Fostering social agency in multimedia learning: Examining the impact of an animated agent’s voice

Robert K. Atkinson\textsuperscript{a,*}, Richard E. Mayer\textsuperscript{b}, Mary Margaret Merrill\textsuperscript{c}

\textsuperscript{a} Division of Psychology in Education, Arizona State University, P.O. Box 870611, Tempe, AZ 85287-0611, USA
\textsuperscript{b} University of California, Santa Barbara, USA
\textsuperscript{c} Louisiana State University-Shreveport, USA

Available online 25 September 2004

Abstract

Consistent with social agency theory, we hypothesized that learners who studied a set of worked-out examples involving proportional reasoning narrated by an animated agent with a human voice would perform better on near and far transfer tests and rate the speaker more positively compared to learners who studied the same set of examples narrated by an agent with a machine synthesized voice. This hypothesis was supported across two experiments, including one conducted in a high school computer classroom. Overall, the results are consistent with social agency theory that posits that social cues in multimedia messages, including the type of voice, can affect how much students like the speaker and how hard students try to understand the presented material.

© 2004 Elsevier Inc. All rights reserved.

* Corresponding author. Fax: +1 480 965 7193.
E-mail address: robert.atkinson@asu.edu (R.K. Atkinson).

The research reported in this study was supported by the Office of Naval Research, Office of Naval Research, Cognitive Sciences Division, under Grant N00014001-1-0459 awarded to Robert K. Atkinson.
1. Introduction

Presently, animated pedagogical agents—human-like characters that provide instruction through verbal and nonverbal modes of communication—are being used by multimedia instructional designers to create simulated human-to-human connections between learners and computers that are intended to help learners accept computers as social partners (Cassell, Sullivan, Prevost, & Churchill, 2000). The available empirical research suggests that animated pedagogical agents nested within multimedia learning environments can enhance the learners’ ability to transfer what was learned to new situations as well as to increase their enjoyment of working with the learning tutorial (Atkinson, 2002; Johnson, Rickel, & Lester, 2000; Mayer, Sobko, & Mautone, 2003; Moreno, Mayer, Spires, & Lester, 2001) while not producing any split-attention effects (Craig, Gholson, & Driscoll, 2002). Despite the potential benefits of employing animated pedagogical agents to visually and aurally guide learners through computer-based learning environments, almost no empirical research has examined the specific effect of an agent’s voice. Thus, the current study examines whether the nature of an animated pedagogical agent’s voice can have an impact on the social connection between learners and computers, and ultimately, on the process and outcome of learning.

2. Social agency theory

One theoretical framework for considering the effectiveness and utility of fostering simulated human-to-human connections in multimedia learning environments is social agency theory (Mayer et al., 2003; Moreno et al., 2001). According to this theory, multimedia learning environments can be designed to encourage learners to operate under the assumption that their relationship with the computer is a social one, in which the conventions of human-to-human communication apply as described by Reeves and Nass (1996). Essentially, the theory posits that the use of verbal and visual social cues in computer-based environments can foster the development of a partnership by encouraging the learners to consider their interaction with the computer to be similar to what they would expect from a human-to-human conversation. For instance, across several empirical studies, Nass and his colleagues (Moon & Naas, 1996; Naas, Moon, Fogg, Reeves, & Dreyer, 1995) documented that individuals react to various verbal social cues present in a text-based exchange with a computer as if the social cues originated from a human. Alternatively the environment might rely on spoken verbal social cues, such as a standard accented voice, or visual social cues, such as an animated agent that utilizes dynamic non-verbal signals (e.g., gaze, gesture, and facial expressions), to encourage learners to approach this situation as if they are engaged in a human-to-human conversation.

Once this social partnership is established, learners can rely on several basic human-to-human social rules that guide their interaction with the multimedia learning environment (Mayer et al., 2003). According to Grice (1975), these social rules include the cooperation principle and its four associated maxims. Specifically, Grice
proposed that in human-to-human conversations, an individual listening to another person speaking will assume that he or she is making a concerted effort to make sense by being informative, accurate, relevant, and concise. Thus, the assumption of social agency theory is that learners will assume that the speaker in the multimedia learning environment—like a typical human speaker—is attempting to make sense.

According to social agency theory, priming the social interaction schema will cause the learner to try to understand and deeply process the computer’s instructional message concerning academic subject matter. Mayer (1999, 2001) has posited that the cognitive processes that learners employ in order to understand an instructional message include: (a) selection of relevant information, (b) organization of patterns of information, and (c) the integration of prior knowledge with newly presented information. The ability to process information with deep levels of understanding—that is, to engage in sense making processes—will affect whether the learner is able to transfer what was learned to related problem solving endeavors.

Additionally, social agency theory seeks to determine the conditions under which learners interpret their interaction with a computer-based learning environment. Specifically, do learners perceive their computer experiences as an instance of social communication or information delivery? The difference between the two descriptions of human–computer interactions affects the learner’s schema activation, levels of cognitive processing, and the quality of learning that takes place. Learners may perceive an interaction as social if they are able to receive the social cues necessary to form a simulated human-to-human conversation with the computer—cues that we posit are provided by friendly on-screen agents who speak in a human voice. Perceiving the computer as a social partner encourages the learner to engage in a sense making process that increases the probability of positive transfer (Mayer et al., 2003).

2.1. Promoting social agency with speaker’s voice

In addition to examining the role of animated agents as social cues in multimedia learning environments, researchers have recently investigated the role of a speaker’s voice as a social cue by varying the nature of the speaker’s voice in a short, 2-min multimedia instructional program designed to convey conceptual information about a science topic, specifically lightning formation (Mayer et al., 2003). Specifically, Mayer et al. were interested in examining the relationship between the nature of a speaker’s voice—whether it was socially appealing or not—and the learner’s attribution of social agency. In the first of two experiments, the narration consisted of a male voice-over (i.e., disembodied voice) with either a standard accented speech—that is, a native speaker of standard American English—or a foreign accented speech—in this case, a non-native speaker of standard American English, one with a Russian accent. The participants in the standard accent condition scored better on a learning transfer test, which required them to solve new problems, than the participants in the foreign accent group, yielding a Cohen’s $d$ statistic of .90 (a large effect). Moreover, participants who listened to the standard accented voice rated the narrator more positively than the participants who listened to the foreign accented voice according to a 15-item instrument adapted from Zahn and Hopper’s (1985)
Speech Evaluation Instrument. Moreover, on the basis of a supplemental experiment, the authors concluded that pattern of results they observed could not be attributed to differences in the intelligibility of the two types of voices since the participants in the standard-accent condition and the foreign-accent condition discerned effectively the same number of words.

In a second experiment, Mayer et al. (2003) compared the social appeal of a human voice to that of a machine synthesized voice. Forty college students were randomly assigned to either a human voice (a male, native speaker of standard American English) group or a machine voice (a male, computer-generated voice) group. Results indicated the participants in the human voice group scored statistically significantly higher on learning performance tests than the machine voice group, yielding a Cohen’s $d$ statistic of .79 (a medium-to-large effect). Participants, as the research suggested, also ascribed more positive social characteristics to the human voice.

In sum, the Mayer et al. (2003) study supports several conclusions regarding the use of voice, specifically a male voice with a standard accent, to support social agency in a multimedia learning environment. First, their research supports the prediction based on social agency theory that participants assigned to the standard accent group would outperform their peers in the foreign accented voice condition on measures of transfer and would rate the speaker’s voice more positively. Second, Mayer et al.’s research supports the notion that a human voice can enhance the process and outcome of learning relative to a machine synthesized voice by providing strong social cues through the use of a familiar, socially appealing voice.

The Mayer et al. study can be criticized on the grounds that the human–machine voice effect is based on a single experiment involving an extremely short presentation (i.e., lasting approximately 2 min) in an artificial laboratory setting. The present study seeks to determine whether the voice effect will occur with a longer, more typical lesson in a realistic classroom setting with high school students as well as a laboratory setting with college students.

Additional evidence for the notion that human voice is associated with larger learning gains than machine synthesized voice can be inferred from a recently conducted study by Graesser et al. (2003). They examined the medium of presentation (i.e., text-only, voice-only, agent + voice, agent + voice + text) in the context of an intelligent tutoring system called AutoTutor designed to improve computer literacy among college students. The voice was machine synthesized using the same software for generating machine synthesized voice as we used in the present study. In contrast to the well-documented modality effect when the voice is human (Atkinson, 2002; Mayer & Moreno, 1998; Mousavi, Low, & Sweller, 1995), Graesser et al. (2003) found no modality effect when the voice in an intelligent tutoring system is machine synthesized. Apparently, the advantage of speech over text is lost when speech does not convey a human quality.

2.2. Promoting social agency with animated agents

In a typical educational setting, a social exchange—including verbal and nonverbal interaction—can naturally occur between a teacher and learner in conjunction
with the presentation of academic material. However, when a learner is engaged in a computer-based learning episode, the opportunity for a social exchange between the learner and the learning environment is often times nonexistent (Mayer et al., 2003). Recently, Moreno et al. proposed a solution to this problem by incorporating animated pedagogical agents into multimedia learning environments in an effort to foster the development of a social relationship between learners and computers. According to the social agency theory, the combination of a multimedia learning environment and an animated agent elicits verbal and visual social cues that create virtual relationships between agents and learners as substitutes for authentic human-to-human interactions—interactions that possess the social properties employed in a human conversation. Moreover, animated agents assume the role of a human teacher giving instruction and feedback as the learner acquires and processes new information. Social agency theory stipulates that the life-like characteristics and behaviors of an animated agent prompt the social engagement of the learner, thus allowing the learner to form a simulated human bond with the agent.

Recent research focusing on the utility of animated agents has provided developers of computer-based learning environments with a means of incorporating motivational and life-like characters to aid in the knowledge and skill acquisition of learners. In theory, an embodied animated agent, with its humanistic communication capabilities, is able to direct a learner’s attention to the appropriate element of a problem-solving task using gestures, gaze, and locomotion. Moreover, multimedia learning environments incorporating animated pedagogical agents offer key features that traditional tutoring programs seem to lack. For instance, animated agents offer the potential to enrich and broaden the communicative relationship between learners and computers as well as provide computers with motivational and affective instructional features that actively engage students (Johnson et al., 2000). Additionally, simply having an animated agent present in a multimedia learning environment can positively influence the learner’s perceptions of their educational experience (Lester et al., 1997). It has been proposed that the combination of an interesting animated agent and a well-structured learning environment can optimize a learner’s active engagement with the task and increase the probability of future interactions with the instructional program (Johnson et al., 2000).

In a recent study, Atkinson (2002) examined whether the presence of an animated agent in a multimedia-based learning environment designed to teach learners how to solve word problems would enhance the process and outcome of learning. Specifically, Atkinson examined whether the delivery method of instructional elaborations (i.e., aurally or textually) in conjunction with the presence or absence of an embodied animated agent—one capable delivering explanations aurally while using non-verbal cues (e.g., gesture, gaze) to direct a learner’s attention—had an affect on learning outcome measures. Findings indicated that the participants who were exposed to the agent in combination with narrated instructions achieved higher scores on both near and far transfer tests than the control participants who were not exposed to an animated agent (i.e., voice-only or text-only). Subsequently, Atkinson attempted to replicate the initial study by placing students in mixed (voice-plus-agent) or single (voice-only or text-only) modality conditions to receive instructional elaborations.
regarding mathematics word problems. Again, students receiving instructions verbally from an agent outscored their peers in the textual condition on near transfer, and outscored both the voice-only and text-only conditions in terms of far transfer performance. Presumably, an interactive relationship between a learner and a surrogate tutor was enabled by the presence of an animated agent with the capacity to narrate explanations of the instructions to the participant.

To explore whether learners will report an increased interest in learning and achieve better transfer performance if they experience a simulated human-to-human connection with the computer via an animated agent, Moreno and her colleagues (Moreno et al., 2001) conducted a series of experiments regarding the presence or absence of an animated agent in conjunction with the delivery of instructions through speech or on-screen text. Across five experiments, learners were asked to work with Design-A-Plant, a computer-based learning program in which they were expected to design a plant from a library of plant structures (e.g., roots, stems, and leaves) that could thrive under specified environmental conditions. In the initial experiment, undergraduate college students who received instruction via an animated agent (i.e., Herman the Bug) scored significantly higher on complex transfer problems than did students who received the same verbal and visual instructional material without the agent. Moreover, participants in the agent group reported an increased interest in the material and a greater willingness to continue interactions with the program. Findings from additional experiments, including one with school-age children, supported the usage of an animated pedagogical agent in conjunction with spoken instruction as a tool for optimizing learning. Thus, this study capitalized on one of the chief premises of the social agency theory, that is, bringing together verbal and visual modalities of instruction with human-like features increases the likelihood that meaningful learning gains can occur through the mediation of a surrogate instructor. Consequently, Moreno et al.’s research provides evidence that a learner can capitalize on a social partnership with an on-screen animated agent, a partnership that can foster both the process and outcome of learning.

3. Overview of experiments

The research on pedagogical agents supports the prediction based on social agency theory that animated on-screen agents are better able to promote social agency in multimedia learning environments than a text-only or voice-only environment. Moreover, research on the role of a narrator’s voice supports the hypothesis that the type of voice can have an impact on social agency. In the present experiments, we examined the impact of an agent’s voice in a realistic mathematics lesson.

Across two experiments, participants received a narrated set of worked-out examples for proportional reasoning word problems spoken by a female native-English speaker (human voice condition) or by a female machine-synthesized voice (machine voice condition). Although we did not expect age or even computer familiarity to be a major factor in the present study, we suspected that proficiency in solving proportional reasoning problems was an important factor. For this reason, we focused on
learners who were not highly proficient on the task but were capable of learning it, namely high school students and beginning-level college students. Experiment 1 was conducted in a university-based computer laboratory with college undergraduates. Experiment 2 was conducted in a computer classroom with high school students. Both learning process and learning outcome measures were collected. The learning process measures included perceived example understanding, perceived example difficulty, and performance on practice problems. The learning outcome measures included a posttest, which contained both near and far transfer items, and a speaker-rating questionnaire designed to detect the social characteristics attributed to speakers. This latter measure was identical to the one used by Mayer et al. (2003).

Overall, this study was designed to provide an important extension of preliminary research conducted by Mayer et al. (2003) by (a) using a new learning environment (i.e., one that relies on an embodied animated agent as the source of aural instructional support rather than a disembodied voice-over), (b) teaching a new type of material (i.e., procedural knowledge rather than conceptual knowledge), (c) incorporating a new domain (i.e., math rather than science), (d) using a new population (i.e., high school students rather than college students), (e) extending the length of the instructional episode (i.e., 40 min rather than 2 min), (f) extending the length of the time to complete posttest (i.e., unlimited time rather than fixed time), (g) incorporating new dependent measures (i.e., performance on practice problems, near and far transfer problem solving), (h) relying on a new independent variable (i.e., female voices rather than male voices), and most importantly, and (i) using an authentic educational context (in Experiment 2 of the present study) rather than a lab setting.

4. Experiment 1: Multimedia learning in a laboratory setting

In this experiment, students received a computer-based mathematics lesson that provided four worked-out examples along with step-by-step descriptions of how to solve them. Narration accompanying the on-screen examples was presented in a human voice or a computer-synthesized voice (machine voice). According to social agency theory, students in the human voice group should produce higher scores than students in the machine voice group on the practice problems and on the near and far transfer tests designed to measure the depth of learner understanding, and rate the speaker more positively while, at the same time, not rating the examples as any more difficult or reporting any differences in understanding than their machine voice counterparts.

4.1. Method

4.1.1. Participants and design

The participants were 50 undergraduate college students recruited from educational psychology courses at Mississippi State University. They were randomly assigned in equal numbers to one of two conditions, with 25 serving in the human voice group and 25 serving in the machine voice group. The percentage of males
was 80% in the human voice group and 84% in the machine voice group; the percentage of juniors and seniors was 80% in the human voice group and 76% in the machine voice group; the percentage of students majoring in education or educational psychology was 80% in the human voice group and 72% in the machine voice group; and the mean GPA was 3.00 for the human voice group and 2.93 for the machine voice group.

4.1.2. Computer-based learning environment

The computer-based materials consisted of two versions (i.e., human voice and machine voice) of a multimedia training program on how to solve proportional reasoning word problems. The training program was created using Director 8.0 (Macromedia, 2000) coupled with Microsoft Agent and XtrAgent 2.0 for deployment within a Windows-based operating system, and was based on an earlier program (Atkinson, 2002; Atkinson & Derry, 2000).

The learning environment, which was 800 by 600 pixels in size, included an instruction pane—for displaying the instructions for the current problem (see top left of Fig. 1), a problem text pane—for displaying the problem on which the worked example was based (see middle left of Fig. 1), a control panel—allowing the user to proceed through the instructional sequence at his/her own pace (see bottom left of Fig. 1), a workspace—for displaying the solution to the example’s problem (see Fig. 1. A frame from the instructional program used in Experiments 1 and 2.
right side of Fig. 1), a calculator (see middle right of Fig. 1), and an animated agent in the form of a parrot named Peedy—an agent capable of 75 animated behaviors, including behaviors specifically designed to direct attention to objects on the screen, such as gesturing and/or looking in specific directions (e.g., up, down, left, and right). The agent was created from several off-the-shelf pieces of software, including Microsoft Agent, a collection of programmable pieces of software designed to support the presentation of the animated agent and XtrAgent 2.0, used to animate the agent within a Director-based learning environment.

The instructional materials presented in the program consisted of four example/practice problem pairs, where each worked example was followed by an isomorphic practice problem. For example, one of the worked examples was the “Bill’s Hometown Furniture Store” problem:

Bill’s Hometown Furniture Store creates custom-ordered furniture. Bill, the owner, received an order for 12 identical kitchen cabinets last week. Bill hired four carpenters to work for five days, and they made 7 cabinets in that time. However, one of the carpenters broke his arm over the weekend and, as a result, will be unable to help finish the order. If Bill has the three healthy carpenters complete the remaining cabinets, how long will it take them to finish the job?

The worked examples were structured to consist of a sequential presentation of problem states and to emphasize problem subgoals. Unlike examples that simultaneously display all of the solution components (i.e., simultaneous examples), the sequential examples used in the present experiment appeared initially unsolved. The learning environment was structured to permit the learner to proceed through each example and watch as problem states were successively added over a series of pages until the final page in the series presented the solution in its entirety. Each solution step was coupled with instructional explanations delivered orally that were designed to underscore what was occurring in that step (e.g., “First, we need to set up a proportional relationship to determine the production rate”). The examples also relied on two explicit cues—the visual isolation and labeling of each subgoal (e.g., “Total Amount 1”)—to clearly demarcate a problem’s subgoals.

Moreover, the learning environment was configurable to run in one of two instructional modes that reflected the two conditions of the present experiment. Since the animation service provided by Microsoft Agent permits audio files to be used for a character’s spoken output, Peedy was programmed in the human voice condition to deliver recoded audio files consisting of instructional elaboration created by a human tutor designed to highlight what is occurring in each of the example’s sequentially presented solution step. These audio files were created by a 29-year-old female graduate student who spoke with a standard North American English accent. The software automatically synchronized Peedy’s mouth to the human tutor’s voice by using the characteristics of the audio file. The machine voice condition was identical in every respect to the human voice condition with one exception: Instead of using voice files containing a human voice to
deliver instructional explanations in the examples, the Lernout and Hauspie TruVoice TTS text-to-speech engine (http://www.microsoft.com/msagent/downloads/user.asp#tts)—a computer-based system able to read text aloud—provided by Microsoft delivered the instructional elaborations orally in North American English. Specifically, the learner in this condition listened to “Mary,” a machine-generated voice based on a 30-year-old female (Model ID #c77c5170-2867-11d0-847B-44455354000) who delivered, along with the presence of an agent’s image, the exact same instructional explanations that were used to highlight the solution steps in the human voice condition. Regardless of which instructional mode the learning environment was configured to employ, Peedy was programmed to move around the workspace, using gesture and gaze to highlight the example’s solution (see right side of Fig. 1).

Following each worked example, two questions—one focused on perceived example understanding and the other addressing perceived example difficulty—were presented to the learners on the computer screen. First, they were asked to respond to the statement “I understood the worked example just presented to me” by selecting a reaction on a balanced five-point rating scale that ranged from “very much agree” (1) to “very much disagree” (5). Second, they were presented with an item adapted from an instrument used by Paas and Van Merriënboer (1993) designed to measure participants’ perceived cognitive load. Specifically, they were asked “please rate the difficulty of the worked example just presented” by selecting a response on a balanced five-point rating scale that ranged from “very easy” (1) to “very difficult” (5).

After rating their understanding of the example and how difficult they perceived it to be, a practice problem was presented on the computer screen that was parallel in structure to the example itself. For example, the practice problem coupled with the “Bill’s Hometown Furniture Store” problem is the following:

A local high school needs 120 classrooms painted over the summer. They hired 5 painters who worked for six days and completed 49 classrooms. Due to a conflict with management, however, 3 painters quit after 6 days of work. If the 2 remaining painters finish the job, how long will it take them to finish painting the classrooms?

The learner was required to enter a response to the practice problem before he or she was given the final answer to the problem. The answers to the practice problems did not include solutions to problem steps or any explanation about the solution. During the presentation of each practice problem, Peedy disappeared and only returned when the subsequent example was presented.

The computer-based environment was deployed on a total of eight Gateway E-1200 computer systems (600mHz, 256 RAM), each equipped with 15-in color monitors and Optimus Nova 80 headphones.

4.1.3. Pencil–paper materials

The paper materials consisted of a participant questionnaire, an 8-page mathematics review booklet, a 15-item speaker survey, and a posttest consisting of four
near transfer items and four far transfer items. The review booklet and the transfer
tests were adopted from Atkinson (2002). The 15-item speaker survey was adopted
from Mayer et al. (2003). The participant questionnaire solicited information con-
cerning the participant’s demographic characteristics including gender and academic
major. The mathematics review booklet provided a brief review of how to solve sim-
ple one-step proportional reasoning word problems; it included three problems that
the participants were encouraged to try, followed by step-by-step descriptions of the
correct solution procedure.

The speaker rating survey was a 15-item instrument adapted from Zahn and Hopper’s (1985) Speech Evaluation Instrument. We adapted Zahn and Hopper’s (1985) speech evaluation instrument because of its effectiveness in detecting the social char-
acteristics attributed to speakers. Instructions at the top of the page asked the par-
ticipant 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: literate–illiterate, unkind–kind, active–passive,
intelligent–unintelligent, cold–warm, talkative–shy, uneducated–educated, friendly–
unfriendly, unaggressive–aggressive, fluent–not fluent, unpleasant–pleasant, confident–
unsure, inexperienced–experienced, unlikeable–likeable, and energetic–lazy. There were 5 items from each of 3 subscales—superiority, attractiveness, and dyna-
mism. According to Zahn and Hopper, superiority “… combines intellectual status
and competence, social status items, and speaking competency items,” attractiveness
captures the social and aesthetic appeal of a speaker’s voice, and dynamism charac-
terizes a speaker’s “… social power, activity level, and the self-presentational aspects
of [his or her] speech” (p. 119).

The near transfer items on the posttest consisted of four proportional reasoning
word problems that were structurally identical to one of the problems presented dur-
ing instruction albeit they had different surface stories. For example, the following
near transfer item is structurally isomorphic to the “Bill’s Hometown Furniture
Store” problem used during instruction:

Mike, a wheat farmer, has to plow 2100 acres. He rented six tractors with peo-
ple to drive for 3.75 days, and they completed 1200 acres. If he rents four trac-
tor/drivers, how long will it take them to complete the plowing?

The far transfer items on the posttest consisted of four proportional reasoning
word problems that were not structurally identical to any of the problems presented
during instruction. For example, the following far transfer problem is not isomor-
phic to the “Bill’s Hometown Furniture Store” problem or any other problem pre-
sented during instruction:

Brian is selling newspapers at a rate of 3 newspapers every 10 minutes on one
side of a downtown street, while Sheila at her newsstand across the street is sell-
ing papers at the rate of 8 newspapers every 20 minutes. If they decide to go
into business together, how many newspapers will they sell in 40 minutes at
these rates?
In accord with classic definitions of problem-solving transfer (Mayer & Wittrock, 1996), the near transfer problems require executing the same solution method as was taught, but applying it to a new problem, whereas far transfer problems require adjusting the solution method that was taught, and applying it to a new problem.

To help control for a possible order effect, four versions of the posttest test were constructed. Within each version, the near and far transfer problems were randomly ordered.

4.1.4. Procedure

In this experiment, participants learned in a single session by working independently in a laboratory containing eight workstations, each located in its own cubicle. During this session, the participants filled out a demographic questionnaire, and then read through the eight-page review on solving proportion problems. When participants completed the review on proportion problems, they began the computer-based lesson in which they studied the four example/practice problem pairs. Based on random assignment, some participants received a program that had a human voice to explain the worked examples whereas others received a program that had a machine voice. Each of the four example/practice problem pairs consisted of a condition-specific worked example along with a paired isomorphic practice problem presented on the computer monitor. The learners were asked to solve the practice problem on paper and then check the accuracy of their solutions using the solution presented in the learning environment. After instruction, the participants were administered the eight-item pencil–paper posttest, which took approximately 50 min to complete. The speaker survey was administered after the posttest, which took approximately 5 min to complete.

4.1.5. Scoring

The two measures collected after each example was presented—perceived example understanding and perceived example difficulty—were scored in the same fashion. The participants’ responses to each of these queries were summed across all four examples and divided by four, thereby generating a measure of average perceived example understanding and a measure of average perceived example difficulty, both with values ranging from 1 to 5.

The protocols generated during practice problem solving as well as the near and far transfer tests were coded for conceptual accuracy according to a set of guidelines for analyzing the written problem-solving protocols designed to help determine where the learner fell along a problem-comprehension continuum. According to these guidelines, each item—the four practice problems, the four near transfer, and the four far transfer—was awarded a conceptual score, ranging from 0 to 3 depending upon the degree to which the participant’s solution was conceptually accurate (e.g., 0 = no evidence of the student understanding the problem; 3 = there is perfect understanding of the problem, ignoring minor computational/copying errors, and the student used a complete and correct strategy to arrive at an answer). For all three measures (i.e., performance on practice problems, near transfer, and far transfer), 12 was the maximum score that a learner could achieve (e.g., 3
points-per-problem × 4 items). To create an average conceptual score, with values ranging from 0 to 3, the conceptual scores awarded on each measure were summed across all four items and divided by 4. Internal consistency reliabilities (Cronbach’s \( z \)) for the practice problem, near transfer, and far transfer measures were .82, .77, and .76, respectively.

One research assistant who was unaware of the participants’ treatment conditions independently scored each problem-solving protocol. To validate the scoring system, a second rater also unaware of the participants’ treatment conditions independently scored a random sample of 20% of the problem-solving protocols. The scores assigned by the two raters to reflect the conceptual accuracy of the participants’ responses across all three measures were consistent 96% of the time. Discussion and common consent were used to resolve any disagreement between coders.

Finally, an overall speaker rating (from 1 to 8) was constructed. This was accomplished by averaging the scores from the three subscales (i.e., superiority, attractiveness, and dynamism); with 1 indicating the most positive rating and 8 indicating the most negative rating. Internal consistency reliability (Cronbach’s \( z \)) for this measure was .90.

### 4.2. Results and discussion

The major research question addressed in this experiment concerned whether learners in the human voice condition reported increased interest in learning and achieved better transfer than learners in the machine voice condition. Table 1 shows the mean score (and standard deviation) for each group in Experiment 1 on the perceived example understanding, perceived example difficulty, performance on practice problems, near transfer test, far transfer test, the speaker rating survey, and instructional time. Separate two-tailed \( t \) tests were conducted on these measures, each at

<table>
<thead>
<tr>
<th></th>
<th>Human voice</th>
<th>Machine voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived example understanding</td>
<td>1.32</td>
<td>1.31</td>
</tr>
<tr>
<td>Perceived example difficulty</td>
<td>2.21</td>
<td>1.99</td>
</tr>
<tr>
<td>Performance on all four practice problems</td>
<td>2.67&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.09</td>
</tr>
<tr>
<td>Posttest—near transfer</td>
<td>2.23&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.62</td>
</tr>
<tr>
<td>Posttest—far transfer</td>
<td>1.32&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.77</td>
</tr>
<tr>
<td>Speaker rating survey</td>
<td>2.29&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.10</td>
</tr>
<tr>
<td>Instructional time (Time-on-Task)</td>
<td>39.2</td>
<td>40.4</td>
</tr>
<tr>
<td>Performance on practice problems 1 and 2</td>
<td>2.66&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.10</td>
</tr>
<tr>
<td>Performance on practice problems 3 and 4</td>
<td>2.68&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.08</td>
</tr>
</tbody>
</table>

<sup>a</sup> Denotes human voice group scored significantly higher than machine voice group at \( p < .05 \); \( n = 25 \) for each group; instructional time is reported in minutes.
\[ z = .05. \] Cohen’s \( d \) statistic was used as an effect size index where \( d \) values of .2, .5, and .8 correspond to small, medium, and large values, respectively (Cohen, 1988).

4.2.1. Does voice affect perceived example understanding?
As revealed in the first row of Table 1, the human voice group (\( M = 1.32, SD = .45 \)) and the machine voice group (\( M = 1.31, SD = .35 \)) did not statistically differ in terms of perceived example understanding, \( t(48) = .87, p = ns \). In sum, the results show that learners reported that the examples were relatively easy to understand, regardless of which voice—human or machine—accompanied the examples.

4.2.2. Does voice affect perceived example difficulty?
As illustrated in the second row of Table 1, the perceived example difficulty (i.e., cognitive load) reported by the participants assigned to the human voice condition (\( M = 2.21, SD = .75 \)) did not differ significantly from the mean ratings of the machine voice group (\( M = 1.99, SD = .58 \)), \( t(48) = 1.16, p = ns \). In general, the results reveal that participants presented with either the human voice or the machine voice perceived the examples to be moderately difficult, with no statistically significant difference between the two conditions.

4.2.3. Does voice affect performance on practice problems?
As shown in the third row of Table 1, the scores associated with the practice problems for participants in the human voice group (\( M = 2.67, SD = .63 \)) were significantly higher than those of their peers in the machine voice group (\( M = 2.09, SD = .85 \)) on, \( t(48) = 2.74, p < .01 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .79 for practice problem-solving performance, which corresponds to a medium-to-large effect. Overall, the results show that human voice fostered better understanding of how to solve the practice problems that accompanied the examples during instruction than did machine voice.

4.2.4. Does voice affect near transfer?
As shown in the fourth row of Table 1, the human voice group (\( M = 2.23, SD = .71 \)) scored significantly higher than the machine voice group (\( M = 1.62, SD = .77 \)) on the near transfer test, \( t(48) = 2.91, p < .01 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .84 for near transfer, which corresponds to a large effect. In general, the results show that human voice fostered better understanding of how to solve problems like those presented during instruction than did machine voice.

4.2.5. Does voice affect far transfer?
As shown in the fifth row of Table 1, the human voice group (\( M = 1.32, SD = .90 \)) scored significantly higher than the machine voice group (\( M = .77, SD = .67 \)) on the far transfer test, \( t(48) = 2.46, p < .05 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .71 for far transfer, which corresponds to a medium-to-large effect.
Overall, the results show that human voice fostered deeper understanding of how to solve problems that were not like those presented during learning than did machine voice.

4.2.6. Does voice affect speaker rating?

As shown in the sixth row of Table 1, the human voice group ($M = 2.29$, $SD = .84$) rated the speaker significantly more favorably than did the machine voice group ($M = 3.10$, $SD = 1.30$) on the speaker rating survey, $t(48) = 2.64$, $p = .01$. Cohen’s $d$ statistic for these data yields an effect size estimate of .76 for speaker rating, which corresponds to a medium-to-large effect. Overall, students in the human voice condition reported a more positive evaluation of the speaker’s attractiveness, dynamism, and superiority than did the machine voice group.

4.2.7. Related issues

To examine whether the pattern of results could be attributed to time-on-task, we compared time spent on the instructional program across the two conditions. We concluded that the pattern of results cannot be attributed to the human voice group spending more time during learning than the machine voice group, since the instructional time for the machine voice group ($M = 40.4$ min, $SD = 17.1$ min) was not significantly different from the human voice group ($M = 39.2$ min, $SD = 9.2$ min), $t(48) = .30$, $p = ns$.

To explore the possibility that the pattern of results could be attributed to a “novelty effect,” we examined the performance on the practice problems in the first half versus the second half of training. According to a novelty effect, there should be large differences between machine and human voices for early practice problems but not later practice problems. That is, at the outset of instruction, the learners might be distracted by the machine voice to the point of decreasing attention to the content (resulting in lower practice problem performance) before adjusting to the machine voice, and thereafter the machine and human voices become equivalent. To test for this possibility, we calculated an average score on the first two and on the last two practice problems for each group. The averages associated with the first two practice problems for participants in the human voice group ($M = 2.66$, $SD = .67$) were significantly higher than those of their peers in machine voice group ($M = 2.10$, $SD = .97$) on, $t(48) = 2.38$, $p < .05$ (Cohen’s $d = .67$). The same pattern emerged for the last two practice problems where the participants in the human voice group ($M = 2.68$, $SD = .64$) were significantly higher than those of their peers in machine voice group ($M = 2.08$, $SD = .85$) on, $t(48) = 2.81$, $p < .01$ (Cohen’s $d = .81$). Thus, the observed differences do not appear to result from a novelty effect.

Finally, to determine whether the results could be attributed to differences in the intelligibility of the human and machine voices, we conducted a supplemental experiment in which participants listened to a word problem spoken in machine voice and wrote down the words and also listened to a question spoken in human voice and wrote down the words. Specifically, 10 undergraduate students (3 males, 7 females; average GPA = 2.98; 6 educational psychology majors, 4 education majors) were presented with the first two worked-out examples of the previously described com-
puter-based learning environment. Using a counterbalanced procedure, the participants listened to the narration accompanying an on-screen worked example spoken in machine voice and wrote down the words and also listened to an example spoken in human voice and wrote down the words. One worked example consisted of 255 words while the other consisted of 192 words. The participants were not expected to solve the accompanying practice problems. The participants correctly recorded an average of 94.5% ($SD = 5.9\%$) of the example's narration when it was a human voice and an average of 93.4% ($SD = 5.0\%$) of the example's narration when it was a machine voice. According to a paired-sample $t$ test ($\alpha = .05$), the percentage of words recorded by the participants did not statistically differ across examples, $t(9) = .34$, $p = ns$. Thus, the pattern of results cannot be attributed to the human voice being substantially easier to discern than the machine voice.

In summary, the human voice condition produced statistically and practically significant differences in terms of practice problem solving (medium-to-large effect), near transfer (large effect), and far transfer (medium-to-large effect) as well as perception of the speaker's voice (medium-to-large effect). Interestingly, despite these performance differences, there did not appear to be any difference in perceived example understanding or perceived example difficulty. Moreover, we conclude that the observed differences should not be attributed to variation in time-on-task, novelty effect, or—according to our supplemental experiment—the intelligibility of the voices.

5. Experiment 2: Multimedia learning in a school setting

In an effort to replicate and extend these findings, we decided to conduct a small-scale field experiment at an area high school with students enrolled in one of several sections of the same college-preparatory mathematics courses. Furthermore, to help ensure the authenticity of the task, the experiment was run in the computer laboratory at the high school with the entire class—including the instructor—present at the time of the experiment. As with the previous experiment, we hypothesized that students in the human voice group should produce higher scores than students in the machine voice group on the practice problems, the near and far transfer tests designed to measure the depth of learner understanding, and rate the speaker more positively while, at the same time, not rating the examples as any more difficult or reporting any differences in understanding than their machine voice counterparts.

5.1. Method

5.1.1. Participants and design

The participants were 40 high school students recruited from several mathematics courses taught by the same instructor at Starkville High School (in Starkville, Mississippi). They were randomly assigned to condition, with 20 serving in the human voice group and 20 serving in the machine voice group. The percentage of males was 75% in the human voice group and 25% in the machine voice group; the percent-
age of juniors and seniors was 70% in the human voice group and 55% in the machine voice group; and the mean GPA was 3.55 for the human voice group and 3.58 for the machine voice group.

5.1.2. Materials and apparatus

The materials and apparatus were identical to Experiment 1. The apparatus consisted of 25 PC computer systems (750mHz, 256 RAM) with 15-in color monitors and headphones.

5.1.3. Procedure

The procedure was similar to Experiment 1. Instead of arriving individually in a laboratory equipped with 8 workstations, in the present experiment, an intact class arrived during two consecutive class periods at a lab containing 25 workstations where they were asked to locate a workstation where they would work independently. All other aspects of the procedure were identical to Experiment 1.

5.1.4. Scoring

The scoring was identical to Experiment 1. As with the previous experiment, a research assistant who was unaware of the participants’ treatment conditions independently scored each problem-solving protocol while a second rater independently scored a random sample of 20% of the protocols. They agreed on scoring 98% of the time. Discussion and consensus were used to resolve any disagreement between raters. Internal consistency reliabilities (Cronbach’s $\alpha$) for the practice problem, near transfer, and far transfer measures were .74, .80, and .79, respectively. Internal consistency reliability for the speaker rating survey was .87.

5.2. Results and discussion

The major research question addressed in this experiment concerned whether the results from Experiment 1, in which learners in the human voice reported increased interest in learning and achieved better transfer performance than learners in the machine voice condition, could be replicated with high school students. Table 2 shows the mean score (and standard deviation) for each group in Experiment 2 on the perceived example understanding, perceived example difficulty, performance on practice problems, near transfer test, far transfer test, speaker rating survey, and instructional time. Separate two-tailed $t$ tests were conducted on these measures, each at $\alpha = .05$.

5.2.1. Does voice affect perceived example understanding?

As indicated in the first row of Table 2, there was no significant difference in perceived example understanding between the participants assigned to the human voice condition ($M = 1.65$, $SD = .45$) and those assigned to the machine voice condition ($M = 1.51$, $SD = .43$), $t(38) = .98$, $p = ns$. In general, across both voice conditions, the results indicate that learners reported that the examples were relatively easy to understand.
5.2.2. Does voice affect perceived example difficulty?

As illustrated in the second row of Table 2, the human voice group (M = 2.44, SD = .75) and the machine voice group (M = 2.40, SD = .79) reported similar levels of perceived example difficulty, \( t(38) = .15, p = ns \). In sum, there was no difference in the perceived difficulty of the examples across conditions.

5.2.3. Does voice affect performance on practice problems?

As shown in the third row of Table 2, the human voice group (M = 2.33, SD = .64) scored significantly higher than the machine voice group (M = 1.80, SD = .86) on solving practice problems, \( t(38) = 2.20, p < .05 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .63, which corresponds to a medium effect. In general, the results show that human voice fostered better understanding of how to solve the practice problems than did machine voice.

5.2.4. Does voice affect near transfer?

As shown in the fourth row of Table 2, the human voice group (M = 2.51, SD = .59) scored significantly higher than the machine voice group (M = 1.84, SD = .86) on the near transfer test, \( t(38) = 2.89, p < .01 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .83 for near transfer, which corresponds to a large effect. As with Experiment 1, the results demonstrate that human voice fostered better understanding of how to solve problems like those presented during instruction than did machine voice.

5.2.5. Does voice affect far transfer?

As shown in the fifth row of Table 2, the human voice group (M = 1.74, SD = .70) scored significantly higher than the machine voice group (M = 1.15, SD = .82) on the far transfer test, \( t(38) = 2.42, p < .05 \). Cohen’s \( d \) statistic for these data yields an effect size estimate of .70 for far transfer, which corresponds to a medium-to-large effect.
As with Experiment 1, the results demonstrate that human voice fostered increased understanding of how to solve problems that were not like those presented during learning than did machine voice.

5.2.6. Does voice affect speaker rating?

As shown in the sixth row Table 2, the human voice group \((M = 3.19, SD = 1.05)\) rated the speaker significantly more favorably than did the machine voice group \((M = 4.23, SD = 1.30)\) on the speaker rating test, \(t(38) = 2.78, p = .008\). Cohen’s \(d\) statistic for these data yields an effect size estimate of .83 for speaker rating, which corresponds to a large effect. As with Experiment 1, students in the human voice condition reported a more positive evaluation of the speaker’s attractiveness, dynamism, and superiority than did the machine voice group.

5.2.7. Related issues

We once again examined whether the pattern of results could be attributed to time-on-task by comparing time spent on the instructional program across the two conditions. We concluded that the pattern of results cannot be attributed to the human voice group spending more time during learning than the machine voice group, since the instructional time for the machine voice group \((M = 42.1\text{min}, SD = 10.3\text{min})\) was not significantly different from the human voice group \((M = 40.7\text{min}, SD = 16.4\text{min})\), \(t(38) = .62, p = ns\).

Again, to explore the possibility that the pattern of results could be attributed to a “novelty effect,” we examined the performance on the practice problems in the first half versus the second half of training. Although the averages associated with the first two practice problems for participants in the human voice group \((M = 2.18, SD = 1.39)\) were not significantly higher than those of their peers in machine voice group \((M = 1.80, SD = 1.04)\) on, \(t(38) = 1.09, p < ns\), the scores for the participants in the human voice group \((M = 2.48, SD = .73)\) were significantly higher than those of their peers in machine voice group \((M = 1.80, SD = 1.15)\) on the last two practice problems, \(t(38) = 2.21, p < .05\) (Cohen’s \(d = .70\)). Thus, the observed differences did not appear to result from a novelty effect since they were in fact opposite of what one would expect if the novelty effect existed.

In sum, the human voice condition once again produced statistically and practically significant difference in terms of practice problem solving (medium effect), near transfer (large effect), and far transfer (medium-to-large effect) as well as perception of the speaker’s voice (large effect). In spite of these performance differences, there did not appear to be any difference in perceived example understanding or perceived example difficulty. Finally, as with Experiment 1, these differences could not be attributed to time-on-task or a novelty effect.

6. Conclusion

In our computer-based learning environment designed to teach mathematics, we attempted to foster a sense of social presence in which learners would be more likely
to interpret the computer-based narrator as a social partner. Overall, across two different experiments with different kinds of participants, we obtained a voice effect in which students achieved better transfer performance when the on-screen agent spoke in a human voice than when the on-screen agent spoke in a machine synthesized voice. Importantly, learners also gave more positive ratings to the on-screen agent who spoke with a human voice rather than a machine voice on an instrument designed to capture the social characteristics of speakers. The participants in both experiments were not expert with the material but were capable of learning it.

6.1. Theoretical implications

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 similar to human-to-human conversation. Although the results are tentative, we found little evidence that our attempts to promote social agency (by using a human voice) increased cognitive load—that is, there were no differences between the two voice conditions in terms of perceived example understanding or in terms of perceived example difficulty (i.e., cognitive load). In particular, the learners who received the human voice showed substantial advantages in solving practice problems during instruction and on solving near and far transfer problems after instruction, as well as reporting a more positive rating of the on-screen agent across a number of speaker dimensions, including attractiveness, dynamism, and superiority. The voice effects we found in these two experiments replicate and extend the voice effect reported by Mayer et al. (2003) by employing new instructional materials (e.g., math instruction rather than science; procedural knowledge rather than conceptual knowledge; 40 min of instruction rather than 2 min; an embodied animated agent as the source of the aural instructional support rather than a disembodied voice; female voices rather than male) and more differentiated dependent measures (i.e., performance on practice problems, near and far transfer problem solving), and by showing that the same effects occur in both laboratory and school settings.

6.2. Practical implications

Our results support a multimedia design principle, which can be called the voice principle: Designers of multimedia learning environments should create life-like on-screen agents that speak in a human voice rather than a machine-synthesized voice. The practical significance of our findings is reflected in the strong and consistent effect sizes: Across the two experiments, the effect sizes for near transfer measures were large and the effect sizes for far transfer measures were medium-to-large. Moreover, the magnitude of effects captured in the present study were comparable to the two voice effects reported by Mayer et al. (2003), namely the medium-to-large effect associated with improved transfer performance of learners exposed to a native-born speaker versus one with a foreign accent and the large effect associated with enhanced transfer performance of learners exposed to a disembodied human voice versus a computer-generated voice. Importantly, we also obtained the
same pattern of results in a laboratory experiment and a field experiment, suggesting the robustness of the voice effect.

6.3. Future directions

Although social agency theory provides one theoretical explanation for these results, an alternative explanation is offered by cognitive load theory (Sweller, 1988; Sweller, Van Merriënboer, & Paas, 1998). Cognitive load is a “multidimensional construct representing the load that performing a particular task imposes on the learner’s cognitive system” (Paas, Tuovinen, Tabbers, & Van Gerven, 2003, p. 64). In particular, cognitive load theory posits that there are three types of cognitive load: (a) intrinsic, which is the amount of cognitive load imposed due to the complexity of the learning material in a domain, where the magnitude is dependent on a learners’ level of prior domain knowledge; (b) germane, which reflects the demands placed on working memory capacity by mental activities that contribute directly to learning; and (c) extraneous, which is the cognitive load caused by mental activities during learning that do not contribute directly to learning. When learners engage in activities that do not contribute to their content understanding, working memory resources are poorly allocated thus leaving little working memory available for actively constructing new schemas. It is plausible that this latter type of load could account for the pattern of results found in this study. Perhaps the novelty of the voice for the learners in the machine voice condition caused them to engage in superfluous cognitive activities that did not contribute directly to learning, thus causing their transfer performance to suffer.

In the present study, we used a direct, subjective measure to assess cognitive load, namely asking learners to rate the difficulty of the instructional material. According the results we obtained from this measure, there was no difference in reported cognitive load between the human and machine voice conditions. This type of measure is, however, only one of four different approaches for measuring cognitive load (Brünken, Plass, & Leutner, 2003). For instance, there are also direct, objective measures of cognitive load including the dual-task approach. This approach involves having learners simultaneously engage in two activities that each require the same amount of mental resources in verbal and/or visual working memory. Dual-task format uses attentional resources allocated to either the primary or the secondary task to measure cognitive load, in order to analyze whether the cognitive load imposed by the primary task interferes with performance on the secondary task and vice versa.

Thus, it appears worthwhile to examine cognitive load more closely, by incorporating other more direct measures of mental effort beyond the item used in the present study (i.e., perceived example difficulty) adapted from an instrument developed by Paas and Van Merriënboer (1993) to measure participants’ perceived cognitive load. One promising technique is an assessment of cognitive load using a dual task methodology (Brünken et al., 2003; Brünken, Steinbacher, Plass, & Leutner, 2002).

Second, as machine synthesized voices improve, it would be important to test their effectiveness to see if they can close the performance gap that this study highlights between human voice and machine voice. In addition, it would be worthwhile to
examine whether the effects are diminished if learners receive more practice with the machine voice.

Third, the present study relies on indirect measures of the degree to which students experience a social relation with the agent, but does not provide direct evidence that simply listening to a computer programmed with a human voice primes the learner to engage in the conventions of human conversation. Future research is needed to test social agency theory using more direct measures of the learner's communicative responses. Social agency theory predicts that students who receive a human voice would be more likely than students who receive the machine voice to display facial expressions and gestures during learning, to directly address the computer using words like “you” when asked to write a type a critique of the instructional program, and to rate the learning experience as more personal and friendly on a post-experimental survey. For example, social agency theory predicts that students who receive a human voice would give higher agreement ratings than students who receive a machine voice to assertions such as, “I felt as if the computer was talking directly to me,” “I felt as if I was in a conversation with the computer,” and “The computer was trying to be cooperative and helpful with me.” Overall, social agency theory needs to be tested more directly in future research.

References


