cause distractions, and thus impose extraneous load in a learner's mind (Davis *et al.*, 2019; Veletsianos, 2012; Wouters *et al.*, 2008).

3. Experiment 1

The purpose of experiment 1 was to assess the effects of an enthusiastic voice (as compared to a calm voice) on a learner's social rating of the speaker, cognitive load, and learning outcome. A multimedia presentation that delivered lessons on programming algorithms to university undergraduates was used as a platform for the voice narrations. Based on the social agency theory, it was predicted that an enthusiastic voice will lead to higher social agency and better transfer performance than the calm voice. While there was no prediction made with respect to the effects of enthusiastic and calm voice on cognitive load; however, this experiment was also conducted to assess this.

3.1. Method

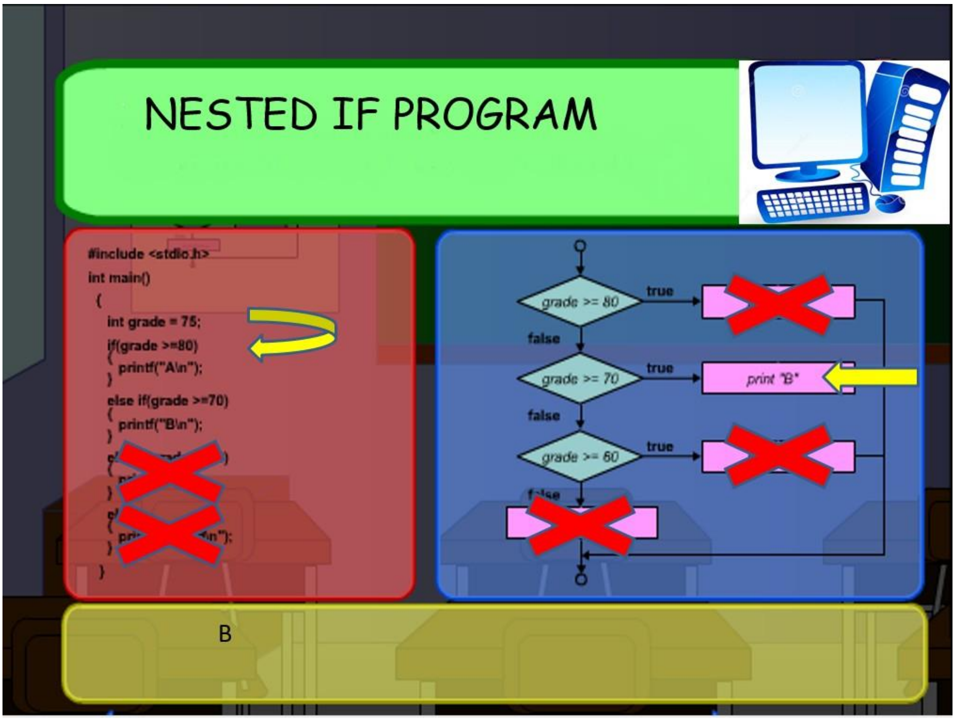
3.1.1. Participants and design

The participants were 76 business major undergraduates who were undertaking a computer-related course in an Asian university (female = 51, male = 25; all aged between 18 and 20). All courses in the university were conducted in English and the entry to the university required results that reflect an intermediate level of English proficiency. Hence, all participants could be assumed to have no difficulty in comprehending English narrations. All participants reported that they had no prior knowledge on programming algorithms. During the experiment, all participants were at the age between 18 and 20. The experiment used a between-group design with 39 participants in the enthusiastic voice group and 37 participants in the calm voice group.

3.1.2. Voice and multimedia learning environment

The authors of this study hired a professional male voice talent to produce two versions of narration based on the same lesson script. For the enthusiastic voice narration, the voice talent was asked to convey a varied, emphatic vocal delivery, excited speech with considerable changes in tone, and large dynamic pitch variation (Collins, 1978; Moè, 2016). Whereas, for the calm voice narration, the voice talent was instructed to convey a pleasant and calm vocal tone, low pitch level, small pitch variations, and expressed no enthusiasm Moè (2016) — the resulting voice narration had a speech style of a newscaster. The enthusiastic voice had an average pitch of 211Hz while the calm pleasant speech had an average pitch of 99Hz. Both voices had speech rate of 138 words per minute. The voices were given post-production treatment to ensure that the voices had appropriate and similar volume levels between them.

The voice narrations were then incorporated into a multimedia learning presentation that was developed using PowerPoint. The multimedia presentation used flowcharts, source code samples, moving arrows, and animated highlights as visuals, which were complemented with the voice narration (either enthusiastic or calm voice)(see Fig.1). The learning outcome of this multimedia presentation was the ability to understand if, if-else, and nested-if algorithms to predict given source-code outputs. Therefore, two versions of multimedia learning presentation were developed — the enthusiastic voice multimedia presentation and the calm voice multimedia presentation. Both of the multimedia learning presentations were program-controlled and system-paced (no user controls other than the start button) and had the duration of about 11 minutes.


NESTED IF PROGRAM

```
#include <stdio.h>
int main()
 {
   int grade = 75;
   if(grade >=80)
   {
     printf("A\n");
   }
   else if(grade >=70)
   {
     printf("B\n");
   }
   el              )
   {
     pr
   }
   el
   {
     pr          \n");
   }
 }
```

```
B
```

FIGURE 1: Multimedia learning environment

3.1.3. *Dependent measures*

Social rating of speaker To assess learner’s social rating of the speaker, the study used a survey with a scale from 1 (strongly disagree) to 7 (strongly agree) which consisted of five items: 1) I like Michael (the speaker); 2) Michael is knowledgeable; 3) Michael is friendly; 4) I can trust Michael; and 5) I am willing to learn with Michael in the future (Cronbach’s alpha = .950). The scores for each item will be totaled and then averaged to produce the overall social rating of the speaker (Liew *et al.*, 2013; Liew and Tan, 2016).

Cognitive load The experiment used the Paas mental effort rating scale ranging from 1 (low) to 9 (high) which asked participants to rate their mental effort used in understand the multimedia lesson (Paas and Merriënboer, 1993).

Transfer test To assess the learning outcome of the participants, the experiment used a near-transfer test which consisted of ten questions. The questions asked participants to correctly predict the output of the given program source codes which represented different algorithms. One mark was awarded for each correct answer; hence, the possible maximum score was ten. The transfer test can be considered as valid and reliable as it was developed by one of the authors of this study who had more than 14 years’ experience in teaching IT and programming concepts.

3.1.4. Procedure

The authors of the study and a research assistant ushered the participants into two adjoining computer laboratories. Each of the computer laboratories had 40 desktop computers with labels that indicated their ordered numbers. To establish randomization, the multimedia presentation with enthusiastic voice narration was installed in each of the computers labeled with odd number, while the multimedia presentation with calm voice narration was installed in each of the computers labeled with even number. Prior to the experiment, the authors of the study and a research assistant ensured that each of the computers had a headphone and volume that was set at 30 percent (optimized for listening comfort).

Once the participants were seated in front of their respective computers, they signed the consent form indicating their agreement to allow their data to be used for research purpose. They were then told to utilize the headphones and checked that the computer volumes were set at 30 percent (optimal volume). The participants were then asked to launch the multimedia presentation by clicking the start button, and were told that they would be tested after the lesson. The 11-minutes multimedia presentations were then viewed and listened by the participants. After the multimedia presentation, participants were given 5 minutes to fill up the survey on social rating of the speaker, and the Paas perceived mental effort (cognitive load). After that, participants were given 15 minutes to answer the transfer test. The session adjourned after the participants were thanked and debriefed. Based on an established scoring rubric, a research assistant scored the transfer tests blind with respect to the conditions.

3.2. Results and discussion

3.2.1. Data analysis

Table 1 shows the means and standard deviations for social rating of the speaker, Paas perceived mental effort, and transfer test scores of each of the conditions. Independent t-tests were performed to compare the scores for each of the dependent measures between calm voice and enthusiastic voice conditions.

TABLE 1
Means and standard deviations of dependent measures between enthusiastic voice group and calm voice group for experiment 1.

	Enthusiastic Voice (n=39) Mean (Standard Deviation)	Calm Voice (n=37) Mean (Standard Deviation)
Social rating of speaker	4.54 (1.53)	3.84 (1.40)
Paas mental effort	4.07 (1.78)	3.83 (1.50)
Transfer test score	6.18 (3.19)	4.38 (3.47)

3.2.2. Does voice enthusiasm affect social rating of the speaker?

Learners who listened to the multimedia presentation narrated by the enthusiastic voice assigned significantly higher social rating of the speaker than did learners who listened to the multimedia presentation narrated by the calm voice, t(74) = 2.07, p = .042 (2-tailed). The effect size for this data using Cohen's d statistic revealed the value of d = .477 which represented a small to medium effect size (Cohen, 2013). Thus, this supported the assumption that an enthusiastic voice can increase the learner's positive social perception regarding the speaker as compared to a calm voice.

3.2.3. Does voice enthusiasm affect perceived mental effort?

The t-test revealed no significant differences for perceived mental effort between the learners in the enthusiastic voice and calm voice conditions, t(74) = .631, p = .53 (2-tailed). Thus, using the Paas mental effort scale, the data illustrates that enthusiastic voice and calm voice did not differently affect the cognitive load of learners.

3.2.4. Does voice enthusiasm affect transfer test score?

Based on the t-test result, it was found that learners who listened to the multimedia presentation narrated by the enthusiastic voice performed significantly better in terms of transfer test score than did learners who listened to the multimedia presentation narrated by the calm voice, t(74) = 2.07, p = 2.354, p = .021 (2-tailed). The calculation of effect size using Cohen's d yielded the value of d = 0.55, which represented a medium to large effect size. In line with social agency theory, this data supported the prediction that a voice that conveys enthusiasm (as compared to a calm voice) can increase social agency and cognitive engagement, which results in better transfer performance.

4. Experiment 2

Experiment 2 aimed to replicate the findings of experiment 1, albeit with some modifications to the dependent measures. First, with respect to the assessment of social rating of the speaker, this experiment used the Agent Persona Instrument (API) which has been validated and extensively used in research examining the persona effect and social rating of voice and pedagogical agent in multimedia learning environment (Ryu and Baylor, 2005; Davis *et al.*, 2019; Craig and Schroeder, 2017; Mayer and DaPra, 2012). The assessment of social rating of speaker via API was more consistent with the current literature; and thus allowed the results to connect more relevantly with prior findings. Second, in line with the suggestion of Davis *et al.* (2019) to use cognitive load measures that separate different type of loads, this experiment used three distinct measures to distinguish between intrinsic load, extraneous load, and germane load.

Based on the social agency theory and supported by findings of Experiment 1, the following predictions were made — 1) the enthusiastic voice should lead to higher social ratings of the speaker than the calm voice, and 2) the enthusiastic voice should lead to better transfer performances than the calm voice. No predictions were made regarding the effects of voice enthusiasm on intrinsic and extraneous load. With regards to germane load, Davis *et al.* (2019) revealed that a human voice with weak prosody prompted higher germane load than a human voice with strong prosody; which was in line with the notion that non-native speakers are less efficient in processing prosodic information associated with the pitch, tempo, stress, intonation, melody, loudness, accent and pause. Given that the participants of this experiment were non-native speakers, it was predicted that the multimedia presentation narrated with a calm voice (weak prosody) will prompt higher germane load than the multimedia presentation narrated with an enthusiastic voice (strong prosody).

4.1. Method

4.1.1. Participants and design

The participants were 52 Information Technology major freshmen who were undertaking a computer-related course in the same university (female = 20, male = 32; all aged between 18 and 20). All of the participants were in their first semester of their freshmen year. Within the context of the educational system here at the time of the experiment, secondary and high schools did not offer programming courses; hence, it could be generally assumed that they had no prior knowledge about programming algorithm. When asked by the experimenters (the authors of this study) prior to the experiment, one participant reported that he had prior programming knowledge through private tuition, and was thus excluded from the data analysis. The rest of the participants confirmed that they had no prior knowledge about programming. One more participant was also excluded from the data analysis as he had mistakenly launched and listened to the wrong multimedia presentation that was irrelevant to the experiment. Hence, the remaining number of participants was 50 — 25 in the enthusiastic voice group and 25 in the calm voice group.

4.1.2. Voice and multimedia learning environment

The voice and multimedia learning environment used in this experiment were the same as Experiment 1.

4.1.3. Dependent measures

Social rating of speaker To assess learner's social rating of the speaker, this experiment adopted the Agent Persona Instrument (API) with 5-points Likert scale consisting of 25 items that assess the four characteristics of the agent (speaker) — facilitating learning, credibility, human-like, and engaging (Ryu and Baylor, 2005).

Cognitive load The intrinsic load survey with 6-points Likert scale asked learners to assign ratings based on the question: "How difficult was the learning content for you?". The extraneous load survey with 6-points Likert scale asked learners to assign ratings based on the question: "How difficult was it for you to learn with the material?" which was based on paper by Kalyuga *et al.* (1998). Adopted from Salomon (1984), the germane load survey with 6-points Likert scale asked learners to assign ratings based on the question:

"How much did you concentrate during learning?". The decision to use these measures was based on an influential study which utilized these same measures to assess intrinsic, extraneous, and germane load (Cierniak *et al.*, 2009)

Transfer test The same transfer test used in Experiment 1 was also utilized in this experiment.

4.1.4. Procedure

The authors of the study and a research assistant ushered the participants into a computer laboratory. Similar to Experiment 1, the computers labeled with odd numbers featured the multimedia presentation with the enthusiastic voice while the computers labeled with even numbers featured the multimedia presentation with the calm voice. Once the participants were randomly seated, they signed the consent form indicating their agreement to allow their data to be utilized for research purpose. They were then asked to listen and view the 11-minutes multimedia presentation. After that, the participants spent 10 minutes to fill up the Agent Persona Instrument and cognitive load surveys, and subsequently did the transfer test for the next 15 minutes. The session adjourned after the participants were thanked and debriefed. Based on an established scoring rubric, a research assistant scored the transfer tests blind with respect to the conditions.

4.2. Results and discussion

4.2.1. Data analysis

Table 2 illustrates the means and standard deviations for API social rating of the speaker, cognitive load — intrinsic and extraneous load, and transfer test scores of each of the conditions. Independent t-tests were performed to compare the scores for each of the dependent measures between calm voice and enthusiastic voice conditions.

TABLE 2
Means and standard deviations of dependent measures between enthusiastic voice group and calm voice group for experiment 2.

	Enthusiastic Voice (n=25) Mean (Standard Deviation)	Calm Voice (n=25) Mean (Standard Deviation)
Facilitating Learning	3.80 (.79)	3.67 (.40)
Credible	4.20 (.62)	4.12 (.51)
Human-like	3.15 (.69)	2.68 (.73)
Engaging	3.49 (.73)	3.12 (.79)
Intrinsic load	2.24 (1.01)	2.68 (1.02)
Extraneous load	2.32 (1.06)	2.40 (1.08)
Germane load	4.08 (1.22)	4.68 (.85)
Transfer Test Score	9.16 (1.84)	8.04 (2.35)

4.2.2. Does voice enthusiasm affect social rating of the speaker?

Based on the t-test results on the four aspects of the Agent Persona Instrument, it was revealed that learners who listened to the multimedia presentation narrated by the enthusiastic voice assigned significantly higher ratings for human-like quality of the speaker than did learners who listened to the multimedia presentation narrated by the calm voice, t(48) = 2.232, p = .024 (2-tailed). Based on Cohen's d statistic, the effect size was d = .66, which represented a medium to large effect size. It was also revealed that learners who listened to the multimedia presentation narrated by the enthusiastic voice assigned significantly higher ratings for engaging quality of the speaker than did learners who listened to the multimedia presentation narrated by the calm voice, t(48) = 1.689, p = .04 (1-tailed). According to Cohen's d statistic, the effect size was d = .48, which represented a small to medium effect size. However, the data indicated no significant differences between the two voice conditions for speaker's qualities with regards to facilitating learning, t(48) = .742, p = .461 (2-tailed) and credibility, t(48) = .44, p = .65 (2-tailed). Taken together, these findings lent support to the prediction that an enthusiastic voice can enhance a learner's social ratings of the speaker as compared to a calm voice, particularly for "human-like" and "engaging" qualities.

4.2.3. Does voice enthusiasm affect intrinsic, extraneous, and germane load?

The t-test results found no significant differences between the voice conditions for intrinsic load, t(48) = 1.524, p = .134 (2-tailed) and extraneous load, t(48) = .263, p = .794 (2-tailed). It was shown that leaners who listened to the multimedia presentation with calm voice reported higher germane load than did leaners who listened to the multimedia presentation with enthusiastic voice, t(48) = 2.013, p = .050 (2-tailed). On the basis of Cohen's d statistic, the effect size was d = 0.57, which represented a medium to large effect size. This result conformed to the argument that for non-native speakers, a weak-prosodic voice (e.g., calm voice) may be more beneficial in terms of germane load than a strong-prosodic voice (e.g., enthusiastic voice) which can be more difficult to process among non-native speakers (Davis *et al.*, 2019) .

4.2.4. Does voice enthusiasm affect transfer test score?

Learners who listened to the multimedia presentation narrated by the enthusiastic voice performed significantly better in terms of transfer test scores than did learners who listened to the multimedia presentation narrated by the calm voice, t(48) = 1.874, p = .03 (1-tailed). On the basis of Cohen's d statistic, the effect size was d = 0.53, which represented a medium to large effect size.

5. General discussion

According to the social agency theory, imbuing multimedia presentation with social cues such as voice, agents, and human images can trigger social responses from learners, which then prompt learners to invest higher cognitive effort during the multimedia learning process. However, extending from the discussion about the mere presence (against the absence) of social cues, recent research has also shown that the characteristics of the social cues can differently influence the level of cognitive efforts by learners (Domagk, 2010; Mayer and DaPra, 2012; Atkinson *et al.*, 2005).

Based on this line of reasoning, this study investigated the enthusiasm effects of a human voice in a multimedia learning environment. Drawing inspiration from the wider literature demonstrating the positive effects of instructors' and pedagogical agents' enthusiasm on affective and cognitive learning (Wang *et al.*, 2019; Moè, 2016; Guo *et al.*, 2014; Liew *et al.*, 2017), the present study examined the voice enthusiasm through the lens of the social agency theory. It was predicted that when compared to a calm voice that expresses no enthusiasm, an enthusiastic voice will increase social agency and thus cognitive engagement, which leads to better transfer performance. This prediction also hinges on the immediacy principle found in literature on enthusiasm, which states that a learner's decision to "approach" or "avoid" the instructor and learning content is based on his/her social interpretation of nonverbal cues such the valence of the voice. The immediacy principle has relevance to social agency theory, insofar that the valence of a social cue (i.e., socially appealing or unappealing cues) can influence the level of learner's cognitive effort. Across two experiments, the data of this study generally supported this assumption; and will be discussed in the following.

Concerning the social rating of the speaker, the results of Experiment 1 showed that the speaker with enthusiastic voice was attributed with higher social ratings than the speaker with calm voice, yielding a small to medium effect size based on Cohen's d statistic (d = .477). The replication of this finding using the Agent Persona Instrument in Experiment 2 demonstrated that the enthusiastic speaker was perceived as more "human-like" (d = .66, medium to large effect size) and "engaging" (d = .48, small to medium effect size) than the calm speaker. However, enthusiastic and calm voices did not differently affect speaker ratings for "facilitating learning" and "credible". When developing the Agent Persona Instrument, Ryu and Baylor (2005) distinguished "informational usefulness" and "affective interaction" as distinct constructs relevant to qualities of a virtual instructor. "Facilitating learning" and "credible" are factors under the informational usefulness construct, and relate to the virtual instructor's skills and knowledge. Whereas, "human-like" and "engaging" are factors under the affective interaction construct, and relate to the emotional expression and communication style of the virtual instructor. This finding aligns with the literature on enthusiasm and immediacy, which states that an enthusiastic speaker will be attributed with desirable social qualities which promote psychological closeness. In contrast, it was found that voice enthusiasm did not influence a learner's perception regarding the speaker's usefulness and helpfulness in terms of contributing toward better learning performance (i.e., informational usefulness — "facilitating learning" and "credible"). Plausibly, this is due to the fact that both enthusiastic and calm voices were recorded using a human voice (same voice talent); and perceived difficulty of learners when discerning the information was not differently affected.

Concerning cognitive load, Experiment 1 revealed that the learners' mental effort ratings did not differ between enthusiastic and calm voice conditions. Experiment 2 found that voice enthusiasm did not affect intrinsic and extraneous load (perceived difficulty). One possible reason is that both of the voices were recorded in a human voice; hence there was no case of poor discernibility which would have imposed extraneous cognitive load. Another possible interpretation can be framed as the following — while the calm voice might have lacked the socially appealing cues as compared to the enthusiastic voice; it did not produce negative social cues (e.g., annoying, irritating, distracting) that would have imposed extraneous cognitive load.

However, it was found that the calm voice prompted higher germane load than the enthusiastic voice (Exp. 2). Given that the enthusiastic voice inherently had strong prosodic cues as compared to the calm voice which inherently had weak prosodic cues, this result supports previous finding demonstrating that non-native speakers who listened to a weak-prosodic human voice in multimedia presentation reported higher germane load than non-native speakers who listened to a strong-prosodic human voice in multimedia presentation Davis *et al.* (2019). Generally, this conforms to the argument put forth by Davis et al. (2019) that non-native speakers are less efficient in processing prosodic cues (pitch, tempo, stress, intonation, melody, loudness, accent and pause).

Similar to prior results (Davis *et al.*, 2019), unexpectedly, the increased germane load did not translate to enhanced transfer performance in this experiment —this result runs counter to the conventional knowledge that germane load will be positively associated with learning performance. In fact, the data of this study showed that learners in the enthusiastic voice group had lower germane load but had higher transfer scores; whereas, learners in the calm voice group had higher germane load but had lower transfer scores. On the basis of this seemingly contradictory result, the use of delayed transfer in further research may bring further clarity to this conundrum, as highlighted by Davis *et al.* (2019) that "researchers should examine voice type and cognitive load with immediate and delayed assessments to evaluate whether increased working memory is more beneficial to the long-term retention of knowledge" (pp.8).

The results from both Experiment 1 and 2 demonstrated positive effects of enthusiastic voice on transfer performance when compared to the calm voice. Experiment 1 which involved business undergraduates as participants (novice learners), showed that enthusiastic voice led to a significantly higher transfer performance as compared to the calm voice, yielding a medium effect size (d = .55). Experiment 2 which involved IT undergraduates as participants (novice learners), replicated this finding by demonstrating that the transfer performances of learners who listened to the enthusiastic voice was superior than those of the learners who listened to the calm voice (d = .53, medium effect to large effect size). This finding can be explained from the social agency theory perspective — the socially appealing cues afforded by the enthusiastic voice can lead learners to engage deeply during the multimedia learning process, thereby producing higher transfer performance.

At this juncture, it is noted that there are other theories that attempt to explain the effects of enthusiasm on learning outcome (Wood, 1998). For instance, Wood (1998) conjectured that enthusiasm may positively affect learning through the increase of selective attention by learners interacting with enthusiastic teachers. That is, enthusiastic cues serve as an attention-getting stimulus that consistently capture the attention of learners during learning, as an enthusiastic stimulus is ever changing in the context of verbal and non-verbal expressions. In contrast, an unenthusiastic teaching style tends to be unchanging and predictable; hence, learners may "tune-out" and stop paying attention to the presented learning content. The present study did not directly test this assumption, albeit the attention-capturing theory of enthusiasm might be relevant in multimedia learning. Nevertheless, it is noted that the data of this study did not find any significant differences between enthusiastic voice and calm voice for the Agent Persona Inventory items of "Facilitating Learning" which are related to the attention engagement — "The speaker kept my attention", "The speaker helped me to concentrate on the presentation", and "The speaker focused me on the relevant information". However, this data should not be taken as direct interpretation that voice enthusiasm did not have any impact on the attentional mechanism of learners, as these items might not have accurately and reliably assessed the attentional process of multimedia learning.

5.1. Implications for instructional design

Based on the current literature, it was shown that enthusiasm cues through body gestures, facial expressions, and voice tones by live-action instructors and pedagogical agents in learning videos can positively affect learners' engagement and learning outcome (Guo *et al.*, 2014; Liew *et al.*, 2017; Wang *et al.*, 2019). The results of this study demonstrated that the positive effects of enthusiasm on learning can also be manifested with disembodied source speaker in a multimedia presentation. In addition to the voice principle which states that instructional designers should favor human voice over machine-synthesized voice, and also choose standard accent over foreign accent; the findings of this study suggest that a human-voice that carries positive emotional valence can also increase social agency that ultimately benefit multimedia learning. Specifically, instructional designers can consider infusing enthusiasm cues into voice narrations; given that enthusiastic voice can enhance social perception and learning performance. This recommendation aligns with the paradigm of voice communication as social communication, rather than purely just a medium of information delivery (Mayer, 2005).

6. Limitations and suggestions for further research

There are some limitations to the results of this study. The first limitation concerned the short duration of the multimedia learning presentation. The results of this study which were obtained with the relatively brief exposure to voice enthusiasm (enthusiastic and calm) might not represent the effects of long-lasting exposure to the voice enthusiasm. It is also possible that learners may feel distracted, annoyed, and weary listening to an enthusiastic voice for a long duration; and thus, giving rise to negative social cues. Future research involving longer duration of the voice enthusiasm can be conducted to clarify this. The second limitation of this study was the small sample sizes of the experiments, particularly in Experiment 2, which affects the statistical power of this study, Future studies can be conducted with larger sample sizes to ensure adequate statistical power.

As utilized in the study by Cierniak *et al.* (2009), the same three one-item measures were used to assess intrinsic, extraneous, and germane load in Experiment 2. Further research may utilize other cognitive load measures that consist of multiple items such as the Leppink scale to distinguish between intrinsic, extraneous, and germane load (Leppink *et al.*, 2013). Moreover, the use of delayed transfer test in future research on voice effects in multimedia learning may clarify the relationship between different types of cognitive load and learning performance related to long-term retention of knowledge (Davis *et al.*, 2019).

7. Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

8. Authors' contributions

All authors were equally involved in the conceptualization, experimentation, data analyses, and interpretation of the findings of this study.

9. Funding

This research was funded by XX.

10. Competing interests

The authors have no competing interests to disclose.

Acknowledgements

This research was supported by XX. The authors would like to thank XX for their assistance in setting up the computer laboratories. The authors thank the editor and anonymous reviewers for their valuable comments. Last but not least, the authors thank the participants for their willingness to participate in the experiments.

References

Ahn, J. (2010), "The Effect of Accents on Cognitive Load and Achievement: The Relationship between Students' Accent Perception and Accented Voice Instructions in Students' Achievement (Doctoral dissertation", .

Akker, E. and Cutler, A. (2003), "Prosodic cues to semantic structure in native and non-native listening", *Bilingualism: Language and Cognition*, Vol. 6 No. 2, pp. 81–96.

Atkinson, R.K., Mayer, R.E. and Merrill, M.M. (2005), "Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice", *Contemporary Educational Psychology*, Vol. 30 No. 1, pp. 117–139.

Bechtold, S.W. (2017), *The Cognitive Theory of Multimedia Learning: The Impact of Social Cues*, 1-14 pp.

Cierniak, G., Scheiter, K. and Gerjets, P. (2009), "Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load?", *Computers in Human Behavior*, Vol. 25 No. 2, pp. 315–324.

Clark, R.C. and Mayer, R.E. (2016), *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*, John Wiley & Sons.

Cohen, J. (2013), *Statistical power analysis for the behavioral sciences*, Routledge.

Collins, M.L. (1978), "Effects of enthusiasm training on preservice elementary teachers", *Journal of Teacher Education*, Vol. 29 No. 1, pp. 53–57.

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

Davis, R.O., Vincent, J. and Park, T.J. (2019), "Reconsidering the Voice Prinicple with Non-native Language Speakers", *Computers & Education*, Vol. 103605.

Domagk, S. (2010), "Do pedagogical agents facilitate learner motivation and learning outcomes?", *Journal of media Psychology*.

Fountoukidou, S., Matzat, U., Ham, J. and Midden, C. (2019), *Effects of a virtual model's pitch and speech rate on affective and cognitive learning*, Springer, Cham, 16-27 pp.

Goh, C.C. (2000).

Grice, H.P., Cole, P. and Morgan, J. (1975), "Logic and conversation", .

Guo, P.J., Kim, J. and Rubin, R. (2014), "How video production affects student engagement: An empirical study of MOOC videos", pp. 41–50.

Johnson, W.L., Rickel, J.W. and Lester, J.C. (2000), "Animated pedagogical agents: Face-to-face interaction in interactive learning environments", *International Journal of Artificial intelligence in education*, Vol. 11 No. 1, pp. 47–78.

Kalyuga, S., Chandler, P. and Sweller, J. (1998), "Levels of expertise and instructional design", *Human*, Vol. 40 No. 1, pp. 1–17.

Keller, M.M., Hoy, A.W., Goetz, T. and Frenzel, A.C. (2016), "Teacher enthusiasm: Reviewing and redefining a complex construct", *Educational Psychology Review*, Vol. 28 No. 4, pp. 743–769.

Kunter, M., Tsai, Y.M., Klusmann, U., Brunner, M., Krauss, S. and Baumert, J. (2008), "Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction", *Learning and Instruction,* Vol. 18 No. 5, pp. 468–482.

Leppink, J., Paas, F., Vleuten, C.P.V.D., Gog, T.V. and Merriënboer, J.J.V. (2013), "Development of an instrument for measuring different types of cognitive load", *Behavior research methods*, Vol. 45 No. 4, pp. 1058–1072.

Liew, T.W. and Tan, S.M. (2016), "Virtual agents with personality: Adaptation of learner-agent personality in a virtual learning environment", pp. 157–162.

Liew, T.W., Tan, S.M. and Jayothisa, C. (2013), "The effects of peer-like and expert-like pedagogical agents on learners' agent perceptions, task-related attitudes, and learning achievement", *Journal of Educational Technology & Society,* Vol. 16 No. 4, pp. 275–286.

Liew, T.W., Zin, N.A.M. and Sahari, N. (2017), "Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment", *Human-centric Computing and Information Sciences,* Vol. 7 No. 1, p. 9.

Linek, S.B., Gerjets, P. and Scheiter, K. (2010), "The speaker/gender effect: does the speaker's gender matter when presenting auditory text in multimedia messages?", *Instructional Science,* Vol. 38 No. 5, pp. 503–521.

Marsh, H.W. and Ware, J.E. (1982), "Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect", *Journal of educational Psychology,* Vol. 74 No. 1, p. 126.

Mayer, R.E. (2005), "Principles of multimedia learning based on social cues: Personalization, voice, and image principles. The Cambridge handbook of multimedia learning", , pp. 201–212.

Mayer, R.E. (2017), "Using multimedia for e-learning", *Journal of Computer Assisted Learning,* Vol. 33 No. 5, pp. 403–423.

Mayer, R.E. and DaPra, C.S. (2012), "An embodiment effect in computer-based learning with animated pedagogical agents", *Journal of Experimental Psychology: Applied,* Vol. 18 No. 3, p. 239.

Mayer, R.E. and Moreno, R. (1998), "A cognitive theory of multimedia learning: Implications for design principles", *Journal of Educational Psychology,* Vol. 91 No. 2, pp. 358–368.

Mayer, R.E. and Pilegard, C. (2005), "Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality principles. The Cambridge handbook of multimedia learning", .

Mayer, R.E., Sobko, K. and Mautone, P.D. (2003), "Social cues in multimedia learning: Role of speaker's voice", *Journal of educational Psychology,* Vol. 95 No. 2, p. 419.

Moè, A. (2016), "Does displayed enthusiasm favour recall, intrinsic motivation and time estimation?", *Cognition and Emotion,* Vol. 30 No. 7, pp. 1361–1369.

Nass, C.I. and Brave, S. (2005), *Wired for speech: How voice activates and advances the human-computer relationship,* MIT press, Cambridge, MA, 9 pp.

Paas, F.G. and Merriënboer, J.J.V. (1993), "The efficiency of instructional conditions: An approach to combine mental effort and performance measures", *Human,* Vol. 35 No. 4, pp. 737–743.

Rey, G.D. and Steib, N. (2013), "The personalization effect in multimedia learning: The influence of dialect", *Computers in Human Behavior,* Vol. 29 No. 5, pp. 2022–2028.

Richmond, V.P., McCroskey, J.C. and Johnson, A.D. (2003), "Development of the nonverbal immediacy scale (NIS): Measures of self-and other-perceived nonverbal immediacy", *Communication Quarterly,* Vol. 51 No. 4, pp. 504–517.

Ryu, J.E.E.H.E.O.N. and Baylor, A.L. (2005), "The psychometric structure of pedagogical agent persona", *Technology Instruction Cognition and Learning,* Vol. 2 No. 4, p. 291.

Salomon, G. (1984), "Television is" easy" and print is" tough": The differential investment of mental effort in learning as a function of perceptions and attributions", *Journal of educational psychology,* Vol. 76 No. 4, pp. 647–647.

Schneider, S., Nebel, S., Pradel, S. and Rey, G.D. (2015), "Introducing the familiarity mechanism: A unified explanatory approach for the personalization effect and the examination of youth slang in multimedia learning", *Computers in Human behavior,* Vol. 43, pp. 129–138.

Sweller, J., Merrienboer, J.J.V. and Paas, F.G. (1998), "Cognitive architecture and instructional design", *Educational psychology review,* Vol. 10 No. 3, pp. 251–296.

Tabbers, H.K., Martens, R.L. and Merriënboer, J.J. (2000), *Multimedia instructions and cognitive load theory: Split-attention and modality effects,* Long Beach, CA.

Thomas, C.E., Richmond, V.P. and McCroskey, J.C. (1994), "The association between immediacy and socio-communicative style", *Communication Research Reports,* Vol. 11 No. 1, pp. 107–114.

Towler, A.J. and Dipboye, R.L. (2001), "Effects of trainer expressiveness, organization, and trainee goal orientation on training outcomes", *Journal of Applied Psychology,* Vol. 86 No. 4, p. 664.

Veletsianos, G. (2012), "How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences", *Computers in Human Behavior,* Vol. 28 No. 1, pp. 275–283.

Wang, Y., Liu, Q., Chen, W., Wang, Q. and Stein, D. (2019), "Effects of instructor's facial expressions on students' learning with video lectures", *British Journal of Educational Technology,* Vol. 50 No. 3, pp. 1381–1395.

Westlund, K., Jacqueline, M., Jeong, S., Park, H.W., Ronfard, S., Adhikari, A. and Breazeal, C.L. (2017), *Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative,* volume 11, 295 pp.

Wood, A.M. (1998), "The Effects of Teacher Enthusiasm on Student Motivation, Selective Attention, and Text Memory. Faculty of Graduate Studies, University of Western Ontario.", .

Wouters, P., Paas, F. and Merriënboer, J.J. (2008), "How to optimize learning from animated models: A review of guidelines based on cognitive load", *Review of Educational Research,* Vol. 78 No. 3, pp. 645–675.