Social Cues in Multimedia Learning: Role of Speaker’s Voice

Richard E. Mayer, Kristina Sobko, and Patricia D. Mautone
University of California, Santa Barbara

In 2 experiments, learners who were seated at a computer workstation received a narrated animation about lightning formation. Then, they took a retention test, took a transfer test, and rated the speaker. There was a voice effect, in which students performed better on the transfer test and rated the speaker more positively if the voice in the narration had a standard accent rather than a foreign accent (Experiment 1) and if the voice was human rather than machine synthesized (Experiment 2). The retention test results were mixed. The results are consistent with social agency theory, which posits that social cues in multimedia messages can encourage learners to interpret human–computer interactions as more similar to human-to-human conversation.

Developers of computer-based multimedia instruction are working hard to create learning experiences in which the learner will accept the computer as a social partner, mainly through the use of on-screen agents that talk to the learner (Cassell, Sullivan, Prevost, & Churchill, 2000). Yet, to date, there has been no systematic research on the role of voice in promoting deep learning from multimedia lessons. In this study, we examine the idea that the speaker’s voice in multimedia lessons carries important social cues that can influence the process and outcome of learning.

Two Interpretations of a Multimedia Instructional Episode

Consider a learning scenario in which a single student sits at a computer station and receives a multimedia instructional message. For example, the student opens a multimedia encyclopedia and types in lightning. In response, a 140-s narrated animation is presented, in which the steps in lightning formation are described orally and the corresponding animation visually depicts the steps. We refer to this scenario as a multimedia learning episode.

At first glance, this episode does not seem to involve social interaction. The simplest characterization is that this episode is a case of information delivery (Mayer, 2001), in which the computer delivers information about lightning to the learner and the learner receives it. However, an alternative analysis of the episode suggests that this can be a case of social conversation (Moreno & Mayer, 2000; Reeves & Nass, 1996), in which the learner and computer are in a sort of conversation in which the learner assumes that the computer is trying to explain something to the learner and the learner is trying to understand.

Social Agency Theory

What does it mean to view a computer-based instructional episode as a social event? Following earlier work by Moreno and Mayer (2000) on social cues in multimedia learning and by Reeves and Nass (1996) on how people interpret computers as social partners, we propose social agency theory. The main thesis in social agency theory is that social cues in a multimedia message can prime the social conversation schema in learners. Once the social conversation schema is activated, learners are more likely to act as if they are in a conversation with another person. Thus, at least to some extent, the social rules of human-to-human communication come into play. These rules include what Grice (1975) referred to as the cooperation principle and its four conversational maxims—that is, listeners assume that the speaker is trying to make sense by being informative, accurate, relevant, and concise.

Once learners interpret their interaction with a computer as social, the rules of human-to-human communication come into play, so they try harder to make sense of what the computer is saying by engaging in deep cognitive processing. Mayer (1999, 2001) analyzed these cognitive processes into selecting relevant information for further processing, organizing the pieces of information into coherent representations, and integrating verbal and visual representations with each other and with prior knowledge. Deep cognitive processing results in meaningful learning outcomes, which enable learners to apply (or transfer) what they have learned to new situations. Thus, the major educational outcome of social cues in multimedia learning is reflected in tests of transfer.

The social agency model is consistent with wider research in discourse processing and classroom teaching. In research on transactional learning theories (Beck, McKeown, Sandra, Kucan, & Worthy, 1996; Schraw & Bruning, 1996), students engaged in deeper elaboration of text when they viewed the author as communicating directly with them. In research on reading stories, Graesser, Bowers, Olde, and Pomeroy (1999) found that learners were more aware of the narrator when the narrator wrote in the first person rather than in the third person. In research on classroom teaching (Cohen, Kulik, & Kulik, 1982; Graesser, Person, &
Magliano, 1995), conversational speech aided learning above and beyond lecturing.

More specifically, social agency theory is based on the idea that the learner can interpret a multimedia learning episode as either a case of information delivery or a case of social communication. Further, the learner’s interpretation of the episode—as social communication versus information delivery—influences the type of schemas that are activated in the learner, the type of cognitive processing that occurs during learning, and ultimately the quality of the learning outcome. First consider what happens for a social-conversation scenario. If the learner receives a multimedia message with strong social cues (such as a human voice speaking in a standard accent), the learner is more likely to interpret the episode as a case of social conversation. In this case, the conventions of human conversational exchange are primed in the learner, including the idea that the computer is obliged to say something that makes sense to the learner and the idea that the learner is obliged to try to make sense of what the computer is saying. Therefore, the learner engages in a sense-making process including selecting relevant information, organizing it into a coherent representation, integrating it with other knowledge, and encoding it in memory. The result of sense-making processing is the construction of a meaningful learning outcome, which supports good performance on transfer tests.

In contrast, consider the events for the information-delivery scenario. If the learner receives a multimedia message with weak social cues (such as a machine voice or a foreign-accent voice), then the learner is more likely to interpret the episode as a case of information delivery. In this case, human-to-human conversational rules are not activated, so the learner uses cognitive processing aimed solely at acquiring information rather than trying to understand it. When the learner views the episode as a case of information delivery, the learner is more likely to engage in a rote learning approach of paying attention to key ideas and trying to store them in memory—processes that can be called selecting and encoding, respectively. This processing leads to rote learning outcomes that lead to poor performance on transfer tests.

Reeves and Nass (1996, p. 5) provided evidence for the idea that “individuals’ interactions with computers . . . are fundamentally social and natural just like interactions in real life.” They proposed the notion of the “media equation” (p. 5), that is, the idea that “media equal real life” (p. 5) or “media experiences equal human experiences” (p. 251). Thus, “people automatically and unconsciously respond socially and naturally to media” (p. 7) and “people automatically use social rules from real life to guide interactions with media” (p. 59).

We seek to extend the study of the media equation in two ways. First, we focus on computer-based educational scenarios. Thus, our major dependent variables are cognitive—concerning how well someone learned—rather than solely affective—concerning how someone felt about a persuasive message or how someone rated a computer-mediated experience. Second, we focus on the conditions under which learners are most likely to use social rules to guide their interactions with computer-based messages. Although Reeves and Nass (1996, p. 5) stated that “it is very easy to foster” a sense of social interaction with a computer, we seek to pinpoint some conditions under which a learner responds to the computer as a social partner.

Social Cues in Multimedia Lessons

How can we promote deep learning from on-line instructional lessons involving narration? In the present set of studies, we turn our attention to the role of the instructor’s voice in multimedia learning. We are motivated by the premise that learning is an inherently social activity because it always involves some sort of transaction between a learner and an instructor. Social agency is established when the learner presumes that the instructor is an intelligent agent who is engaging the learner in a conversation. We focus on one aspect of a multimedia lesson that might increase the learner’s attribution of social agency: when the speaker has a socially appealing voice (rather than a nonsocially appealing voice). To vary the social appeal of the instructor’s voice, we compare multimedia lessons, in which the voice comes from someone who is or is not a native speaker of standard American English (i.e., standard-accented speech vs. foreign-accented speech in Experiment 1) or when the speech is produced by a native speaker of standard American English or a machine-synthesized voice (i.e., human vs. machine voice in Experiment 2).

We propose that students try harder to understand a multimedia instructional message when social agency is primed than when it is not—that is, when students view the computer as a conversational partner rather than a device for presenting information. Furthermore, we propose that this extra effort to make sense of the presented material results in deeper learning, as measured by tests of problem-solving transfer. On the basis of social agency theory, we predict that students will perform better on tests of transfer—which are designed to measure deep learning—when the narration in a narrated animation comes from a standard-accented voice rather than a foreign-accented voice (Experiment 1) and when the narration comes from a human voice rather than a machine voice (Experiment 2). On the basis of social agency theory, we also predict that students will produce higher ratings of the speaker on socially desirable characteristics when the speaker uses a human voice with a standard accent rather than a human voice with a foreign accent (Experiment 1) or a machine voice (Experiment 2).

Voice carries a great deal of information beyond the nominal instructional message in the words of a narration—including information regarding the speaker’s suitability as a conversational partner. There is a growing body of research showing that people interpret computer-based speech from on-screen agents in much the same way as conversations with humans. Although not focusing on learning outcomes per se, Nass and Lee (2001) reported that Web site users recognized and preferred certain personality cues in computer-generated speech. Nass, Isbister, and Lee (2000) reported that students rated computer-based instructional agents as more competent and useful when they displayed the student’s ethnicity and personality characteristics. Oviatt and Adams (2000) found that children adjusted their speech to be slower and clearer when talking with a computer-based animated character, indicating that they viewed the computer character as a sort of at-risk learner. Taken together, these kinds of results are consistent with the idea that people interpret computer agents—that often communicate through speech—as social partners. Similarly, Lester and his colleagues (Lester, Towns, Callaway, Voerman, & Fitzgerald, 2000) have argued for persona effects with animated pedagogical agents, in which the social characteristics of pedagogical agents influence how well people like interacting with them. However, developers
of conversational agents generally have not examined how voice affects learning (Cassell et al., 2000).

Although voice may produce social effects in multimedia learning, it also may yield cognitive effects. According to cognitive load theory, students must allocate fewer cognitive resources to processing a human voice speaking with a standard accent than to processing a human voice speaking with a foreign accent or a machine-synthesized voice, leaving more cognitive resources for deep processing of the instructional message. As a result, cognitive load theory makes the same predictions concerning problem-solving transfer performance as does social agency theory (but does not make any predictions concerning speaker ratings).

According to the cognitive effort theory, students work harder in a cognitively challenging situation (such as trying to comprehend a machine voice or foreign-accented human voice) than in an easier situation (such as a standard-accented human voice). As a result, cognitive effort theory predicts better performance on the retention and transfer tests for the machine voice or foreign-accented voice than the standard-accented voice.

Experiment 1

In Experiment 1, students received a narrated animation that explained how lightning forms, in which the words of the narration were spoken by a male native-English speaker (standard-accent group) or by a male speaker who had a Russian accent (foreign-accent group). According to social agency theory, students in the standard-accent group should produce higher scores than students in the foreign-accent group on transfer tests designed to measure the depth of learner understanding and on rating scales designed to measure the learner’s evaluation of the speaker.

Method

Participants and design. The participants were 68 college students recruited from the Psychology Subject Pool at the University of California, Santa Barbara. Thirty-four participants served in the standard-accent group and 34 participants served in the foreign-accent group. The mean age (in years) of the participants was 18.9 for the standard-accent group and 18.8 for the foreign-accent group. The percentage of women was 69% for the standard-accent group and 65% for the foreign-accent group. The mean combined SAT score was 1180 for the standard-accent group and 1192 for the foreign-accent group. Each treatment group was divided into four subgroups, with each subgroup consisting of 18 participants. Within each treatment group, we also varied the wording of an introduction, but we have not included that factor in this report because it produced no effects on any measures.

Materials and apparatus. The paper-based materials consisted of a participant questionnaire, a retention test, four transfer test questions, and a speaker-rating survey. The participant questionnaire solicited demographic information. The retention test contained the following sentence at the top of the sheet: “Please write down an explanation of how lightning works.” Each of the four transfer test sheets contained one of the following questions at the top of the sheet: “What could you do to decrease the intensity of lightning?” “Suppose you see clouds in the sky but no lightning, why not?” “What does air temperature have to do with lightning?” “What causes lightning?” At the bottom of each sheet were the words “PLEASE KEEP WORKING UNTIL YOU ARE TOLD TO STOP.” The speaker-rating survey was a 15-item instrument adapted from Zahn and Hopper’s (1985) Speech Evaluation Instrument. Instructions at the top of the page asked the participant to circle a number from 1 to 8 indicating how the speaker sounded along each of 15 dimensions. For each dimension, the numbers 1 through 8 were printed along a line with one adjective above the 1 and an opposite adjective above the 8. The 15 adjective pairs were as follows: literate–illiterate, unkind–kind, active–passive, intelligent–unintelligent, cold–warm, talkative–shy, uneducated–educated, friendly–unfriendly, aggressive–aggressive, fluent–not fluent, unpleasant–pleasant, confident–unsure, inexperienced–experienced, unlikeable–likeable, and energetic–lazy. There were five items from each of three subscales—Superiority, Attractiveness, and Dynamism. An overall speaker rating (from 1 to 8) was constructed by averaging the scores from the three subscales, with 1 indicating the most negative rating and 8 indicating the most positive rating. We adapted Zahn and Hopper’s instrument because of its effectiveness in detecting the social characteristics attributed to speakers.

The computer-based materials consisted of two versions of a multimedia program on lightning formation created using Director 6.0 (Macromedia, 1997). Both versions of the program consisted of a 140-s narrated animation, in which the animation depicted 16 steps in the formation of lightning, and the accompanying narration described the steps as shown in the Appendix. For the standard-accent version, the narration was recorded by a man who spoke with a standard American accent; the voice was identical to that used by Mayer and Moreno (1998). For the foreign-accent version, the narration was recorded by a man who said exactly the same words but with a noticeable Russian accent. On the basis of a supplemental study, we found that participants could discern 100% of the words in the standard-accent version and 97% of the words in the foreign-accent version.

The apparatus consisted of four Macintosh computer systems with 17-in. color monitors and Sony headphones.

Procedure. Participants were tested in groups of 1 to 4 per session and were randomly assigned to a treatment group. Each participant was seated at an individual cubicle containing work space and a Macintosh G3 computer system. First, participants completed the participant questionnaire. Second, the experimenter told the participants that they would receive an explanation of how lightning storms develop and that afterwards they would be asked some questions. They were then told to put on the headphones and press the space bar to begin the presentation. On the basis of random assignment, participants in the standard-accent group received the standard-accent version of the program, and participants in the foreign-accent group received the foreign-accent version of the program. Third, after the presentation ended, students were instructed to take off the headphones; the experimenter passed out the retention test and asked students to keep writing until told to stop. Fourth, after 4 min, the retention test was collected and the first transfer test sheet was distributed. After 2.5 min, it was collected and the next transfer sheet was distributed and so on until all four transfer sheets had been completed. Next, participants completed the speaker-rating survey at their own rates. Finally, participants were thanked and debriefed.

Scoring. All scoring was done by one or more scorers who were unaware of the participants’ treatment condition, following the same procedure as used by Mayer and Moreno (1998). The retention test was scored by tallying the number of idea units from the narration (with a maximum of eight) that the participant had written down, regardless of specific wording. The idea units were as follows: (a) air rises, (b) water condenses, (c) water and crystals fall, (d) wind is dragged downward, (e) negative charges fall to the bottom of the cloud, (f) the leaders meet, (g) negative charges rush down, and (h) positive charges rush up. We consider this a retention test because participants are asked to recite the information that was presented in the lesson.

The transfer test was scored by tallying the number of acceptable answers given across all four questions (with a maximum obtained score of 8). Some acceptable answers for the first question about decreasing the intensity of a lightning storm included statements about removing positively charged particles from the ground or placing positively charged particles near the cloud. Some acceptable answers for the second question about seeing clouds but no lightning included stating that the top of the cloud might not be above the freezing level or that there are no negatively charged particles in the cloud. Some acceptable answers for the third
question about air temperature included stating that the air must be cooler than the surface of the earth or that the top of the cloud must be colder than the bottom. Some acceptable answers for the fourth question about the causes of lightning included stating that there must be a difference of electrical charge within the cloud or between the cloud and the ground.

The fourth transfer question ("What causes lightning?") appears to be similar to the retention question, but the questions differ in what is required of the learner. In the retention question, the learner is required to list the main steps in lightning formation. In the transfer question, the learner is required to be even more selective in describing the steps and in organizing them into a coherent explanation, for example, to abstract out the idea that a central cause of lighting is that the positive particles on the ground are attracted to the negative particles in the bottom of the cloud.

We consider these items to constitute a transfer test because participants must select and adapt what was presented in the lesson to fit the requirements of each question. Instead of simply being cued to recall what was presented, participants must judge which aspects of the presented material are relevant to the question and must determine how to link that material to their answer. In short, the transfer questions require the participants to go beyond simply recalling the explanation presented in the lesson, although recalling relevant portions of the lesson is certainly a component in the solution. For example, in answering the first question, the participant must recall a specific aspect of the lightning explanation, namely that negative particles in the cloud meet positive particles from the ground. Furthermore, they must adapt this information to the requirements of the question by suggesting that positive particles be removed from the earth’s surface. Although this idea is never presented in the lesson, it can be inferred and therefore represents a form of transfer.

Results and Discussion

The means and standard deviations for the two groups on the retention, transfer, and speaker-ratering measures are shown in Table 1. Separate t tests for independent means were conducted on the retention, transfer, and speaker-rating data.

**Does accented voice affect retention?** Students who received a presentation with an accented voice performed as well on the retention test as students who received a presentation with a nonaccented voice, t(66) = 0.26, p = .794, yielding an effect size of –0.06. Apparently, the foreign-accent group was able to devote cognitive capacity to hearing, encoding, and remembering the main facts in the presentation. Additional, deeper cognitive processing is required for the transfer test, so we are most interested in transfer as our major measure of meaningful learning outcome.

**Does accented voice affect transfer?** Consistent with social agency theory and cognitive load theory, students who received a narrated animation with a nonaccented voice performed significantly better on tests of problem-solving transfer than students who received a narrated animation with an accented voice, t(66) = 4.07, p < .001. The effect size was 0.80, indicating a relatively strong effect of accented voice on transfer. This result suggests that students in the standard-accent group put more effort into building meaningful learning outcomes than did students in the accent group.

**Does accented voice affect social rating of the speaker?** Consistent with the social agency theory, students who received a nonaccented voice rated the speaker more positively than did students who received an accented voice, t(66) = 3.27, p < .01. The effect size was 0.71. Our results show that learners in the standard-accent and foreign-accent groups made different social judgments about the speaker.

### Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Retention Test</th>
<th>Transfer Test</th>
<th>Speaker Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Standard accent</td>
<td>3.9</td>
<td>1.4</td>
<td>5.5</td>
</tr>
<tr>
<td>Foreign accent</td>
<td>3.8</td>
<td>1.4</td>
<td>3.7</td>
</tr>
</tbody>
</table>

*Standard-accent group scored significantly higher than foreign-accent group at p < .05.*

In Experiment 2, students received a narrated animation that explained how lightning forms, in which the words of the narration were spoken by a male native-English speaker (human-voice group) or by a machine-synthesized male voice (machine-voice group). According to social agency theory, students in the human-voice group should produce higher scores than students in the machine-voice group on transfer tests designed to measure the depth of learner understanding and on speaker-rating scales designed to measure the learner’s evaluation of the speaker.

### Method

**Participants and design.** The participants were 40 college students recruited from the Psychology Subject Pool at the University of California, Santa Barbara. Twenty-one participants served in the human-voice group and 19 served in the machine-voice group. The mean age was 19.2 for the human-voice group and 19.6 for the machine-voice group.

The percentage of women was 52% in the human-voice group and 56% in the machine-voice group. The mean combined SAT score was 1169 for the human-voice group and 1141 for the machine-voice group.

**Materials and apparatus.** The paper-based materials and apparatus were the same as in Experiment 1, except that a two-item presentation survey was included. The first item asked the participant to rate the difficulty of the presentation: "How difficult was it for you to learn about lightning from the presentation you just saw? Please place a check mark in one of the boxes below to indicate your rating." Seven alternatives were printed below the question: very easy, somewhat easy, slightly easy, medium, slightly hard, somewhat hard, and very hard. This item was adapted from an instrument used by Paas and Van Merrienboer (1993) designed to measure the participant’s perceived cognitive load during learning. Participants were assigned a score of 1 (very hard) to 7 (very easy) based on which alternative they marked. The second item asked the participant to rate their difficulty in deciphering what the narrator was saying: “Apart from the content of the presentation, how difficult was it to hear what the narrator was saying (i.e., how easy was it to understand the voice)?” The same alternatives and scoring scheme were used as for the first question, and the question was intended to evaluate the participant’s perceived difficulty in sensory processing of the narration.

The computer-based materials consisted of two versions (human voice and machine voice) of a narrated animation on lightning formation. The human-voice program was identical to the standard-accent program in Experiment 1, except that there was no introductory paragraph describing the credibility of the speaker. The machine-voice program was identical to the human-voice program, except that the voice was constructed using a machine simulated voice called Bruce (high quality) provided in the voice
folder of Macintosh G4 computer systems. We chose Bruce (high quality) because it was the most neutral and clearest male voice available in the voice folder and was typical of currently available low-cost synthesized voices. In a supplemental study, participants were able to discern 95% of the words spoken in the machine voice and 100% of the words spoken in the human voice.

Procedure. The procedure was the same as in Experiment 1, with two notable exceptions: (a) Participants in the human-voice group received the human-voice program and participants in the machine-voice group received the machine voice program, and (b) after the transfer test, participants were asked to complete the presentation survey.

Results and Discussion

Table 2 shows the mean score (and standard deviation) on retention, transfer, and speaker-rating measures for each group in Experiment 2. Separate t tests for independent means were conducted on the retention, transfer, and speaker-rating data.

Does mechanization of voice affect retention? As shown in the left portion of Table 2, the human-voice group scored higher than the machine-voice group on retention, t(38) = 2.09, p < .05, yielding an effect size of 0.66. Students in the machine-voice group were less likely to encode and/or remember the incoming material than were students in the human-voice group.

The retention score is not a traditional measure of total amount learned because it focuses only on recall of conceptually central events. Thus, the retention score taps some of the same aspects of meaningful learning outcome as does the transfer score, particularly the degree to which the learners actively focused on the important material in their attempts to make sense of the lesson. It is therefore not surprising that the correlation between the retention and transfer scores was .31 in Experiment 1 (p < .05) and .39 in Experiment 2 (p < .05).

Does mechanization of voice affect transfer? As shown in the middle portion of Table 2, the human-voice group scored higher than the machine-voice group on transfer, t(38) = 2.57, p < .02, yielding an effect size of 0.81. Consistent with both social agency theory and cognitive load theory, students in the machine-voice group were less likely to construct meaningful learning outcomes than were students in the human-voice group.

Does mechanization of voice affect speaker rating? As shown in the right portion of Table 2, the human-voice group rated the speaker more favorably than did the machine-voice group, t(38) = 4.19, p < .001, yielding an effect size of 1.45 for speaker rating. The less favorable rating by the machine-voice group can be summarized by saying that the machine-voice group viewed the speaker as less dynamic, less attractive, and less superior than did the human-voice group. This result is consistent with the social agency theory in which the learner attributes social characteristics to the computer-based presenter.

Overall, the human-voice group learned more, was better able to apply what was learned to solve new problems, and generally liked the speaker better than did the machine-voice group. The absolute level of the scores in Experiments 1 and 2 cannot be directly compared because the studies were conducted with different participants, a different experimenter, and during a different time period.

Does mechanization of voice affect perceived difficulty? The human-voice group (M = 4.0, SD = 1.4) rated the material in the program as easier to learn than did the machine-voice group (M = 2.8, SD = 1.7), t(38) = 2.58, p < .02. Paas and Van Merrienboer (1993) have shown that this kind of difficulty rating is an indication of the learner’s perceived cognitive load. Consistent with cognitive load theory, the human-voice group appears to have experienced less cognitive load than the machine-voice group. If the machine voice increases cognitive load, then learners in the machine-voice group would be less able to attend to relevant information and build connections, which in turn would lead to poorer performance on retention and transfer tests.

Finally, the human-voice group (M = 5.3, SD = 1.1) rated the voice as easier to understand than did the machine-voice group (M = 1.1, SD = 0.8), t(38) = 14.38, p < .001. Students in the machine-voice group may have had to exert considerable cognitive energy in listening to the voice and therefore may have had less energy available for processing the incoming material.

Related issues. To determine whether differences between the groups in Experiments 1 and 2 could be attributed to the intelligibility of the speech, we conducted a supplemental study, in which we asked 12 participants to listen to each of 12 sentences and write down the words. In a counterbalanced within-subject design, 4 sentences were spoken in a human voice (corresponding to the standard-accent group in Experiment 1 and the human-voice group in Experiment 2), 4 sentences were spoken in a machine voice (corresponding to the machine-voice group in Experiment 2), and 4 sentences were spoken in a foreign-accent voice (corresponding to the foreign-accent group in Experiment 2). Participants correctly recorded an average of 100% of the human voice (or standard-accent voice) sentences, 97% of the foreign-accent voice sentences, and 95% of the machine-voice sentences. The majority of errors were attributable to two words: leader and updraft. Overall, almost all of the words were discernable by learners regardless of the speaker’s voice.

Table 2
Mean Retention, Transfer, and Speaker-Rating Scores (and Standard Deviations) for Two Groups in Experiment 2

<table>
<thead>
<tr>
<th>Group</th>
<th>Retention M</th>
<th>Retention SD</th>
<th>Transfer M</th>
<th>Transfer SD</th>
<th>Speaker rating M</th>
<th>Speaker rating SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human voice</td>
<td>4.7*</td>
<td>1.3</td>
<td>4.2*</td>
<td>1.6</td>
<td>4.9*</td>
<td>0.9</td>
</tr>
<tr>
<td>Machine voice</td>
<td>3.9</td>
<td>1.3</td>
<td>2.9</td>
<td>1.7</td>
<td>3.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

* Human-voice group scored significantly higher than machine-voice group at p < .05.

Conclusion

Theoretical Implications

We focused on problem-solving transfer as the key measure of student understanding of a scientific explanation. Across two experiments, students performed better on problem-solving transfer when the voice in the multimedia message was from a human speaking with a standard accent rather than a human speaking with a foreign accent (Experiment 1) or a machine voice (Experiment 2). These results are highly inconsistent with the predictions of
cognitive effort theory and allow us to reject it. These results are highly consistent with the predictions of social agency theory and cognitive load theory, allowing us to consider the possibility that both might be correct.

Additional support for social agency theory comes from survey ratings in which students rated the speaker higher on social dimensions when the voice was from a human speaking with a standard accent than from a human speaking with a foreign accent (Experiment 1) or a machine voice (Experiment 2). Taken together, the transfer and speaker-rating results suggest that voice can be a factor in creating a sense of social presence in which learners would be more likely to interpret the computer-based narrator as a social partner. Furthermore, students try harder to understand and therefore learn more deeply from a multimedia message when the speaker primes a social conversation schema in the student.

Additional support for cognitive load theory comes from the difficulty ratings in which students in the human-voice group rated the lesson as easier than did students in the machine-voice group (Experiment 2). On the basis of cognitive load theory, we can interpret the negative effects of foreign-accented voice and machine voice as creating extra cognitive load. Learners must devote more cognitive capacity to deciphering the incoming narration, and therefore they have less cognitive capacity leftover to build connections among pieces of information as needed for meaningful learning.

Given that the efficacy of cognitive load theory has been well established in multimedia research (Mayer, 2001; Sweller, 1999), a major new contribution of the present study concerns the potential viability of social agency theory as an adjunct to cognitive load theory. In our view, cognitive load theory and social agency theory are not mutually exclusive, because both cognitive and social factors can contribute to how students learn from multimedia messages. Further research is needed to untangle the relative contributions of cognitive and social factors, such as by comparing multimedia messages in which one speaker has a more socially desirable foreign accent and another speaker has a less socially desirable foreign accent or in which one speaker has a machine voice without stress on appropriate syllables and another speaker has a machine voice with human-like stress on appropriate syllables. These studies would allow a comparison of social features of the speaker while keeping the cognitive features the same (e.g., both are accented or both are machine voices).

The present study represents an instantiation of the media equation in the domain of computer-based instruction and shows that modest social cues—the voice of the speaker—can have major effects (with effect sizes for transfer of at least 0.80). Although previous studies have examined how voice affects the users’ feelings of attraction (e.g., Nass & Lee, 2001), the present set of studies is the first documented evidence for voice effects on deep learning in multimedia environments. However, additional work is needed to pinpoint which aspects of voice are most important in promoting deep learning.

In this article, we have proposed a social mechanism—social agency theory—that might work in conjunction with a cognitive mechanism—such as cognitive load theory (Chandler & Sweller, 1991; Sweller, 1999)—to explain some multimedia effects. Should we abandon social agency theory in favor of cognitive load theory? In response, we note that it is possible that both social and cognitive forces are at work, with social cues affecting how much cognitive energy is put into the task of understanding the multimedia explanation and cognitive load affecting much how of that cognitive energy is diverted away from meaningful cognitive processing. The transfer results of both experiments are consistent with both theories. The speaker-rating results in both experiments are consistent with the social agency theory, and the difficulty ratings in Experiment 2 are consistent with cognitive load theory. Thus, the results of our voice studies give us a rationale to propose social agency theory as a complement to cognitive theory, which has been established in other multimedia learning tasks (Mayer, 2001; Sweller, 1999).

**Practical Implications**

This work allows us to offer a newly discovered principle for the design of multimedia instructional messages. The *voice principle* is that students learn more deeply when the narration in a multimedia lesson is spoken by a standard-accented human voice rather than a foreign-accented human voice or a machine voice. When designing multimedia messages, designers should consider the role of social cues such as the speaker’s voice. This prescription is particularly relevant to the design of animated pedagogical agents (Lester et al., 2000; Lester, Voerman, Towns, & Callaway, 1999; Moreno, Mayer, Spires, & Lester, 2001; Rickel & Johnson, 1999, 2000).

Our research extends the media equation into the realm of multimedia instructional episodes. Overall, it appears that the voice used in computer-based multimedia lessons can have an important effect on learners’ conceptions of the speaker and ultimately on their learning outcomes. Therefore, in building virtual learning environments, instructional designers should choose voices—such as human voices with standard accents—that prime a sense of social partnership in learners. However, this research should not be taken to denigrate the instructional effectiveness of instructors who have foreign accents because more sophisticated factors are involved in students’ sense of social partnership with real instructors.

**References**


Appendix

Narration Script for Lightning Lesson in Experiments 1 and 2

Cool, moist air moves over a warmer surface and becomes heated. Warmed, moist air near the earth’s surface rises rapidly. As the air in this updraft cools, water vapor condenses into water droplets and forms a cloud. The cloud’s top extends above the freezing level, so the upper portion of the cloud is composed of tiny ice crystals. Eventually, the water droplets and ice crystals become too large to be suspended by the updrafts. As the raindrops and ice crystals fall through the cloud, they draw some of the air in the cloud downward, producing downdrafts. When downdrafts strike the ground, they spread out in all directions, producing the gusts of cool wind people feel just before the start of the rain. Within the cloud, the rising air currents cause electrical charges to build. The charge results from the collision of the cloud’s rising water droplets against heavier, falling pieces of ice. The negatively charged particles fall to the bottom of the cloud, and most of the positively charged particles rise to the top. A stepped leader of negative charges moves downward in a series of steps. It nears the ground. A positively charged leader travels up from such objects as trees and buildings. The two leaders generally meet about 165 feet above the ground. The negatively charged particles then rush from the cloud to the ground along the path created by the leaders. It is not very bright. As the leader stroke nears the ground, it induces an opposite charge, so positively charged particles from the ground rush upward along the same path. This upward motion of the current is the return stroke. It produces the bright flash that people notice as a flash of lightning.

Received August 9, 2001
Revision received January 14, 2002
Accepted March 13, 2002