Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect

Article *in* Journal of Educational Computing Research · September 2018

DOI: 10.1177/0735633118802877

2 authors:

Scotty D Craig
Arizona State University
138 PUBLICATIONS **4,483** CITATIONS
SEE PROFILE

Noah L. Schroeder
University of Florida
75 PUBLICATIONS **1,814** CITATIONS
SEE PROFILE

Text to Speech Software and Learning: Investigating the Relevancy of the Voice Effect

Scotty D. Craig[1] *

[1] Arizona State University, Human Systems Engineering, 7271 E. Sonoran Arroyo Mall Mesa, AZ 85212 USA, **(scotty.craig@asu.edu)**

Noah L. Schroeder[2]

[2] Wright State University, College of Education and Human Services, Leadership Studies in Education and Organizations, 442 Allyn Hall, 3640 Colonel, Glenn Hwy, Dayton, OH 45435 USA, (noah.schroeder@wright.edu)

Craig, S. D., & Schroeder, N. L. (2019). Text-to-speech software and learning: Investigating the relevancy of the voice effect. *Journal of Educational Computing Research, 57*(6), 1534-1548.

* Corresponding author:
Scotty D. Craig
7271 E Sonoran Arroyo Mall
Santa Catalina Hall, Ste. 150
Mesa, AZ 85212
Phone: 480-727-1006
Fax: 480-727-1538
Email: Scotty.Craig@asu.edu

Text to Speech Software and Learning: Investigating the Relevancy of the Voice Effect

Abstract:

Technology advances quickly in today's society. This is particularly true in regards to instructional multimedia. One increasingly important aspect of instructional multimedia design is determining the type of voice that will provide the narration; however, research in the area is dated and limited in scope. Using a randomized pretest-posttest design, we examined the efficacy of learning from an instructional animation where narration was provided by either an older text-to-speech engine, a modern text-to-speech engine, or a recorded human voice. In most respects, those who learned from the modern text-to-speech engine were not statistically different in regards to their perceptions, learning outcomes, or cognitive efficiency measures compared to those who learned from the recorded human voice. Our results imply that software technologies may have reached a point where they can credibly and effectively deliver the narration for multimedia learning environments.

Keywords: Voice Effect, synthesized voice, narration, multimedia learning

Text to Speech Software and Learning: Investigating the Relevancy of the Voice Effect

Teachers and instructional designers are increasingly designing multimedia-based instruction for online learning environments. Regardless of the form of online multimedia, whether it is websites, videos, or even smartphone applications, there is a plethora of evidence-based best practices to help guide design decisions depending on a variety of contextual factors, such as the population being taught and the qualities of the learning system. Much of this work has stemmed from the cognitive theory of multimedia learning (Mayer, 2014a), which describes a cognitive model of how people learn with multimedia, and cognitive load theory (Paas & Sweller, 2014; Sweller, 2010), which describes how people learn novel information regardless of teaching modality. Research investigating best practices for designing instruction based on these theories has led to a number of instructional design principles that can be applicable in a variety of situations.

These evidence-based instructional design principles can potentially become the corner stone of many instructors' design decisions as they create instructional materials. However, not all of the design principles have been thoroughly researched in a wide variety of situations, and while the cognitive impacts of many of these principles are well established, the impacts on the learners' perceptions have not been thoroughly researched in many cases. An example of this is the voice effect, or voice principle (Mayer, 2014b). Mayer's (2014b) review describes the voice principle as stating that narration should be provided by "a standard-accented human voice rather than … a machine voice" (p. 358) based on the research literature (median d = .74). Accordingly, a designer would likely avoid using text-to-speech generated voices when designing their instruction. For the purposes of this paper, the terms text-to-speech voice, machine-voice, and computer-generated voices are used interchangeably since they refer to the same genre of technologies.

Critically examining the evidence for the voice effect, one will note a few areas where the scope of the existing literature base is limited (Mayer, 2014b). Notably, the few published studies around the principle are largely dated with two of the three studies having been published in the early to mid- 2000's. The studies found consistent results, having showed that the use of a human voice led to improved learning scores compared to machine-generated voices, and the human voices were perceived as being significantly more favorable compared to the machine-generated voices (Atkinson, Mayer, & Merrill, 2005; Mayer, Sabko, & Mautone, 2003).

These results are easily explained through cognitive load theory (Kalyuga, 2011; Paas & Sweller, 2014; Sweller, 2010) and social agency theory (Atkinson et al., 2005; Mayer et al., 2003). Cognitive load theorists may suggest that the recorded human voice was superior for learning because the machine voice caused extraneous cognitive load (Mayer et al., 2003). Essentially, this explanation hinges on the notion that the computer-generated voice required additional mental effort to comprehend, mental effort that could have been used for learning if the voice was not difficult to understand, which thereby impeded the learning process.

An alternative explanation for the voice effect is social agency theory. Social agency theorists suggest that social cues in a learning environment can encourage the learner to try to

learn the material (Atkinson et al., 2005; Mayer et al., 2003), although recent research has shown that unlikeable social cues can actually impede learning (Domagk, 2010). Based on this evidence, we question if the machine voices used in early studies of the voice effect were not well-liked by the participants. If that were the case, Domagk's findings would suggest that the participants' dislike of the voice could explain why the machine-voice groups did not perform well on learning outcome tests.

As technology has advanced, there has been a dearth of research around the voice effect and the use of text-to-speech software compared to recorded human voices. However, Mayer and DaPra (2012) revisited the voice effect by comparing modern text-to-speech software compared to a recorded human voice. Interestingly, their results showed no significant differences in learning outcomes between groups. A second study by Craig and Schroeder (2017) examined the influence of a virtual human that communicated through a modern computer-generated voice compared or a recorded human voice, and largely found no significant differences between the two. However, the modern computer-generated voice provided benefits on the transfer test compared to other conditions. Based on these recent findings, we question, as Craig and Schroeder did, if the voice effect could presumably be an artifact of available technologies. As noted, most of the studies around the voice effect occurred between 2000 and 2010. The text-to-speech technology available to researchers has drastically changed over this timespan, and computer-generated voices are now common-place. High quality text-to-speech generators can be purchased through the internet for reasonable fees, and in some cases may even be available free of charge.

Hypotheses and Predictions

Due to the increase of accessible and affordable high quality text-to-speech technologies, recent findings which indicate that the voice effect may no longer exist (Craig & Schroeder, 2017; Mayer & DaPra, 2012), and the dearth of recent literature around the voice effect in general, it is necessary to explore whether the voice effect still holds across multiple content domains, learner characteristics, and multimedia formats. The purpose of this study is to begin this exploration through the investigation of three different types of voices, including a text-to-speech generated voice that was used in a prior study in the area (Atkinson, Mayer, & Merrill, 2005), a modern text-to-speech generated voice, and a recorded human voice. While recent studies have investigated different types of voices with the aid of a virtual human (Craig & Schroeder, 2017), in this study we examine the effect in a multimedia animation without a virtual human presence, as presumably their presence could influence the impact of the voice effect. We posit that the voice effect is a byproduct of technological limitations rather than a binding, persistent limitation an instructional designer should be concerned with.

The current study investigated the impact of the voice used for narration within a multimedia learning video on the learner's experience in terms of perceptions, learning, and perceived cognitive load. Based on the literature, three separate hypotheses were tested. The voice effect (Atkinson et al., 2005; Mayer, 2014b; Mayer et al., 2003) would claim that decreased quality and monotone intonation increase extraneous processing (Kalyuga, 2011; Paas & Sweller, 2014; Sweller, 2010), meaning that those learning from computerized voices will

perform worse than those learning from a human voice. Thus, narration by a human would result in higher perceptions, learning, and better mental effort measures than any synthesized voice. However, the quality of synthesized voices has increased and it is possible that the only the quality of the voice matters (Remez, Rubin, Pisoni, & Carrell, 1981). In this case, if the mental effort required understanding the voice quality was reduced, then the learning gains predicted by the voice effect would disappear. Therefore, it could be predicted that learning differences for high quality synthesized voice could be equal to human voice and that both would outperform lower quality synthesized voice for learning. This is not to say that participants cannot tell the difference between the human and synthesized voice, just that it will not significantly impact learning. The third possible hypothesis (null) was that the voice providing the narration does not matter. In this case, it would be predicted that all three conditions would be statistically equal for all measures.

Methods

Participants and Design

This study implemented a randomized pretest-posttest design. Participants (n = 150) were randomly assigned to view a presentation using a classic text-to-speech software (n = 47, same voice software as used in Atkinson et al., 2005), a modern text-to-speech software (n = 53, same voice software as Craig & Schroeder, 2017), or a recorded human voice condition (n = 50).

Participants were users of Amazon.com's Mechanical Turk (MTurk) that voluntarily took part in the study. Study participation was limited to MTurkers who had at participated in at least 50 tasks and maintained a 95% or above Human Intelligence Task (HIT) approval rating to encourage collection of reliable data (see Paolacci & Chandler, 2014). Participants were limited by location to within the United States, because a standard American English voice was implemented for the human voice condition. Participation in the study resulted in $1.00 US as compensation for each participant. In total, 55 male and 95 female MTurkers participated in the study, and the modal age range was 26-34.

Materials

The learning materials, adapted from Moreno and Mayer's (1999) materials on lightning formation, were identical between the three conditions with the only difference being the voice that provided the narration. The basic material presented visual images on the formation of lightning along with 19 statements, which describe the formation. The text of the narration can be found in the Appendix of Moreno and Mayer (1999). The images were a recreation of the images originally implemented in the Moreno and Mayer article. They are available upon request to the corresponding author.

Three different voices were used to present the material. The voices mirrored those used by Craig and Schroeder (2017). The classic text-to-speech software condition used "Mary", the Microsoft speech engine voice as was used in Atkinson et al.'s (2005) study. While understandable to the listener, this voice had a digital quality with clipped or choppy production

and no inflection. A video clip with the voice can be viewed at the following link: https://youtu.be/rZl7N_xPYFw. The modern text-to-speech software used was Neospeech (neospeech.com), and the specific voice used was "Kate". This voice engine, while still computer-generated without inflection or prosody, does not have the synthesized tone and has a smoother voice presentation. A video clip with the voice can be viewed at the following link: https://youtu.be/PSJY1wbnM4I. Finally, the human voice was recorded by a female with an American accent. The human voice was recorded at a similar speed as the computerized voice engines using a HD microphone at 705 kbps. A video clip with the voice can be viewed at the following link: https://youtu.be/9BilX7wzHSI.

Agent Persona Instrument (API). The API consists of 25-items. The instrument is scored using a 5-point Likert scale (1 = strongly disagree, 3 = neutral, 5 = strongly agree) and is designed to measure participants' perceptions of virtual humans and other pedagogical agents (Ryu & Baylor, 2005). However, the instrument may also hold relevance for evaluating a software agent's persona even in the absence of a visual representation (i.e., the agent is not visually present). First, it should be noted that not all software agents are visually represented (Moreno, 2005), and the questions on the API do not specifically refer to the agent's visual representation (see Ryu & Baylor, 2005). In addition, the API's subscales measure participant's perceptions of how well the learning was facilitated (10 questions; Cronbach's $\alpha = .95$), how credible it was (five questions; Cronbach's $\alpha = .90$), its human-likeness (five questions; Cronbach's $\alpha = .93$), and how engaging it was (five questions; Cronbach's $\alpha = .90$), which are all salient constructs for this study. Finally, the instrument has been used in prior work in the area (Craig & Schroeder, 2017), thus allowing for comparisons to previous research of the voice effect. Hence, the instrument was used to evaluate participants' perceptions of the presentation from different types of voice, and responses for the questions corresponding to the four subscales of the API were summed to give a composite score.

Learning Measures. A pretest measure included demographics survey (age, gender) and a general meteorological knowledge test that included a check list of seven items and a self-report rating meteorological knowledge. This pretest measure can be found in Moreno and Mayer (1999). Three different learning outcome measures were used, a multiple-choice test, a retention test, and a transfer test. The multiple-choice test consisted of six questions (Craig, Gholson, & Driscoll, 2002). The retention test consisted of one free recall question and the transfer test consisted of four open-ended questions (Mayer & Moreno, 1998). Two coders scored each answer. The two raters had an inter-rater reliability (Cohen's Kappa) of $\kappa = .68$ for retention and a $\kappa = .62$ transfer questions, which is considered substantial agreement (Cohen, 1960; McHugh, 2012). Disagreements were reconciled with consultation of a third rater.

Mental Effort Scale. The mental effort scale (Paas, 1992) was answered by participants twice during the experiment. The scale is a one-item scale frequently used as a self-reported measure of perceived mental effort. Participants completed the item for the first time immediately after the learning phase of the study (*In studying the preceding video I invested)*. The participants also answered the question after completion of the testing phase (*In solving or studying the preceding problems I invested*). Participants responded to both items using Paas's (1992) 9 point scale ranging from 1 (Very Very low mental effort) to 9 (Very Very High Mental Effort).

Scores on these items were used to calculate both training and testing efficiency scores. The formulas for calculating training and instructional efficiency can be found in Paas, Tuovinen, Tabbers, and van Gerven (2003). In short, training efficiency refers to the mental effort invested into the learning phase in relation to learning performance, while instructional efficiency refers to the mental effort invested into the testing phase in relation to learning performance (Paas et al., 2003).

Procedure

Participants were provided a link to the experimental website (Qualtrics.com) for data collection. Participants received an online informed consent followed by a demographics and pretest survey. They were then randomized into one of the conditions and watched the two-minute video. This was followed by the mental effort question. Then the learning assessments were given in the following order: the multiple choice test (not timed), the retention question (five minutes) and four transfer question presented one at a time for three minutes each. The learning assessments were followed by another mental effort question and the API questionnaire was completed as the final assessment.

Results

Initial tests of variance

Levene's test of Homogeneity of Variances was performed for each dependent measure. These tests indicated that variances were equal among conditions for all measures. So, ANOVA tests were used to assess potential differences.

Perceptions

Participant's ratings of their perceptions indicated differences among the three voice conditions (Table 1). There was a significant difference in participant's ratings on the human-like subscale among the voices, $F(2,87) = 14.08$, $p = .00$; $\eta_p^2 = .25$. As would be expected, the human voice received significantly higher ratings than the classic voice engine ($M_d = 4.42$; $p = .001$; *Cohen d* = 1.41) and the modern voice engine ($M_d = 5.95$; $p = .001$; *Cohen d* = 0.95). There were no significant differences between the computerized voices ($M_d = 1.53$; $p = .19$; *Cohen d* = 0.34). However, the means were in the expected direction with the modern voice engine being rated higher.

The same significant pattern was seen for rating on the subscale that measured engagement, $F(2,87) = 3.56$, $p = .03$; $\eta_p^2 = .08$. The human voice received significantly higher ratings than the classic voice engine ($M_d = 2.69$; $p = .03$; *Cohen d* = 0.63) and the modern voice engine ($M_d = 2.68$; $p = .02$; *Cohen d* = 0.61). There were no differences between the computerized voices ($M_d = .01$; $p = .99$; *Cohen d* =0.00).

There were no significant differences found in the participant's perceptions of how well the voice facilitated learning ($F(2,87) = 2.25$, $p = .11$; $\eta_p^2 = .05$) or its credibility ($F(2,87) = 1.85$, $p = .16$; $\eta_p^2 = .04$).

Table 1.

Means and standard deviations for participants' ratings on the subscales of the Agent Persona Instrument separated by condition.

Voice Condition	*N*	Facilitates Learning M(SD)	Credibility M(SD)	Human-Like M(SD)	Engaging M(SD)
Classic	28	30.14(9.87)	17.29(4.12)	9.29(3.96)	11.64(4.76)
Modern	32	29.91(9.20)	17.28(4.12)	10.81(4.90)	11.66(4.88)
Human	30	34.47(9.15)	19.13(4.66)	15.23(4.42)	14.33(3.78)

Learning Measures

An ANOVA was conducted across the participant's four learning measures (pre-test, retention test, multiple choice test, and transfer test) to determine differences between conditions. All means and standard deviations can be found in Table 2. The ANOVAs performed on the pretest ($F(2,147) = 0.31$, $p = .73$; $\eta_p^2 = .00$), the retention test ($F(2,147) = 0.56$, $p = .57$; $\eta_p^2 = .00$), and the transfer test ($F(2,147) = 0.86$, $p = .98$; $\eta_p^2 = .08$) did not indicate any significant differences among conditions. However, a significant difference was found among conditions for the multiple choice test, $F(2,147) = 6.34$, $p = .002$; $\eta_p^2 = .08$. LSD post hoc tests indicated that participants learning from the classic voice engine were outperformed by participants learning from the modern voice engine ($M_d = .15$; $p = .01$; *Cohen d* = .59) and the human voice ($M_d = .18$; $p = .001$; *Cohen d* = .65). However, there were no significant learning differences observed between participants receiving the presentation with the modern voice engine and the human voice ($M_d = .02$; $p = .64$; *Cohen d* = .10).

Table 2.

Means and standard deviations for participants' scores on multiple choice (proportion correct), retention, and transfer tests separated by condition.

Voice Condition	*N*	Pretest M(SD)	Multiple choice M(SD)	Retention M(SD)	Transfer M(SD)
Classic	47	4.70(2.12)	.48(.28)	1.79(2.24)	1.15(1.49)
Modern	53	4.36(2.12)	.64(.25)	2.21(2.45)	1.06(1.23)
Human	50	4.58(2.42)	.66(.28)	2.26(2.51)	1.16(1.45)

Cognitive Efficiency Measures

A series of ANOVAs were performed on the calculated training and instructional efficiency scores to determine if there were cognitive efficiency differences across conditions. There were significant differences observed for training efficiency with multiple-choice measures, $F(2, 147) = 4.07$, $p = .02$; $\eta_p^2 = .05$. LSD post hoc tests indicated that participants learning from the classic voice engine reported significantly lower efficiency than participants learning from the modern voice engine ($M_d = .36$; $p = .01$; *Cohen d* = .58). Participants in the human voice condition reported better efficiency than the classic voice, but not significantly so ($M_d = .24$; $p = .07$; *Cohen d* = .35). There were no significant learning differences observed between participants receiving the presentation with the modern voice engine and the human voice ($M_d = .01$; $p = .31$; *Cohen d* = .21). The participants' training efficiency scores were not significant among conditions for retention, $F(2, 147) = 1.05$, $p = .35$; $\eta_p^2 = .01$, or transfer tests, $F(2, 147) = 0.50$, $p = .61$; $\eta_p^2 = .007$. Additionally, the ANOVAs performed on each of the instructional efficiency measures were not significant (Multiple choice: $F(2, 88) = 1.16$, $p = .32$; $\eta_p^2 = .03$, Retention: $F(2, 88) = 1.10$, $p = .34$; $\eta_p^2 = .02$, Transfer: $F(2, 88) = 1.90$, $p = .16$; $\eta_p^2 = .04$). Descriptive statistics are provided in Table 3.

Table 3.

Means and standard deviations for participants' training and instructional efficiencies based for each learning measure separated by condition.

		Training Efficiency				Instructional Efficiency		
		Multiple choice	Retention	Transfer		Multiple choice	Retention	Transfer
Voice Condition	**N**	**M(SD)**	**M(SD)**	**M(SD)**	**N**	**M(SD)**	**M(SD)**	**M(SD)**
Classic	47	-.19(0.68)	-.06(0.64)	.02(0.69)	28	.01(0.68)	.15(.58)	.42(0.76)
Modern	53	.17(0.58)	.12(0.70)	.07(0.69)	33	.20(0.56)	.22(.67)	.25(0.69)
Human	50	.04(0.67)	-.05(0.72)	-.06(0.70)	30	-.01(.58)	-.01(.65)	.08(0.57)

Discussion

The analyses of the API subscales showed that human voices were rated significantly higher than the computerized voices in relation to perceived human-likeness and engagement, while no differences were found in regards to the facilitation of learning or credibility. These results are somewhat contradictory to those of Craig and Schroeder (2017). While Craig and Schroeder also found that the human voice outperformed the synthesized voices in regards to perceived human-likeness and engagement, they found that the classic voice was rated worse than the modern or human voices in regards to facilitation of learning and credibility. The findings could differ because in this study there was not a visually represented virtual human within the learning environment. Thus, the results of the two studies taken together extend social agency theory by highlighting how the appearance of virtual humans within a learning environment may change the nature of how certain aspects of a social interaction may be

perceived by the learner, thus adding valuable data to the under-researched area of how learners perceive different sources of audio narration in different contexts.

In terms of learning performance, participants receiving information with a human voice outperformed the older voice engine on the multiple-choice measure; a result which replicates previous findings (Atkinson et al., 2005; Mayer et al., 2003), but only for this one learning measure. However, we note that participants in the modern voice engine condition also outperformed those in the older voice engine condition, and those in the human voice condition did not significantly outperform those in the modern speech engine on any of the learning measures. These results contradict the predictions of the voice effect (Mayer, 2014b), support the findings of recent studies (Craig & Schroeder, 2017; Mayer & DaPra, 2012), and support earlier results showing that the natural variations in human voices may not be needed for comprehension (Remez et al., 1981). Furthermore, while previous evidence had been found in two studies when virtual humans deliver the narration (Craig & Schroeder, 2017; Mayer & DaPra, 2012), this study provides consistent evidence in the context of multimedia animations without the physical presence of a virtual human. If these results replicate across additional contexts, it would indicate that modern text-to-speech voice engines, which often are still missing the inflection and cadence of standard human voices, may have reached a sufficient level of clarity to provide equivalent learning to human voices in multimedia environments.

Consistent with our findings in regards to learning outcomes, the cognitive efficiency measures do not provide support for the voice effect. Participants in the human voice conditions had similar efficiency scores to both voice engine conditions for all cognitive efficiency measures except one. In the one measure where there was a difference, training efficiency for the multiple-choice measure, the human voice condition had significantly worse efficiency scores than the modern voice engine. This provides evidence that the modern voice engine does not appear to require significantly more mental effort in relation to the learning outcomes achieved.

The current study also a replication of a previous finding favoring a high quality synthetic voice, which suggests that high quality synthetic voices may provide similar, or in some cases improved, learning outcome scores compared to recorded human voices. Using the same voice engines and human voice in the current study, Craig and Schroeder (2017) found that a high quality synthetic voice paired with a virtual agent performed better on learning transfer measures than a virtual human using a human voice. While the current effect was observed for recall-based multiple-choice questions and not observed in retention and transfer measures, it does provide further evidence for a synthetic voice effect for improving learning in some cases. While both studies showed equivalent differences in learning between modern voice engines and human voices and improved learning compared to older voice engine comparison, the two studies differed in the type of learning impacted. While this difference deserves additional research, it is possible the difference is due to the presence of a virtual human. A previous study found that a likeable virtual human had a greater influence on transfer outcomes then retention outcomes, and a second experiment showed that a virtual human’s appearance influenced transfer but not retention outcomes compared to a no-agent condition (Domagk, 2010).

Previous research lends additional support toward explaining this synthetic voice effect. Human voice patterns have a naturally changing sequence of linguistic elements (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), which makes them very different from the flat monotone intonation of many computer-generated voices. However, this difference is not required for listener's identification of utterances (Remez et al., 1981). The current findings seem to be in agreement with the Liberman et al.'s and Remez et al.'s results. The data from the current study and the previous Craig and Schroeder study (2017) would seem to indicate that learners were sensitive to the human voice and felt the voice was more engaging. However, as evidenced by the scores on the credibility and facilitated learning scales, the learners did not feel that the higher quality modern voice impacted overall quality in terms of perceived ability to convey information. The lack of learning differences provides further evidence of this.

The evidence presented above does not mean that synthetic voices should not continue to be improved. Addressing the differences between synthetic voices and human voices could further allow for personalization and support further improvements for learning and perceptions of the learning environments. Dialect is an excellent example of this potential. While the intonation of most modern voice engines are often still monotone, dialect of the voice presenting information can have an impact on academic performance when the dialect reflects the student's own dialect (Finkelstein, 2015; Finkelstein, Yarzebinski, Vaughn, Ogan, & Cassell, 2013). In Finkelstein et al's study, the science performance of third grade students was measured after interacting with a "distant peer" technology that employed different dialect use patterns. They found that all native speakers of African American Vernacular English (AAVE) demonstrated the strongest science performance when the technology used AAVE features consistently throughout the interaction. Accordingly, the literature indicates that continued investigation into different types of synthesized voices and their influence on learning and perceptions will continue to be important into the future.

A limitation of the current study is that the API analysis had missing data, with a total of only 90 participants instead of the 150 for the full study. This missing data appeared to be random in relation to condition. The most likely cause was that the API was presented at the end of the study and could easily be skipped, unlike the learning measures that had a timer. Additionally, the API with 25 questions was a fairly long test. This finding could point toward the need for additional research on the measure to determine if a shorter test could be constructed. Future studies using the instrument should consider instrument placement within the experimental procedures for optimal data collection.

In addition, the current study used Moreno and Mayer's (1998) narrative on lightning formation. This resulted in a presentation was a little over two minutes in length. While these short duration videos are common within the multimedia learning research area (Mayer, 2009), it is not known if the current results would hold for longer or more complicated material. However, the current findings are promising since e-learning environments such as MOOCS tend to use short video clips, or mini-lectures, instead of longer video (Scagnoli, McKinney, & Moore-Reynen, 2015) with many of online lectures being only minutes long (Crook & Schofield, 2017).

In online learning courses, when longer videos were provided students only engaged with them for six minutes regardless of the length of the video (Guo, Kim, & Rubin, 2014).

Participants in the current study were recruited using Amazon MTurk. The question here is whether the potential risk from the participant's reduced attentional effort from the lack of experimental control is a detriment to the study. Previous work has shown limited differences between MTurk populations and comparison populations on performance (Casler, Bickel, & Hackett, 2013; Mason & Suri, 2011), so this population does not pose a likely threat to the current findings. Within the Amazon MTurk population, the risk of decreased attention can be lessened by selecting reliable participants that have a 95% or above Human Intelligence Task (HIT) approval rating (Paolacci & Chandler, 2014). Further, Hauser and Schwarz (2015) found that the Amazon MTurk participants were more attentive to instructions when attentiveness to instructions was compared between Amazon MTurk and subject pool populations. Thus, the potential risk from the lack of experimental control in the MTurk population is not viewed as a plausible threat to the current findings. Because participants had control of their own learning environment, it seems plausible that the environment could be similar to what an online learner may experience. Hence, the limited control over the MTurker's learning environment may not be a detriment, but rather a benefit as it could more closely replicate an online learning situation than a laboratory-based study may be able to accomplish.

Conclusion

The voice effect suggests that using recorded human voices to provide narration in multimedia learning environments will provide better learning outcomes than using computer-generated voices (Mayer, 2014b). However, for those who design learning content, finding a content expert confident enough to discuss the topics and willing to record numerous narratives for inclusion in learning environments can be challenging, hence highlighting the benefits of a text-to-speech tool if they are consistently as effective as recorded human voices.

Using a randomized pretest-posttest design, it was found that, by and large, there were minimal differences in the ways that participants perceived and learned from a modern computer-generated voice compared to a recorded human voice. As expected, those learning with the recorded human voice perceived it to be more engaging and human-like than the machine generated voices, but there were no significant differences in the participants' perceptions of how well the voices facilitated learning or how credible they were. Our analysis of the learning and cognitive efficiency measures showed mixed support for the voice effect, with most results not providing significant support for the effect. However, it should be noted that the current study was in a non-interactive multimedia environment. It is possible that these results might not replicate in interactive environments where there could be higher expectations of responsiveness to the dynamic interaction. Additionally, this study was performed with one example of a modern text-to-speech voice engine and human voice. Hence, additional studies are required to determine the generalizability of these findings. However, if this pattern replicates in different contexts, it will show that voice engines have reached an acceptable level of performance for use within learning technologies. This finding could result in the creation of

more dynamic and less expensive learning technologies, as currently research would suggest that any narrative should be a recorded human voice.

References

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30*, 117-139.Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156-2160.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46

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

Crook, C. & Schofield, L. (2017). The video lecture. *The Internet and Higher Education, 34*, 56-64.

Domagk, S. (2010). Do pedagogical agents facilitate learning motivation and learning outcomes? *Journal of Media Psychology, 22*(20), 84-97.

Finkelstein, S. (2015, June). Educational Technologies to Support Linguistically Diverse Students, and the Challenges of Classroom Integration. In International Conference on Artificial Intelligence in Education (pp. 836-839). Springer.

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013, July). The effects of culturally congruent educational technologies on student achievement. In International Conference on Artificial Intelligence in Education (pp. 493-502). Springer Berlin Heidelberg.

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. Proceedings of the First ACM Conference on Learning @ Scale Conference (pp. 41–50). New York, NY, USA: ACM.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48(1)*, 400-407.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*(1), 1-19.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review, 74(6)*, 431-461.

Mayer, R. E. (2009). Multimedia Learning (2nd ed.). New York: Cambridge University Press.

Mayer, R. E. (2014a). *The Cambridge handbook of multimedia learning (2nd ed.)*. New York: Cambridge University Press.

Mayer, R. E. (2014b). Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In. R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning (2nd ed.)*(pp. 345-368.). New York, NY: 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), 239-253.

Mayer, R. E., Sabko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*(2), 419-425.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*¸ 276-282.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods, 44(1), 1-23.

Moreno, R. (2005). Multimedia learning with animated pedagogical agents. In R. E. Mayer (Ed.). *The Cambridge Handbook of Multimedia Learning* (pp. 507-523). New York, NY: Cambridge University Press.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91(2),* 358-368.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429-434.

Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning (2nd ed.)*(pp. 27-42.). New York, NY: Cambridge University Press.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science, 23*, 184–188.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212(4497),* 947-949.

Ryu, J., & Baylor, A. L. (2005). The psychometric structure of pedagogical agent persona. *Technology Instruction Cognition and Learning, 2(4),* 291-314.

Scagnoli, N. I., McKinney, A., & Moore-Reynen, J. (2015). Video lectures in eLearning. In F. M. Nafukho & B. J. Irby (Eds.), Handbook of Research on Innovative Technology Integration in Higher Education (pp.115-134). Hershey, PA: IGI Global.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, *22*(2), 123-138.