The Effects of an Embodied Pedagogical Agent's Synthetic Speech Accent on Learning Outcomes

Tiffany D. Do
University of Central Florida
Orlando, FL, USA
tiffanydo@knights.ucf.edu

Mamtaj Akter
University of Central Florida
Orlando, FL, USA
mamtaj.akter@knights.ucf.edu

Zubin Choudhary
University of Central Florida
Orlando, FL, USA
zubinchoudhary@knights.ucf.edu

Roger Azevedo
University of Central Florida
Orlando, FL, USA
roger.azevedo@ucf.edu

Ryan P. McMahan
University of Central Florida
Orlando, FL, USA
rpm@ucf.edu

ABSTRACT

Modern text-to-speech engines can be an effective speech choice for embodied virtual pedagogical agents. However, it is not known how synthesized accents influence learning outcomes and perceptions of the agent. In this paper, we conducted a between-subjects experiment (n=60) to determine the effect of a pedagogical agent's machine synthesized text-to-speech accent (United States English or Indian English) on learning outcomes and perceptions of the agent for students in the United States. Our results indicate that learner gender interacts with synthesized speech accent to significantly affect learning outcomes and perceptions of the agent. Our results reveal that a foreign synthetic speech accent may affect the learning outcomes of female university students (n=30), but not male university students (n=30). Finally, our results indicate that learner gender interacts with synthesized speech accent to affect perceptions of the pedagogical agent's human-likeness. We provide novel insights on the differences between male and female learners for interactions with pedagogical agents with synthetic TTS accents.

CCS CONCEPTS

• **Applied computing → Computer-assisted instruction**; **E-learning**; • **Human-centered computing → User studies**.

KEYWORDS

pedagogical agents; synthetic speech; speech accent

ACM Reference Format:
Tiffany D. Do, Mamtaj Akter, Zubin Choudhary, Roger Azevedo, and Ryan P. McMahan. 2022. The Effects of an Embodied Pedagogical Agent's Synthetic Speech Accent on Learning Outcomes. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556587

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3536221.3556587

1 INTRODUCTION

Pedagogical agents refer to anthropomorphic characters with a virtual presence in the learning environment that are designed to facilitate learning [18, 39]. Pedagogical agents in the form of embodied virtual humans can be beneficial for learning [31, 33, 39, 56] and have become more prevalent in the classroom and virtual platforms due to advances in computer hardware and accessibility [18]. Prior work indicates that the design (e.g., speech, appearance, gender) of a pedagogical agent can influence agent perception and student learning outcomes [35, 40].

An important design aspect for virtual agents is speech [43]. The use of machine synthesized text-to-speech (TTS) has seen interest in the education field (e.g., [6, 16, 18]) since it can be quicker and more convenient to record than using human voice actors. Yet, synthetic speech has been avoided when learning from a virtual human due to its negative effects on learning outcomes and perception [6, 43]. However, recent work found that this effect should be reconsidered due to the substantial advancements in synthetic speech technology, and suggests that modern synthetic speech engines are now as effective as human speech for virtual pedagogical agents [16, 18].

Recent advancements in TTS have also allowed for the advent of realistic regional synthetic accents (e.g., United States English, United Kingdom English, India English). However, these advancements are still limited, as not all regional accents are available (e.g., Mexico English), especially in commercial products like Microsoft Azure. While previous work has investigated the *quality* of a pedagogical agent's synthetic speech [18, 40], there is little work that investigates the *accent* of an agent's synthetic speech, which can prime social cues and affect learning outcomes and instructor perception. For example, prior work reported that human speech with non-native accents negatively affect both learning outcomes and instructor perception [29, 43]. This effect has been observed in multiple regions, such as Sweden [11] and the United Kingdom [64]. According to the "voice effect" principle [40], a student will prefer an instructor with a familiar accent and learn more deeply. However, it is not clear whether the voice principle applies to pedagogical agents with TTS, which can be perceived differently from human voices [22].

Prior work suggests that these effects may be attributed to the cognitive load theory [43, 51]. Students must allocate fewer cognitive resources to process a familiar accent, which thereby enhances their learning experience, compared to a foreign accent. Additionally, prior work [43] posited that learning is an inherently social activity (i.e. social agency theory), and that a socially appealing voice would promote deep cognitive learning. Since user gender can influence social cues, we also investigated disaggregated data by user gender, unlike previous studies investigating speech accents. For example, previous work found that women may be more sensitive to an agent's personal attributes (e.g., age, ethnicity, gender) [36].

As online education platforms grow in popularity around the world [20], understanding the effects of TTS accents can help developers decide whether additional effort should be taken to align a pedagogical agent's speech accent to a learner's demographic to foster learning. Current commercially available TTS engines do not represent all regional accents and provide limited options in terms of accents. Hence, for some regional accents, custom TTS models would need to be developed in order to align an agent's speech accent to the learner's regional representation. Thus, it is important to understand how synthesized speech accents can positively or negatively affect students. Our work sheds light on the effect of TTS accents and provides insight for developers. We investigate the following research questions:

- **RQ1:** How does the TTS accent of a pedagogical agent affect learning outcomes?
- **RQ2:** How does the TTS accent of a pedagogical agent affect learners' perception of an agent?

To answer these questions, we used an established multimedia lesson regarding lightning formation adapted from Mayer and Moreno [41]. Through a between-subjects study (n=60), we investigated the effect of a pedagogical agent's TTS speech accent (US English or Indian English) on learning outcomes and perception. In our experiment, participants watched a multimedia video lesson and completed questionnaires that measured their learning outcomes and perception of the agent. This paper makes two primary contributions:

(1) We provide the first user study investigating the voice effect principle using machine synthesized accents. Our results indicate an interaction effect between user gender and synthetic accent on learning recall.
(2) We show that synthesized accent may interact with user gender to affect the perception of human-likeness of an embodied pedagogical agent.

2 RELATED WORK

2.1 Pedagogical Agents as Social Actors

Design aspects such as voice, appearance, and gender can influence the perception and effect of virtual humans and agents (e.g., [9, 10, 18, 30, 42]). Many of these design aspects are social, which could help explain their influence on learning outcomes and agent perception. A large body of work indicates that human interactions with computers are social. Early work by Nass et al. [49] introduced the CASA (Computers Are Social Actors) paradigm, which posited that human interactions with computers are fundamentally social and that people apply social rules, norms, and expectations to computers.

Nass and Brave [50] reported that people accept computers as social partners and that social cues prime social responses in learners. Moreno et al. [46] furthered this body of work in the field of education by introducing social agency theory, which refers to the idea that social cues in multimedia instructional messages can prime a social response in learning and influence cognitive processing and learning outcomes. Moreno et al. indicated that pedagogical agents can promote meaningful learning in multimedia lessons due to social agency. Domagk [23] expanded on the social agency theory by investigating the valence of social cues. Domagk reported that pedagogical agents with appealing social cues can promote increased transfer performance, but also that unappealing social cues might even hinder learning.

Martha and Santoso [39] provided an excellent review on the impact of design principles for pedagogical agents and suggested that good design elements can make students more involved in learning. Likewise, Lester et al. [38] argued that animated pedagogical agents have persona effects and described how the social characteristics of agents influence how much people enjoy interacting with them.

2.2 Effects of an Instructor's Speech Accent

The effects of an instructor's speech accent on learning outcomes and perception is a well-studied topic. Early work by Gill [29] argued that different accents yield different perceptions of teachers. Gill also noted that teachers with similar accents to the student are perceived more favorably (e.g. more intelligent and dynamic). Similarly, Ahn and Moore [1] found that US college students preferred instructors with American accents. Both Gill [29] and Ahn and Moore [1] found that US college students preferred American accents over European accents, and that European accents were preferred over Asian accents. Kang and Wilson [34] analyzed a substantial amount of work (e.g., [5, 27, 60]) and noted that US undergraduates have an inherent, and usually subconscious prejudice against international teaching assistants with accented English. Although many of these studies are based on students in the US, this phenomenon has also been observed across other regions and languages (e.g., [11]).

Foreign accents can also influence learning outcomes in addition to instructor perception. An early study by Gill [29] reported that British-accented and Malaysian-accented English impaired learning outcomes for US students. Mayer et al. [43] also reported that Russian-accented human speech had significantly lower learning outcomes for US students, possibly due to the cognitive load theory. Further supporting this theory, Munro et al. [48] found that US students took more time to evaluate Mandarin-accented utterances compared to utterances of native English speakers, which may influence meaningful learning. Foreign accents can also sometimes be helpful for pedagogical agents. For example, Galluccio [28] found that an animated pedagogical agent with Spanish-accented English increased motivation for learning in a Spanish language class. This can likely be attributed to the topic as foreign language learners typically prefer instructors to have accents that reflect the language being taught [64]. Although these studies found that students typically learn better from instructors with familiar accents, it is not clear whether these learning effects apply to pedagogical agents with TTS accents, which our study investigates.

Table 1: Overview of previous studies investigating speech accents.

	Platform	Synthesized Accents	Pedagogical	Agent Type
Dahlback et al. [19]	Audio Recording	×	×	Voice
Sandygulova & O'Hare [54]	Physical Robot	×	×	Humanoid Robot
Ahn et al. [1]	Web Browser	×	✓	Voice
Cao Ngoc [14]	Web Browser	×	✓	Voice
Galluccio [28]	Desktop	×	✓	Virtual Human
Gill [29]	Audio Recording	×	✓	Voice
Mayer et al. [43]	Audio Recording	×	✓	Voice
McCrocklin et al. [44]	Desktop	×	✓	Voice/Static Photo
Chan et al. [15]	Desktop	×	✓	Voice
Baird et al. [8]	Web Browser	✓	×	Voice
Krenn et al. [37]	Desktop	✓	×	Voice
McGinn & Torre [45]	Desktop	✓	×	Voice/Static Image
Tamagawa et al. [61]	Audio Recording	✓	×	Voice
Ours	Web Browser	✓	✓	Virtual Human

2.3 Effects of Agent Speech Type

Qualities of an agent's speech can affect learning outcomes and instructor perception [18, 39, 40]. Early research found that TTS negatively affected learning outcomes in animated pedagogical agents [6, 43]. However, more recent work indicates that synthetic speech has considerably improved and may be as effective as human speech for an animated pedagogical agent [16, 18]. We extend upon these prior works by replicating their robust methodology in the context of accented TTS. We investigate accented TTS due to the effect of possible differences between human speech and synthetic speech, such as differences in trust [22] and prosody [13].

Seaborn et al. [58] provided a thorough review on the effects of agent voice on a wide scope of human-agent interactions. Seaborn et al. reported that in many studies, agents with accents that match the user's were more positively perceived. This effect was consistent throughout different regions, such as New Zealand [61] and Ireland [54]. Additionally, some studies involving human-robot interactions have investigated the use of machine-synthesized accents in robots and found that synthesized accents influence social stereotypes [45, 54, 61]. Perhaps most relevantly, Tamagawa et al. [61] found that listeners in New Zealand preferred a health care robot with a synthesized New Zealand accent over other synthesized accents.

These results can possibly be explained by the similarity-attraction principle [12], which posits that individuals are attracted to other people that are similar to themselves. Although we investigated a pedagogical agent within a multimedia learning environment instead of robots, we expected that students would apply social cues to interactions with the agent according to the social agency theory [46] and prefer accents that are similar to their own. The key difference of our work from these studies is that we are interested in investigating the voice principle effect (i.e., foreign accents affect learning) using synthesized accented voices. Table 1 shows a brief overview of prior studies that also investigated agent accents.

3 METHODS

We conducted a between-subjects experimental study to evaluate how a pedagogical agent's speech accent affects learning outcomes and learners' perception of the agent. We examined two synthetic accents (US English or Indian English). We drew from the methods of Chiou et al. [16], who adapted material from early work by Mayer and Moreno [41] into the context of animated pedagogical agents. We used the same script for the multimedia lesson, which taught users about lightning formation. All participants were classified as having low experience in meteorology on the basis of a pretest questionnaire.

3.1 Research Hypotheses

We proposed the following hypotheses regarding the synthetic speech accent of a virtual pedagogical agent:

- **H1 (Learning outcomes):** Participants (from the US) will have higher scores of Recall, Retention, and Transfer for the agent with a US accent compared to the agent with an Indian accent, considering previous results indicating that a foreign accent negatively affects learning outcomes [43].
- **H2 (Agent perception):** Participants (from the US) will rate the agent with a US accent more favorably, considering previous results indicating that learners prefer teachers with a familiar accent [1, 11, 29].

3.2 Dependent Variables

3.2.1 *Learning outcomes.*

We measured learning outcomes using recall, retention, and transfer tests. We measured a participant's information Recall using a six-item multiple choice questionnaire developed by Craig et al. [17]. This questionnaire contained four deep level conceptual knowledge questions (e.g., "Why does it get colder right before it rains?") and two shallow conceptual knowledge questions (e.g., "What part of the cloud are the positively charged particles located in?"). This questionnaire can be found in Appendix A.

Our retention and transfer tests were the same as previous studies using the same learning content (c.f. [18, 41, 47]. Two raters scored each Transfer and Retention test without knowledge of the participant's condition. A third rater reconciled any disagreements. We measured a participant's Retention using a one-question

Retention test that asked participants to "Please write down an explanation of how lightning works." Scores were computed by counting the number of major idea units present in the participant's answer based on a scoring system presented by Mayer and Moreno [41]. According to this scoring system, each idea unit was awarded one point, regardless of wording. The maximum possible score was 19.

The Transfer test consisted of the following four open-ended questions: "What could you do to decrease the intensity of the lightning?", "Suppose you see clouds in the sky, but no lightning. Why not?", "What does air temperature have to do with lightning?", and "What causes lightning?". This test was scored by counting the number of acceptable answers across the four questions according to the rubric described by Mayer and Moreno [41]. The maximum possible score was 12 [41].

3.2.2 Agent Perception. We measured learners' perception of the agent using the Revised Agent Persona Inventory (API-R) [57], which is a 25-item five-point Likert scale questionnaire measuring a participant's perception of a pedagogical agent. The API-R questionnaire measures four constructs: the agent's ability to facilitate learning (Facilitates Learning), perceptions of credibility of the agent (Credibility), the level to which the agent is perceived as human-like (Human-like), and how engaging the agent was during the presentation (Engaging). The Facilitates Learning construct combines ten questions, while the other three combine five questions.

3.3 Learning Materials

3.3.1 Multimedia presentation. Participants listened to a system-paced video narrated by a virtual pedagogical agent. The agent presented a series of images describing the formation of lightning, which were adapted from the images used in [16]. The instructional video lasted around 140 seconds and described the formation of lightning using a 600-word passage (16 different sentences) paired with 14 images. These sentences can be found in the Appendix of Moreno and Mayer's [47] work. The only difference between the videos was the agent's speech accent (US or Indian). A screenshot of the multimedia lesson is shown in Figure 1.

3.3.2 Virtual agents. The agents' voices were created using Microsoft Azure's TTS engine. The US English condition used the "English (United States) - Jenny (Neural)" voice, while the Indian English condition used "English (Indian) - Neerja (Neural)" voice. Based on previous studies, we used the character "Susan" from the software "Media Semantics Character Builder" to create and animate our agent [16, 18]. We chose to maintain the same agent appearance for both conditions since Rakic et al. [53] reported that accent is predominant over appearance for ethnic categorization. Furthermore, Galluccio [28] found that a mismatched ethnic appearance and accent had no effect on learning outcomes or perception, while accent alone did. Hence, we chose to maintain the same agent appearance despite manipulating the agent's speech accent.

3.4 Procedure

The following procedure was reviewed and approved by our university Institutional Review Board (IRB). Due to the COVID-19 pandemic, the study was held fully online. The study consisted of one online Qualtrics survey that lasted approximately 35 minutes. Participants completed a consent form and a demographics survey that captured their self-reported age, gender, native language, and education level. Afterwards, they completed the pretest regarding their meteorological knowledge. In order to include only students with low meteorological experience, we excluded the data of any participants who scored above a 6 [47].

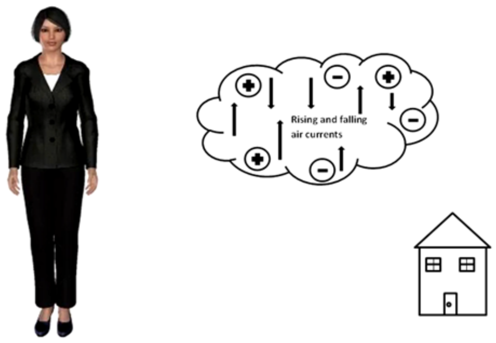

Figure 1: An image of the virtual human agent used in our multimedia lesson.

Participants then completed a sound check, which instructed them to listen to an audio code that was required for them to move forward. Once this code was entered, participants were then randomly assigned to one of the two speech accent conditions and were presented with instructions to please pay attention as the information is presented. Condition assignment was controlled by gender such that genders were evenly distributed across conditions. Next, participants watched the corresponding multimedia video lesson. After finishing the video, the participants completed the multiple choice Recall test (5 minutes), the free response Retention test (5 minutes), the free response Transfer test (3 minutes per question), and the API-R measure. Participants were given a code for a $5 Amazon gift card at the conclusion of the experiment.

3.5 Participants

A total of 60 participants (30 female, 30 male) were recruited to take part in the study after all exclusions and pre-screening procedures. All participants were current university students that were born in and currently reside in the United States. Participants were recruited through university listservs and reported no visual, audio, or neurological/learning disabilities. Additionally, all participants reported proficiency in the English language, with 95% reporting English as their first language. We excluded participants that had experience with an Indian language. Male participants had a mean age of 20.40, within a range of 18 to 24. Female participants had a mean age of 20.57, within a range of 19 to 24.

4 RESULTS

We used non-parametric tests to investigate our hypotheses since our dependent variables were not normally distributed. We disaggregated data by gender according to best practices [62] as well as possible influences of user gender on perception of pedagogical agents [36]. Since we anticipated a possible interaction effect between learner gender and condition, we used the Aligned Rank Transform (ART) tool to perform non-parametric 2 x 2 factorial analyses [63] and post-hoc contrast (ART-C) tests [25].

4.1 Learning Outcomes

The means and standard deviations for all learning outcome measures can be found in Table 2. We first analyzed pretest scores to determine that there was no significant difference between speech conditions, $F(1,56) = 1.68, p = 0.20$ or participant gender, $F(1,56) = 0.319, p = 0.57$.

We found a significant interaction effect between learner gender and agent accent for Recall scores, $F(1,56) = 4.11$, $p = 0.047$, $\eta^2 = 0.07$. A post-hoc test (ART-C) found that female learners had significantly higher Recall scores when they were taught with the US accent condition compared to the Indian accent condition, $t(29) = 3.37$, $p < 0.01$. On the other hand, male learners had no significant difference between conditions, $t(29) = 0.21$, $p = 0.97$. We found no significant differences between any other pairwise comparisons. Figure 2 shows the interaction plot of Recall scores between male and female learners. H1 (Learning outcomes) was partly supported because female learners had significantly higher Recall scores when taught by the agent with a US accent.

Table 2: Means and standard deviations of all learning outcome scores separated by learner gender (M or F) and agent accent condition (US or IN).

	Pretest M (SD)	Recall M (SD)	Retention M (SD)	Transfer M (SD)
M-US	3.53 (1.60)	0.59 (0.28)	4.27 (2.92)	2.60 (1.77)
M-IN	2.80 (1.61)	0.56 (0.23)	4.60 (2.32)	2.67 (1.23)
F-US	3.33 (1.45)	0.62 (0.17)	4.53 (2.07)	2.53 (1.36)
F-IN	2.93 (0.96)	0.36 (0.18)	4.60 (3.56)	2.80 (1.90)

Upon collapsing gender, we also found a significant main effect of agent accent on Recall scores, $F(1,56) = 5.92$, $p = 0.02$, $\eta^2 = 0.10$. A post-hoc (ART-C) test revealed that the US accent condition had significantly higher Recall scores, $t(56) = 2.43$, $p = 0.02$. We did not find any significant interaction effects or main effects on Pretest, Retention or Transfer scores.

4.2 Agent Perception

All means and standard deviations of API-R constructs can be found in Table 3. We analyzed construct scores from the API-R (Facilitates Learning, Credibility, Human-like, and Engaging). We found a significant interaction effect between learner gender and agent accent for the Human-like construct, $F(1,56) = 8.14$, $p < 0.01$, $\eta^2 = 0.13$. A post-hoc (ART-C) test found no significant pairwise comparison differences between groups. However, we found a trend where female learners rated the agent with the US accent as marginally more Human-like than the agent with the Indian accent, $t(29) = 2.55$, $p = 0.08$. Figure 3 shows the interaction plot of Human-like ratings between male and female learners. H2 (Agent perception) was not supported, since learners did not rate the agent with a US accent more favorably than the agent with an Indian accent.

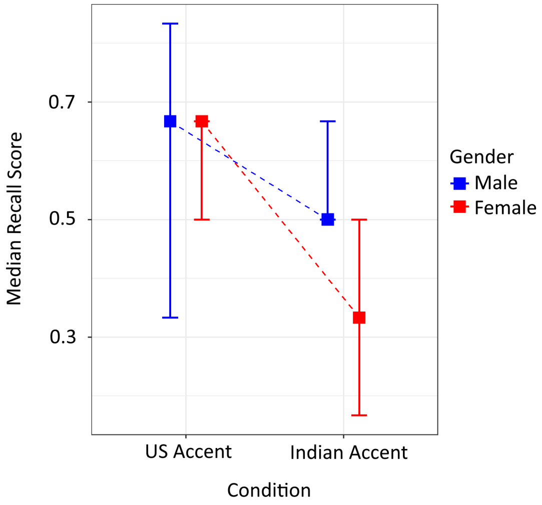

Figure 2: Recall median scores between male and female learners with 95% confidence interval of the median.

Table 3: Means and standard deviations of all agent persona index construct scores separated by learner gender (M or F) and agent accent (US or IN).

	Facilitates Learning M (SD)	Credible M (SD)	Human-like M (SD)	Engaging M (SD)
M-US	30.67 (11.10)	16.80 (4.81)	13.87 (5.13)	15.40 (5.44)
M-IN	27.47 (7.56)	14.73 (3.70)	10.87 (5.08)	13.07 (4.86)
F-US	28.32 (11.41)	16.47 (5.96)	9.60 (4.00)	12.00 (5.11)
F-IN	32.73 (7.75)	18.27 (3.22)	14.00 (4.83)	13.73 (5.18)

5 DISCUSSION

In this section, we first discuss implications for research and implications for pedagogical agent design based on our results. We conclude the section by discussing the limitations of our work and paths for the future.

5.1 Synthetic Accent's Effect on Recall May Be Affected By User Gender

Our results indicate that effect of an agent's synthetic accent on recall may be affected by the learner's gender. We found a significant interaction effect between learner gender and agent accent on learning recall. Our results indicate that the recall of female learners may be negatively affected by the foreign synthesized accents, while male learners may not be affected. These results differ from those of Mayer et al. [43], who reported a significant main effect of foreign accents on learning outcomes of university students using the same multimodal multimedia lesson.

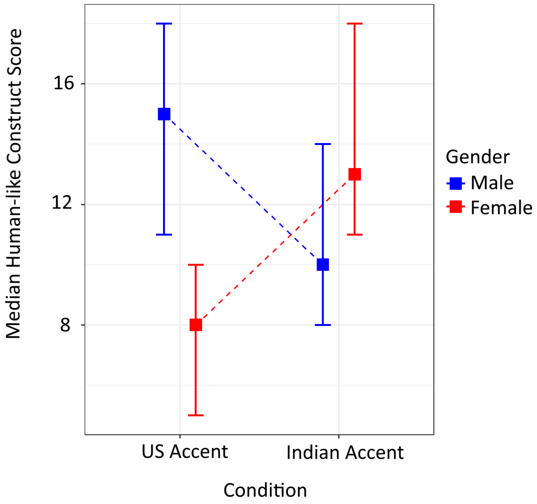

Figure 3: Human-like construct median scores between male and female learners with 95% confidence interval of the median.

There may be several reasons behind these differences. For instance, since men and women may perceive synthetic voices differently [22], these differences may be due to the use of synthetic speech, while Mayer et al. used recorded human speech. The effect of synthetic accents may be different from the effect of recorded human speech accents. Additionally, unlike our current study, Mayer et al. did not gender balance across conditions, and the majority of their participants were female (65%). These differences may be attributed to the distinct effect of accented speech on female learners, rather than an effect on all learners. Although we also found a significant main effect of accent on Recall scores when gender was collapsed, disaggregating data by gender revealed key differences between men and women.

Our results may possibly be explained by differing modality preferences and skills. For example, An and Carr [3] reported that individual differences in verbal and visual skills can help predict academic achievement. Prior work found that men have an inclination towards visual skills, while women have an inclination towards auditory skills [2, 52, 59]. The multimedia lesson used in our study was multimodal and included both visual elements and verbal narration. In our study, female participants may have relied more on the agent's narration to learn, which therefore caused their learning outcomes to be negatively affected by the foreign accent. On the other hand, male participants may have relied more on the visuals of the lesson (i.e., the 14 image sequence) and were thus unaffected by the agent's accent. Furthermore, our results indicate that research investigating pedagogical agents should consider gender balancing participants. In order to properly interpret results, it can be important to disaggregate data by gender due to possible differences among genders [62].

5.2 User Gender May Affect Perception of Agents

We also found a significant interaction effect between learner gender and agent accent on perception of the agent. As seen in Figure 3, the Human-like ratings of male and female learners intersect across agent accent conditions. Although we found no significant differences for pairwise comparisons, it is interesting to note that male and female learners trended in opposite directions. While male learners tended to rate the agent with a US accent as more Human-like than the one with an Indian accent, the opposite ratings trended for female learners.

We hypothesize that this rating trend may be attributed to differences in viewing patterns. Hewig et al. [32] found that both male and female viewers first fixate on a woman's face, and then the rest of the body from top to bottom. However, they reported that male viewers took a longer time to scan the figure compared to female viewers, and focused on the torso more. We hypothesize that male learners may have looked at the agent's attire more than female learners. Since the virtual agent wore a formal Western suit (see Figure 1), male learners may have felt that the US accent better aligned with the appearance. On the other hand, female viewers may have fixated on the facial features of the virtual agent, and felt that these features better aligned with the Indian accent. Further work, such as the use of eye tracking may be helpful to determine which areas of the agent the participants are looking at.

We used the same agent appearance for both conditions, considering previous results indicating that neither agent appearance nor mismatched appearance/accent affects agent perception or learning outcomes [28]. However, when participant gender was collapsed, we also did not find any significant differences regarding agent perception measures. These results only become apparent upon disaggregating data by gender. Thus, we recommend that future studies gender balance participants and disaggregate data when investigating perception of pedagogical agents.

5.3 Guideline for Pedagogical Application Developers

Prior researchers suggested that personalizing pedagogical agents might enrich learning experiences [21, 35]. We found that the voice effect principle is applicable to synthetic voices, at least for female students. While more research is necessary to validate and fully understand the results of our current study, we do recommend that pedagogical application developers should make an effort and take steps to align the accents of any pedagogical virtual agents with the country of the learner. With commercially available TTS engines, like Microsoft Azure's, developers can update the accents of any virtual agents with synthesized speech in real time. Hence, by accounting for the native country of the learner (e.g., US, UK, India), developers can easily personalize the synthesized accents to match, which should reduce the likelihood that a learner would perceive an accent to be foreign. Furthermore, while we have investigated virtual humans in a web browser, we believe this guideline is likely applicable for the broad range of pedagogical agents, from voice-only agents to embodied agents in immersive environments.

However, it is also important to note that not all synthetic accents are currently available for use in commercial services, and thus accommodating for the native country of a student may not be possible. For example, both AWS and Azure currently lack accents for Mexico English and China English, which are the native countries of the two largest immigrant populations in the United States [24]. Due to these reasons, we encourage TTS developers to create voices with more diverse accents in order to foster learning for all students.

5.4 Future Work and Limitations

It is important to note that our results are potentially limited to our specific study design. First, we have only investigated one educational lesson so far (i.e., the formation of lightning) and one presentation format (i.e., a multimedia video of the virtual pedagogical agent and a series of 14 images). Different educational lessons (e.g., quantum mechanics, calculus) or different presentation formats (e.g., voice only, virtual reality) may yield different results. Additionally, our lesson was relatively short (around 140 seconds), and longer lessons may also yield different results [4]. Second, we have only investigated the effects of a female virtual agent on US-based university students. We do not know whether our results would hold for university students native to another country for the female virtual agent, let alone a male version. Furthermore, our results may be different if the students were middle or high school students, which can be important since commercial TTS engines are also used for online education platforms for primary and secondary schools.

Additionally, although accents can vary by region, we analyzed only one Indian accent that was provided by commercial TTS engines. Unfortunately, Microsoft Azure does not provide information on the specific region that the synthetic accent is based on, and only lists the accent as "Indian". However, since our participants are not familiar with any Indian languages, we may be able to assume that all Indian accents would be unfamiliar to them.

In the future, we plan to run a comparative study using a male pedagogical agent with male speech. Although there is some evidence that instructor gender does not affect instructional videos [26, 55], more research is required to understand possible interaction effects with synthetic speech accents [7]. Additionally, we plan to utilize eye tracking software to determine where the participants are looking at during the lesson. This could help provide insights on our results, such as why male and female learners have different ratings of Human-likeness. Finally, since no students reported as non-binary, we only performed factorial analyses on male and female participants. Future work should consider recruiting non-binary participants and performing factorial analyses by gender.

6 CONCLUSION

We investigated the effect of a virtual pedagogical agent's synthetic speech accent on learning outcomes and perception. We chose to investigate synthetic speech due to its convenience in development and efficient performance. In particular, we chose to investigate the effect of TTS accents because accents can prime social cues and affect learning. Surprisingly, we found an interaction effect between learner gender and synthetic speech accent on recall scores, which indicates that a foreign synthetic accent may only affect the recall of female learners. Our results may be attributed to differing modality skills of male and female learners. Furthermore, we found an interaction effect between learner gender and TTS accent on ratings of human-likeness. While male learners tended to rate the agent with a US accent as more human-like, female learners tended to rate the agent with an Indian accent as more human-like. Overall, our results suggest that user gender plays an important role in the perception and effectiveness of virtual pedagogical agents.

ACKNOWLEDGMENTS

This material is based on work partially supported by the National Science Foundation under Grant No. 2021607 – "CAREER: Leveraging the Virtualness of Virtual Reality for More-Effective Training."

REFERENCES

[1] Jeahyeon Ahn and David Moore. 2011. The relationship between students' accent perception and accented voice instructions and its effect on students' achievement in an interactive multimedia environment. *Journal of Educational Multimedia and Hypermedia* 20, 4 (2011).

[2] Magnus Alm and Dawn Behne. 2015. Do gender differences in audio-visual benefit and visual influence in audio-visual speech perception emerge with age? *Frontiers in Psychology* 6, July (2015). https://doi.org/10.3389/fpsyg.2015.01014

[3] Donggun An and Martha Carr. 2017. Learning styles theory fails to explain learning and achievement: Recommendations for alternative approaches. *Personality and Individual Differences* 116 (2017), 410–416. https://doi.org/10.1016/j.paid.2017.04.050

[4] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/2696454.2696464

[5] Ekaterina Arshavskaya. 2015. International Teaching Assistants' Experiences in the U.S. Classrooms: Implications for Practice. *Journal of the Scholarship of Teaching and Learning* 15, 2 (2015), 56–69. https://doi.org/10.14434/josotl.v15i2.12947

[6] Robert K. Atkinson, Richard E. Mayer, and Mary Margaret Merrill. 2005. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology* 30, 1 (2005), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001

[7] Roger Azevedo, François Bouchet, Melissa Duffy, Jason Harley, Michelle Taub, Gregory Trevors, Elizabeth Cloude, Daryn Dever, Megan Wiedbusch, Franz Wortha, et al. 2022. Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning with an Intelligent Tutoring System. *Frontiers in Psychology* (2022), 1656.

[8] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. 2017. Perception of paralinguistic traits in synthesized voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*. 1–5.

[9] Amy Baylor. 2003. Effects of Images and Animation on Agent Persona. *Journal of Educational Computing Research* 28, 4 (2003), 373–394.

[10] Amy L Baylor and Yanghee Kim. 2003. The role of gender and ethnicity in pedagogical agent perception. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education. Association for the Advancement of Computing in Education (AACE)*.

[11] Sally Boyd. 2003. Foreign-born Teachers in the Multilingual Classroom in Sweden: The Role of Attitudes to Foreign Accent. *International Journal of Bilingual Education and Bilingualism* 6, 3-4 (2003), 283–295. https://doi.org/10.1080/13670050308667786

[12] Donn Byrne and Don Nelson. 1964. Attraction as a function of attitude similarity-dissimilarity: the effect of topic importance. *Psychonomic Science* 1, 1-12 (1964), 93–94. https://doi.org/10.3758/bf03342806

[13] Sung-Woo Byun and Seok-Pil Lee. 2021. Design of a Multi-Condition Emotional Speech Synthesizer. *Applied Sciences* 11, 3 (Jan 2021), 1144. https://doi.org/10.3390/app11031144

[14] Lam Vien Cao Ngoc. 2014. *Effects of speaker's accent in a multimedia tutorial on non-native students' learning and attitudes.* Ph. D. Dissertation.

[15] Kit Ying Chan, Claire Lyons, Lo Lo Kon, Katelyn Stine, Melissa Manley, and Anthony Crossley. 2020. Effect of on-screen text on multimedia learning with native and foreign-accented narration. *Learning and Instruction* 67, August 2018 (2020), 101305. https://doi.org/10.1016/j.learninstruc.2020.101305

[16] Erin K. Chiou, Noah L. Schroeder, and Scotty D. Craig. 2020. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers*

and Education 146 (2020). https://doi.org/10.1016/j.compedu.2019.103756

[17] Scotty D. Craig, Barry Gholson, and David M. Driscoll. 2002. Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology* 94, 2 (2002), 428–434. https://doi.org/10.1037/0022-0663.94.2.428

[18] Scotty D. Craig and Noah L. Schroeder. 2017. Reconsidering the voice effect when learning from a virtual human. *Computers and Education* 114 (2017), 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

[19] Nils Dahlbäck, Seema Swamy, Clifford Nass, Stanford Ca, and Jörgen SkågEby. 2001. Spoken Interaction with Computers in a Native or Non-Native Language - Same or Different ? *Human Computer Interact '01* (2001), 294–301.

[20] Shivangi Dhawan. 2020. Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems* 49, 1 (2020), 5–22. https://doi.org/10.1177/0047239520934018

[21] Tiffany D. Do. 2021. Designing Virtual Pedagogical Agents and Mentors for Extended Reality. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 476–479. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00112

[22] Tiffany D. Do, Ryan P. McMahan, and Pamela J. Wisniewski. 2022. A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 424, 11 pages. https://doi.org/10.1145/3491102.3517564

[23] Steffi Domagk. 2010. Do pedagogical agents facilitate learner motivation and learning outcomes?: The role of the appeal of agent's appearance and voice. *Journal of Media Psychology* 22, 2 (2010), 84–97. https://doi.org/10.1027/1864-1105/a000011

[24] Susan Eckstein and Giovanni Peri. 2018. Immigrant Niches and Immigrant Networks in the U.S. Labor Market. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 4, 1 (2018), 1–17. https://doi.org/10.7758/RSF.2018.4.1.01 arXiv:https://www.rsfjournal.org/content/4/1/1.full.pdf

[25] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '21)*. 754–768. https://doi.org/10.1145/3472749.3474784 arXiv:2102.11824

[26] Logan Fiorella and Richard E. Mayer. 2018. What works and doesn't work with instructional video. *Computers in Human Behavior* 89 (2018), 465–470. https://doi.org/10.1016/j.chb.2018.07.015

[27] Fred Fitch and Susan E. Morgan. 2003. "Not a Lick of English": Constructing the ITA Identity Through Student Narratives. *Communication Education* 52, 3-4 (2003), 297–310. https://doi.org/10.1080/0363452032000156262

[28] R. G. P. Galluccio. 2008. Animated Pedagogical Agents as Spanish Language Instructors: Effect of Accent, Appearance, and Type of Activity on Student Performance, Motivation, and Perception of Agent. 69, 12-A (2008), 4697.

[29] Mary M. Gill. 1994. Accent and Stereotypes: Their Effect on Perceptions of Teachers and Lecture Comprehension. *Journal of Applied Communication Research* 22, 4 (1994), 348–361. https://doi.org/10.1080/00909889409365409

[30] Arthur C. Graesser. 2016. Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 124–132. https://doi.org/10.1007/s40593-015-0086-4

[31] Jason M. Harley, Michelle Taub, Roger Azevedo, and François Bouchet. 2018. "Let's Set Up Some Subgoals": Understanding Human-Pedagogical Agent Collaborations and Their Implications for Learning and Prompt and Feedback Compliance. *IEEE Transactions on Learning Technologies* 11, 1 (2018), 54–66. https://doi.org/10.1109/TLT.2017.2756629

[32] Johannes Hewig, Ralf H. Trippe, Holger Hecht, Thomas Straube, and Wolfgang H.R. Miltner. 2008. Gender differences for specific body regions when looking at men and women. *Journal of Nonverbal Behavior* 32, 2 (2008), 67–78. https://doi.org/10.1007/s10919-007-0043-5

[33] W. Lewis Johnson and James C. Lester. 2016. Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 25–36. https://doi.org/10.1007/s40593-015-0065-9

[34] Okim Kang and Meghan Moran Wilson. 2019. Enhancing Communication Between ITAs and US Undergraduate Students. A Transdisciplinary Approach to ITA Research: Perspectives from Applied Linguistics. In *Multilingual Matters*.

[35] Yanghee Kim and Amy L. Baylor. 2016. Research-Based Design of Pedagogical Agent Roles: A Review, Progress, and Recommendations. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 160–169. https://doi.org/10.1007/s40593-015-0055-y

[36] Yanghee Kim and Jae Hoon Lim. 2013. Gendered Socialization with an Embodied Agent: Creating a Social and Affable Mathematics Learning Environment for Middle-Grade Females. *Journal of Educational Psychology* 105, 4 (2013), 1164–1174. https://doi.org/10.1037/a0031027

[37] Brigitte Krenn, Stephanie Schreitter, and Friedrich Neubarth. 2017. Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI and Society* 32, 1 (2017), 65–77. https://doi.org/10.1007/s00146-014-0569-0

[38] James C Lester, Stuart G Towns, Charles B Callaway, Jennifer L Voerman, and Patrick J Fitzgerald. 2000. Deictic and Emotive Communication in Animated Pedagogical Agents. In *Embodied Conversational Agents*. MIT press, Cambridge, MA, 123–154. https://doi.org/10.7551/mitpress/2697.003.0007

[39] Ati Suci Dian Martha and Harry B. Santoso. 2019. The design and impact of the pedagogical agent: A systematic literature review. *Journal of Educators Online* 16, 1 (2019). https://doi.org/10.9743/jeo.2019.16.1.8

[40] Richard E. Mayer. 2014. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. In *The Cambridge Handbook of Multimedia Learning, Second Edition*. Number May 2017. 345–368. https://doi.org/10.1017/CBO9781139547369.017

[41] Richard E. Mayer and Roxana Moreno. 1998. A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology* 90, 2 (1998), 312–320. https://doi.org/10.1037/0022-0663.90.2.312

[42] Richard E. Mayer and Celeste Pilegard. 2014. Principles for managing essential processing in multimedia learning: Segmenting, pre-training, and modality principles. In *The Cambridge Handbook of Multimedia Learning, Second Edition*. 316–344. https://doi.org/10.1017/CBO9781139547369.016

[43] Richard E. Mayer, Kristina Sobko, and Patricia D. Mautone. 2003. Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology* 95, 2 (2003), 419–425. https://doi.org/10.1037/0022-0663.95.2.419

[44] Shannon M McCrocklin, Kyle P Blanquera, and Deyna Loera. 2017. Student Perceptions of University Instructor Accent in a Linguistically Diverse Area. In *Proceedings of the 9th Pronounciation in Second Language Learning and Teaching conference*. 141–150.

[45] Conor McGinn and Ilaria Torre. 2019. Can you Tell the Robot by the Voice?: An Exploratory Study on the Role of Voice in the Perception of Robots. *ACM/IEEE International Conference on Human-Robot Interaction* (2019), 211–221. https://doi.org/10.1109/HRI.2019.8673279

[46] Roxana Moreno, Richard E. Mayer, Hiller A. Spires, and James C. Lester. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction* 19, 2 (2001), 177–213. https://doi.org/10.1207/S1532690XCI1902_02

[47] Roxana Moreno and Richard E. and Mayer. 1999. Cognitive principles of multimedia learning: the role of modality and contiguity. *Journal of Educational Psychology* 91, 2 (1999), 358–368.

[48] Murray J. Munro and Tracey M. Derwing. 1995. Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. *Language and Speech* 38, 3 (1995), 289–306. https://doi.org/10.1177/002383099503800305

[49] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are Social Actors. *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), 72–78. https://doi.org/10.1109/VSMM.2014.7136659

[50] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.

[51] Fred Paas and Jeroen J.G. van Merriënboer. 2020. Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. *Current Directions in Psychological Science* 29, 4 (2020), 394–398. https://doi.org/10.1177/0963721420922183

[52] Franz Pauls, Franz Petermann, and Anja Christina Lepach. 2013. Gender differences in episodic memory and visual working memory including the effects of age. *Memory* 21, 7 (2013), 857–874. https://doi.org/10.1080/09658211.2013.765892

[53] Tamara Rakić, Melanie C. Steffens, and Amélie Mummendey. 2011. Blinded by the Accent! The Minor Role of Looks in Ethnic Categorization. *Journal of Personality and Social Psychology* 100, 1 (2011), 16–29. https://doi.org/10.1037/a0021522

[54] Anara Sandygulova and Gregory M.P. O'Hare. 2015. Children's perception of synthesized voice: Robot's gender, age and accent. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9388 LNCS (2015), 594–602. https://doi.org/10.1007/978-3-319-25554-5_59

[55] Claudia Schrader, Tina Seufert, and Steffi Zander. 2021. Learning From Instructional Videos: Learner Gender Does Matter; Speaker Gender Does Not. *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.655720

[56] Noah L. Schroeder, Olusola O. Adesope, and Rachel Barouch Gilbert. 2013. How effective are pedagogical agents for learning? a meta-analytic review. *Journal of Educational Computing Research* 49, 1 (2013), 1–39. https://doi.org/10.2190/EC.49.1.a

[57] Noah L. Schroeder, Fan Yang, Tanvi Banerjee, William L. Romine, and Scotty D. Craig. 2018. The influence of learners' perceptions of virtual humans on learning transfer. *Computers and Education* 126 (2018), 170–182. https://doi.org/10.1016/j.compedu.2018.07.005

[58] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction. *Comput. Surveys* 54, 4 (2021), 1–43. https://doi.org/10.1145/3386867

[59] Sneha Shetty. 2016. Influence of gender in learning style preference in undergraduate medical students. In *Proceedings of ASAR-IJIEEE International Conference*. 28–29. https://www.digitalxplore.org/up_proc/pdf/253-147911407928-29.pdf

[60] N. C. Subtirelu and S. Lindemann. 2014. Teaching First Language Speakers to Communicate Across Linguistic Difference: Addressing Attitudes, Comprehension, and Strategies. *Applied Linguistics* (2014), 1–20. https://doi.org/10.1093/applin/amu068

[61] Rie Tamagawa, Catherine I. Watson, I. Han Kuo, Bruce A. Macdonald, and Elizabeth Broadbent. 2011. The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics* 3, 3 (2011), 253–262. https://doi.org/10.1007/s12369-011-0100-4

[62] Cara Tannenbaum, Robert P. Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. 2019. Sex and gender analysis improves science and engineering. *Nature* 575 (2019), 137–146. https://doi.org/10.1038/s41586-019-1657-6

[63] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*. 143–146. https://dl.acm.org/doi/pdf/10.1145/1978942.1978963

[64] Ruth Wong. 2018. Non-native EFL Teachers' Perception of English Accent in Teaching and Learning: Any Preference? *Theory and Practice in Language Studies* 8, 2 (2018), 177. https://doi.org/10.17507/tpls.0802.01

A RECALL QUESTIONNAIRE

Recall questionnaire from [17]. correct answers are denoted with ^{a}.
Explicit—deep

(1) What causes a flash of lightning?
  (a) The return stroke^a
  (b) Negatively charged leader
  (c) Positively charged leader
  (d) Negative charges rushing from the cloud

(2) When do downdrafts occur?
  (a) When air is dragged down by rain^a
  (b) When air currents cool and fall back to earth
  (c) When cold air hits the ground
  (d) When there are unbalanced electrical charges between the ground and the clouds

Explicit—shallow

(3) The upper portion of the cloud is made up of what?
  (a) Water droplets
  (b) Cold air
  (c) Ice crystals^a
  (d) Water vapor

(4) What part of the cloud are the positively charged particles located in?
  (a) Bottom part
  (b) Center of the cloud
  (c) Outside edge
  (d) Top part^a

Implicit—deep

(5) Why does lightning strike buildings and trees?
  (a) They are higher than the ground
  (b) A build-up of positive charges
  (c) It is the point where the negative leader ends
  (d) Positive leader starts at these points^a

(6) Why does it get colder right before it rains?
  (a) Positive charges are absorbed into the clouds
  (b) Warm moist air rushes upward into the clouds
  (c) Cold downdrafts of air fall from the clouds^a
  (d) Warm surface air rapidly cools