Human Communication Research ISSN 0360-3989

ORIGINAL ARTICLE

When a Talking-Face Computer Agent is Half-Human and Half-Humanoid: Human Identity and Consistency Preference

Li Gong[1] & Clifford Nass[2]

1 School of Communication, Ohio State University, Columbus, OH 43210
2 Department of Communication, Stanford University, Stanford, CA 94305

*Computer-generated anthropomorphic characters are a growing type of communicator that is deployed in digital communication environments. An essential theoretical question is how people identify humanlike but clearly artificial, hence humanoid, entities in comparison to natural human ones. This identity categorization inquiry was approached under the framework of consistency and tested through examining inconsistency effects from mismatching categories. Study 1 (*N = *80), incorporating a self-disclosure task, tested participants' responses to a talking-face agent, which varied in four combinations of human versus humanoid faces and voices. In line with the literature on inconsistency, the pairing of a human face with a humanoid voice or a humanoid face with a human voice led to longer processing time in making judgment of the agent and less trust than the pairing of a face and a voice from either the human or the humanoid category. Female users particularly showed negative attitudes toward inconsistently paired talking faces. Study 2 (*N = *80), using a task that stressed comprehension demand, replicated the inconsistency effects on judging time and females' negative attitudes but not for comprehension-related outcomes. Voice clarity overshadowed the consistency concern for comprehension-related responses. The overall inconsistency effects suggest that people treat humanoid entities in a different category from natural human ones.*

doi:10.1111/j.1468-2958.2007.00295.x

Computer-generated characters are increasingly used as digital communicators on Web sites and in computer applications and computer games. They are referred to as "agents" when they autonomously interact with users, as personal assistants (Maes, 1994), virtual news anchors (e.g., http://www.ananova.com/video), virtual educators (Lester, Stone, Converse, Kahler, & Barlow, 1997), e-commerce agents (Tanaka, 2000), or online interviewers (e.g., http://www.aboutefile.com). They are commonly called "avatars" when they represent human users in computer games or virtual

Corresponding author: Li Gong; e-mail: gong.33@osu.edu
This study is based on the first author's doctoral dissertation, for which the second author was the primary advisor. This paper was accepted under the editorship of James P. Dillard.

reality (Biocca, 1997). They are also often used to visually represent interactants in computer-mediated communication (E.-J. Lee, 2004; E.-J. Lee & Nass, 2002). Many of the computer agents or avatars are anthropomorphic, for example, virtually embodied with computer-generated faces. Often, these synthetic faces talk with computer-generated speech or prerecorded natural speech using lip synchronization and facial animation technology (Cassell, Sullivan, Prevost, & Churchill, 2000).

Apart from evidence that people are amazed by the technological feat of simulating human faces and speech, researchers do not know much about how people perceive the identity of anthropomorphic talking faces with respect to humanness. Are they considered human, as renderings of people in photos or on TV are? Or, do people consider them nonhuman because they are clearly artificial and computer synthesized? Or, is there a new identity category for them? And how will the identity categorization of talking-face agents impact people's interaction with them?

Human faces and voices are two essential biosocial attributes defining most humans (Nass & Brave, 2005). Although computer-generated anthropomorphic faces and speech are clearly humanlike rather than animallike, they can be quickly detected as artificial by most people. People easily detect the artificial quality in the graphics of computer-generated faces (Cassell et al., 2000). Computer-generated speech lags behind natural human speech in both clarity and prosody (Gong & Lai, 2003; Olive, 1997; van Santen et al., 2000). "Humanoid" is adopted as the working term in this article to refer to the state of being humanlike but bearing the clear artificiality of computer synthesis. Veridical renderings of real humans such as in photos and videos are considered human.

Research is lacking in the literature for addressing the human identity issue of humanoid entities. The extant research on anthropomorphic agents and avatars has mainly focused on how attributes and processes about humans apply to agents and avatars, such as gender and ethnicity (Baylor & Kim, 2003) and personality (Nass & Lee, 2001), or on attributes unique to computer characters such as the degree of anthropomorphism (Nowak, 2004). Research grounded in the Computers Are Social Actors Paradigm (Reeves & Nass, 1996) has demonstrated that people follow the same social rules and heuristics when they interact with computers and other media as when they interact with other humans. However, this line of research primarily compared one computer to another and did not involve the direct comparison of human and computerized humanoid representations. The research presented in this article provides a direct human identity test for humanoid entities on computers and examines the effects on people's communication with them.

Identity categories become apparent when people perceive mismatching. A red patch is considered different from a blue patch when they are juxtaposed. Gender dichotomy becomes salient when a man puts on a woman's dress. When inconsistency or mismatching is detected, it suggests the existence of distinct categories in the perceiver's cognitive model.

Humans are born with a *consistently* paired package of a face and a voice except in the case of aphasia. That is, the *consistent* packages of human faces

and voices are the cognitive models that we have developed through repeated exposures to human faces and voices in face-to-face or technology-mediated interactions.

For computer-generated talking faces, however, media designers pick the face and the voice. The patterns of face–voice pairings in digital characters may differ from genetically determined patterns that people are used to. For example, recorded natural human speech is paired with a technologically sophisticated, three-dimensional humanoid face, which can be switched to be skinless and rotated to show the artificial palate and teeth from different angles to teach deaf children how to speak (Massaro, 1998). In another application, the image of a human face has been used as the face of a virtual translator that autonomously switches among different computer-generated languages (Ritter, Meier, Yang, & Waibel, 1999). Similarly, a video-captured human face was texture mapped to talk with a computer-synthesized humanoid voice in two experimental studies, which aimed to examine social responses to human faces as computer agents but failed to conceptualize the consistency issue between the face and the voice (Kiesler, Sproull, & Waters, 1996; Sproull, Subramani, Kiesler, Walker, & Waters, 1996).

The natural pairing of a human face and a human voice is usually not an option for autonomous agents because video recording a human requires manual and planned effort, and technology does not yet enable automatic real-time coupling of originally separate natural facial movements and natural speech. Thus, technologists mostly follow a *modular excellence philosophy* in that for each modality, the best among what is available is used, often without considering the issue of consistency between modules or modalities (Nass & Gong, 1999). This philosophy explains why a human voice is used with a computerized face in Massaro's (1998) case: Teaching syllables, words, and phrases involves limited content that can be prerecorded, but only a computer face can become skinless. It also explains why a human face and computer-generated speech are used in Ritter et al.'s (1999) case: Only computer speech engines can autonomously convert text to speech in multiple languages without needing a human translator.

Although a talking face can show palate movement or utter translated spoken words, does the pairing of a humanoid face with a human voice or a human face with a humanoid voice create mismatching and perceptual inconsistency? This question was not applicable until the advent of interactive and autonomous agents. Drawings and photos do not talk. Films, videos, and TV programs usually capture the facial movements and speech of the same person. In the case of dubbing, synchrony is trained to be perfect. When dubbing presents visual-auditory asynchrony or crossing in social identity categories such as gender or accent, people identify such inconsistencies effortlessly and immediately. Nonhuman cartoon characters are also dubbed with human voices (albeit unusual voices; see Nass & Brave, 2005). Would there be a different choice if we ever could understand the language of animals? Commonly, voice actors adjust their vocal profiles to simulate the vocal features of nonhuman characters, such as Mickey Mouse.

For the first time in the history of human technological creation, computer-synthesized faces and voices can move their own faces and speak their own voices. Do humans then draw a human identity line to differentiate these artificial humanoid entities from us? Or, do people classify them within the human category?

One might argue that face and voice are two different modalities and that they can enjoy some degrees of freedom for taking on different identity categories. However, this argument is untenable because vast literatures on speech processing and person perception clearly demonstrate that people inherently integrate face and voice in their perception as well as production. The well-established McGurk effect shows that speech perception and production are inherently bimodal (McGurk & MacDonald, 1976). For example, seeing a "ga" in a person's lip movement but hearing a "ba" in his/her voice results in the perception of "da." The McGurk effect has been shown to be very robust. The ability to detect a McGurk effect was found in infants as young as 22 weeks old (Rosenblum, Schmuckler, & Johnson, 1997). At the level of social identity perception, research also found inconsistency effects from mismatching the ethnic cues in a person's face versus voice (Nass & Brave, 2005). In the case of human versus humanoid, the human face talking with the humanoid voice in Kiesler et al.'s study (1996) received more negative social evaluation than the humanoid voice by itself. This finding indicated that adding the human face to the humanoid voice did not facilitate but instead dampened people's judgment. The most potent explanation is the apparent inconsistency between the human face and the humanoid voice.

If exemplars from different categories are mixed within one entity, inconsistency ensues in the beholder's perception. Conversely, the perception of inconsistency reveals the distinction between the involved categories. If pairing a humanoid face with a human voice or pairing a human face with a humanoid voice activates inconsistency in people's perception and responses, it would suggest that people hold "human" and "humanoid" as distinct categories. If no inconsistency is perceived, humanoid faces or voices may be considered in the same category as natural human faces or voices.

Effects of inconsistency

Research on perception, social cognition, and communication all demonstrate that people desire consistency in perceiving objects and perceiving and interacting with other people. The classic Gestalt theory in psychology proposes a holistic processing model for perceptual organization and includes principles such as similarity grouping (Pomerantz & Kubovy, 1986). Circles are grouped with circles, squares with squares. Similarity breeds consistency in the perceptual organization. Asch (1946) extended the Gestalt theory to person perception by postulating that people process a person as a single psychological unit and form a unified and consistent impression of the person. Abundant research has proven that mixing elements from different categories in an object, a person's traits, and a person's communication cues causes

inconsistency, requiring more processing effort from the perceiver and leading to negative attitudinal and behavioral responses.

A large body of perceptual research has repeatedly demonstrated the "Stroop" effect (Stroop, 1935), by showing that inconsistency in the perceptual target causes interference in perception and results in longer response time. Stroop found that people took a longer time to name an ink color aloud when the word printed in that ink meant a different color—for example, the word "blue" in red ink—than when only a color square was on the card. Numerous studies replicated color-word compound stimuli by comparing congruent pairings and incongruent pairings, by varying one card to two cards and the positioning of the cards, or by varying the hue or brightness of the color, and all demonstrated the Stroop inconsistency interference effect (MacLeod, 1991). The Stroop effect was also confirmed with other visual stimuli, such as the images of an ankle and a hand (Golinkoff & Rosinski, 1976) and auditory stimuli. The words "high" and "low" inconsistently spoken in a low pitch (110 Hz) or high pitch (175 Hz) caused longer time for people to repeat the word than when they were spoken in the consistent pitch (Green & Barber, 1981; Hamers & Lambert, 1972). In another study, the participants were slower to identify the speaker's gender when the male voice said the word "girl" and when the female voice said the word "man" than when the voice and the word matched in gender (Green & Barber).

Similarly, studies in person perception found that inconsistent traits presented about a target person delay people's judgment of the person. Schul, Burnstein, and Martinez (1983) had undergraduate students rate how suitable a hypothetical person would be for an occupation. Two traits of a person were presented, which were either consistent or inconsistent. They found that participants took significantly longer time to reach judgment when the two traits were inconsistent than when they were consistent. Fiske and Taylor (1991) claimed that inconsistency in traits requires more effort to process. Fiske and Neuberg's (1990) continuum model of impression formation postulates that inconsistent information is processed as pieces, whereas consistent information is processed as a whole.

In addition to more processing effort and longer processing time, research has also documented negative attitudes resulting from trait inconsistency. "It was assumed that inconsistency perceived as intrinsic to the character of a person is unpleasant, and that such persons will be relatively disliked" (Hendrick, 1972). Hendrick found that college students reported less liking of a hypothetical person when the person was presented with inconsistent traits than when the person was presented with consistent traits.

Extending person traits to personality traits of computer agents, recent studies have produced findings of negative effects of inconsistency. A stick-figure agent was found to be less persuasive and evaluated more negatively when it had inconsistent personality cues between its posture and its verbal message than when they were consistent (Isbister & Nass, 2000). A computer-generated voice agent was considered more trustworthy when its personality cues were consistent between its vocal

parameters and its verbal messages than when they were inconsistent (Nass & Lee, 2001). Again, in the human versus humanoid dimension, the human face talking with the humanoid voice received more negative social evaluation than the humanoid voice alone in Kiesler et al.'s (1996) study, suggesting an inconsistency effect.

A wealth of research focusing on processing of verbal and nonverbal communication cues presents clear evidence that inconsistency in the valence of verbal versus nonverbal cues causes perceptual and judgment confusion, negative attitudes, suspicion of deception, and avoidant behaviors. In a study, the participants had more difficulty in judging the intended attitude of a mother toward her child when the audiotapes of her messages contradicted with her demeanors in the pictures in terms of valence (Roy & Sawyers, 1990). Argyle, Alkema, and Gilmour (1971) found that a person who showed hostile or friendly attitude inconsistently between verbal content and nonverbal style (face, vocal tone, and posture) was rated as less stable, less sincere, and more confusing. Compared to consistency, inconsistency in the valence of verbal and nonverbal cues was demonstrated to upset children and result in their avoidance of the source person in one study (Volkmar & Siegel, 1979) and to cause greater physical distance between college students and a counselor in another study (Graves & Robinson, 1976).

In particular, inconsistency between verbal and nonverbal channels is found to be a major indicator of deception (Ekman & Friesen, 1969, 1974). People are found to be most capable of manipulating their verbal messages, much less so for facial expressions, and least so for their body movements. Facial expressions and body language are the main leakage outlets of lies when they convey a different valence from the verbal messages. Mehrabian (1971), for example, found that when college students were asked to deceive in an experiment, they managed to be verbally assertive but showed more negative nonverbal cues such as slower speech, more speech errors, and less immediacy in distance.

Hence, preference for consistency appears a strong human propensity. Inconsistency leads to more processing effort and more negative attitudinal and behavioral responses than consistency. Inconsistency resulting from mixing indicates that the elements mixed belong to different categories. Given the current state of the technology for computer-generated faces and voices and the obvious artificiality of their humanoid quality, we hypothesized that people treat humanoid and human faces or voices in different categories and this distinction will be reflected in inconsistency effects when a face and a voice from these two categories are mixed into one talking-face agent. Based on the evidence in the literature that inconsistency causes negative attitudes and incurs longer response time in making judgment of the perceptual target, the postulated inconsistency effects were tested through examining people's negative attitudes and time in making judgment of a talking face. In particular, the literature has shown that inconsistency hurts trust and activates suspicion of deception. Trust is considered a critical factor in human–agent interaction, for example, when an agent solicits information from users and offers opinions, products, and services (Bickmore & Cassell, 2001; Grabner-Kräuter & Kaluscha, 2003; Nass & Lee,

2001). Therefore, three hypotheses were proposed to test trust, negative attitudes, and judging time, respectively.

H1: People trust a talking-face agent less when it is a human face talking with a humanoid voice compared to a human voice and when it is a humanoid face talking with a human voice compared to a humanoid voice.

H2: People have more negative attitudes toward a talking-face agent when it is a human face talking with a humanoid voice compared to a human voice and when it is a humanoid face talking with a human voice compared to a humanoid voice.

H3: People take longer time in making judgment of a talking-face agent when it is a human face talking with a humanoid voice compared to a human voice and when it is a humanoid face talking with a human voice compared to a humanoid voice.

It will be reasonable to assume that people by default expect a human face talking with a human voice except in the case of speaking impairment. What is interesting and less predictable is how people would respond to a humanoid face talking with a human voice versus a humanoid voice. The framework of consistency preference supports the superiority of pairing a humanoid face with a humanoid voice, as proposed in the hypotheses. But the approach of modular excellence would call for pairing a humanoid face with a human voice because natural human speech is by far clearer and more pleasant than synthetic humanoid speech.

Gender differences in processing nonverbal communication

Along the traditional line of verbal and nonverbal distinction in communication, face–voice pairing would fall into the nonverbal domain because talking-face agents of all pairings could deliver the same verbal messages. The nonverbal literature shows substantial differences between women and men in processing nonverbal cues. Women are found to pay more attention to nonverbal cues, be more accurate in decoding nonverbal cues, and be more susceptible to the influence of nonverbal cues than men. For example, women gaze more than men (Argyle et al., 1971) and paid more attention to actions, whereas men paid more attention to the verbal content of conversations (Mazanec & McCall, 1976). Hall (1978, 1984) conducted meta-analyses of two nonoverlapping groups of studies on decoding nonverbal cues and indicated that women are more accurate in decoding nonverbal cues than men regardless of the age or gender of the stimulus person or the age of the participant. Research has also shown that women are more susceptible to the influence of nonverbal cues than men. Compared to men, women relied more on nonverbal cues than verbal cues when rating stimulus people in the dominance-submissiveness dimension in one study (Argyle, Salter, Nicholson, Williams, & Burgess, 1970) and rating spoken messages in the friendly–hostile dimension in other studies (Zahn, 1973). Therefore, women were hypothesized to be more attentive to the consistency issue in pairing face and voice for a talking-face agent and be more susceptible to the

effects of inconsistency because face–voice pairing is a nonverbal rather than a verbal issue.

H4: Inconsistency in the face–voice pairing of a talking-face agent causes greater negative effects on women than on men.

Study 1

Method

A 2 (human face vs. humanoid face) × 2 (human voice vs. humanoid voice) × 2 (male vs. female users) between-participants experiment was conducted. The face and voice factors constituted all four possible face–voice pairings: a human face with a human voice, a human face with a humanoid voice, a humanoid face with a humanoid voice, and a humanoid face with a human voice. The experimental task was a self-disclosure interview-type task framing the talking-face agent as an interviewer. The purpose was to stress the trust factor because self-disclosure of personal information has been found to be positively related to trust (Wheeless & Grotz, 1977).

Participants

Eighty undergraduate students (40 men and 40 women) enrolled in a large communication course at a private, West Coast U.S. university were recruited to participate in the study. Only English native speakers were selected to avoid potential language problems in understanding the speech. Ten male and 10 female participants were randomly assigned to each of the four combinations of human and humanoid faces and voices. Participants received extra course credit for their participation in the study.

Procedure

The study was labeled as testing a prototype interviewing system. Nine out of 11 open-ended intimate self-disclosure questions developed by Moon (2000) were used (the two questions about sexual fantasy and arousal were omitted because they appear to be of a different nature from the other disclosure questions). Two sample self-disclosure questions are "What have you done in your life that you feel most guilty about?" and "What characteristics of your best friend really bother you?" As in Moon's (2000) study, four nonintimate warm-up questions about age, gender, hometown, and hobbies were used prior to the actual self-disclosure questions. Each question was one sentence long and was orally presented by the talking face, one at a time. The video of the talking face, 9.60 × 7.15 cm in its frame, was displayed in the upper center of a 17-inch computer screen. The participants typed their answers in a text box underneath the video for each question. Participants could click a "repeat" button to hear the question again to minimize any possible comprehension problem. The repeat button was clicked a total of three times in 1,040 (80 participants × 13

questions) opportunities, indicating that the short disclosure questions were not difficult for the participants to understand in any of the face–voice combinations. After the participants submitted their answers to the last disclosure question, they answered a posttask questionnaire on the computer.

Manipulation of the face and the voice stimuli

The humanoid face used in the study was a computer-generated and computer-animated face named "Baldi." Baldi is based on a computer-generated facial animation model (Parke & Waters, 1996). It was provided as a module in the Center for Spoken Language Understanding (CSLU) Toolkit.[1] The Baldi face was designed with male and Caucasian features (Massaro, 1998). The humanoid voice was synthesized by the male American English model of the Festival computer speech engine, which is also a module in the CSLU Toolkit. The humanoid face looks obviously computer generated and synthetic, and the humanoid voice by the Festival speech engine sounds obviously synthetic as well (Massaro; Nass & Brave, 2005). A male Caucasian graduate student from the Department of Drama at the university was recruited to be the actor providing the human face and the human voice. Figure 1 presents the pictures of the humanoid face and the human face.

The actor was digitally videotaped with his natural speaking to create the stimulus of the human face with the human voice. The audio track was fed into the CSLU Toolkit, which animated and lip synchronized the Baldi face with the actor's voice. The lip synchronization of the Baldi face by the CSLU Toolkit was excellent and represented the state of the art (Massaro, 1998). The animation and lip synchronization feature of the CSLU Toolkit was also used to create the version of the humanoid face with the humanoid voice. For the stimulus of the human face with the humanoid voice, the actor was trained to lip synchronize with the humanoid voice and then was videotaped talking over the humanoid voice. The resulting video file was dubbed with the recorded sound files of the humanoid voice. The lip synchrony was extremely good. This approach was preferable to texture mapping the actor's face onto Baldi because the mapping would not be sufficiently lifelike and the lip

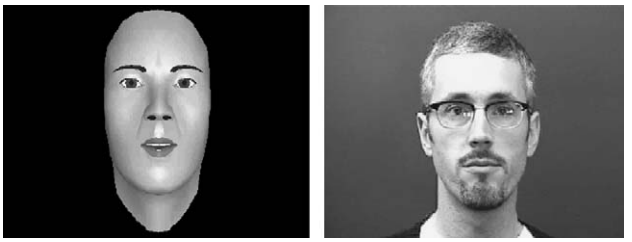

Figure 1 Still pictures of the humanoid face and the human face.

Note: The pictures were captured from the videos of the talking faces. The original videos were in color.

synch would not be improved. By speech experts' judgment, all four face–voice combinations had excellent and comparable lip synchronization. A perceptual study to test people's judgment of lip synchronization is not appropriate because it is impossible to separate the human–humanoid factor, which was self-evident in the face including the lip and the voice, from the pure speech–lips synchronization factor.

The human actor presented the materials with neutral facial expressions and tone. The humanoid face and voice used the default neutral parameters in the CSLU Toolkit.

Measures

Manipulation check and explicit measure of consistency

The end of the posttask questionnaire included questions checking the manipulation of the experimental stimuli and an explicit measure of consistency between the face and the voice. Two questions asked the participants whether the face and the voice of the agent were "*computer synthesized*" or "*of a real person*," respectively. The explicit measure of consistency contained two 10-point semantic differential scales for judging the face and the voice as being *inconsistent–consistent* and *mismatched–matched.* The responses to these two items were averaged; the index was very reliable (Cronbach's $\alpha = .89$). All participants reported no prior exposure to the faces or the voices.

Trust-related measures

To test H1, trust was assessed behaviorally by examining the participants' answers to the self-disclosure questions and attitudinally with a scale in the questionnaire. Participants' self-disclosure was analyzed along the dimensions of amount and intimacy, two important dimensions of self-disclosure (Dindia, Fitzpatrick, & Kenny, 1997; Gibbs, Ellison, & Heino, 2006; Wheeless, 1978).

An objective word-count method and a subjective coder-rating method were used to capture the amount of self-disclosure. As in Moon (2000), the number of words in the answer to each self-disclosure question was counted. Such answers as "I don't know," "I don't want to tell you," or "pass" counted as zero words. The word-count measure of the amount of self-disclosure showed good reliability across the nine items (Cronbach's $\alpha = .80$). Hence, the items were averaged to create a single objective amount measure.

In addition, two coders were hired and trained to code each response for the amount and intimacy of self-disclosure. The coders were blind to the study's purpose and the experimental conditions and did the ratings independently.

The coders first rated the amount of self-disclosure for all the responses of each participant on a 7-point scale (1 = *none*, 7 = *a lot*). The intercoder reliability across items was high (average $r = .83$, r ranged from .73 to .89, all ps $< .001$). The two coders then rated each item for each participant on intimacy on a 7-point scale (1 = *not intimate at all*, 7 = *very intimate*), consistent with Dindia et al. (1997).

The intercoder reliability for the intimacy ratings was good (average $r = .77$, r ranged from .70 to .82, all $ps < .001$). Because of the high reliability, the two coder's ratings were averaged for each item.

Moon's self-disclosure instrument was created to tap relatively intimate topics with multiple homogenous questions. Responses to the nine self-disclosure questions were expected to be similar in amount and in intimacy. Interitem reliability was good for the subjective amount ($\alpha = .85$) and intimacy ratings ($\alpha = .78$). Thus, the instrument was reliable, and the ratings were averaged across the nine items to create an overall subjective amount rating and intimacy rating.

The word-count measure and the subjective rating for the amount of self-disclosure were highly correlated ($r = .81$, $p < .001$), as were the amount and intimacy ratings ($r = .72$, $p < .001$). Nonetheless, we analyzed the three separately because of the measurement distinctions between objective and subjective measures of amount of disclosure and conceptual differences between amount and intimacy of self-disclosure.

Wheeless and Grotz's (1977) Individualized Trust Scale was adapted to provide an attitudinal measure of the trust of the talking-face agent and was placed at the beginning of the questionnaire. Two pairs of adjectives (*exploitive–benevolent* and *faithful–unfaithful*) in the original scale were excluded because these two items did not appear to have face validity for measuring the talking-face agent. Via a principal-components factor analysis, six pairs of adjectives were selected to constitute the trust measure: *trustworthy–untrustworthy* (reverse coded), *unreliable–reliable*, *distrustful of the agent–trustful of the agent, confidential–divulging* (reverse coded), *dangerous–safe*, and *respectful–disrespectful* (reverse coded). The original 7-point semantic differential format was retained. The responses to these six items were averaged to form the measure of self-reported trust of the talking-face agent. The index was very reliable ($\alpha = .87$).

In addition, three items measured the participants' felt hesitance to disclose as a trust-related measure. The participants indicated how three adjectives described their feelings: *guarded, hesitant,* and *questioning* on 10-point Likert scales (1 = *describes very poorly*, 10 = *describes very well*). The responses to these three items were averaged; the index was reliable ($\alpha = .76$).

Negative attitudes toward the talking face

Negative attitudes toward the talking-face agent were assessed to test H2. Two indices were created. Negativity of the agent was measured by three adjectives: *disconcerting, disturbing,* and *frustrating*. Strangeness of the agent was measured by *strange*, *surprising*, and *unusual*. The participants rated how well these adjectives described the agent on 10-point Likert scales (1 = *describes very poorly*, 10 = *describes very well*). The responses to the items within each index were averaged. Both indices were reliable ($\alpha = .80$ and $\alpha = .75$, respectively).

To further explore the interaction between the face and voice modalities and the potential impact of face–voice pairing on judgment of the face and the voice independently, separate attitudinal measures were assessed for the face and the voice.

Participants evaluated how much they liked the face on five 7-point semantic differential scales: *pleasant–unpleasant* (reverse coded), *likeable–annoying* (reverse coded), *cold–warm, disturbing–pleasing,* and *nice–awful* (reverse coded). The responses to these items were averaged; the index was reliable ($\alpha = .81$). The participants also rated how much they liked the voice on three 7-point semantic differential scales: *jarring–soothing, unpleasant–pleasant*, and *awful–nice*. The responses to these items were averaged; the index was very reliable ($\alpha = .86$).

Time in judging the talking face

To test H3, the computer serving the questionnaire recorded the time that each participant spent on every page of the questionnaire. The total time spent on a questionnaire page, recorded in milliseconds, was divided by the number of items on that page to form an average per item time measure. The time data were then converted into seconds because all the time data were more than 1 second. This time measure is similar to the time measure used by studies judging a target person (Schul et al., 1983).

Results

Because this study primarily concerned the consistency issue between the face and the voice of the talking face, two-way analysis of variance (ANOVA) for the face and voice factors were at first conducted for male and female participants, respectively. Then, three-way ANOVA were run including the gender factor to discern possible interaction effects due to gender. The face–voice interaction was tested again in the three-way ANOVA's. Table 1 presents the means and standard deviations of all the dependent measures. Table 2 presents the interaction results for the two-way face–voice ANOVAs and the three-way face–voice–gender ANOVAs.

Manipulation check and the explicit measure of consistency

All participants correctly identified the human face and voice as "of a real person" and the synthetic face and voice as "computer synthesized" in all conditions. Thus, mixing human and humanoid faces and voices did not confuse the participants about the nature of the face or the voice. The explicit measure of consistency yielded significant face–voice interaction effects for both males and females. The participants perceived the human face as more consistent with the human voice compared to the humanoid voice and the humanoid face as more consistent with the humanoid voice as compared to the human voice. The face–voice interaction effect remained significant in the three-way ANOVA. The three-way interaction was not significant.

Testing H1: inconsistency effects on trust-related measures

Significant face–voice interaction effects supporting the consistency hypothesis were found for both measures of the amount of self-disclosure and the trust of the agent for both gender groups and for hesitancy to disclose for female participants. Such an interaction effect was found at the marginal significance for intimacy of self-disclosure among all participants and particularly among females.

Table 1 Means and Standard Deviations of Dependent Measures in Study 1

	Males				Females			
	Human Face		Humanoid Face		Human Face		Humanoid Face	
	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice
Consistency	7.65 (1.94)	2.65 (1.72)	4.76 (1.68)	7.35 (2.47)	7.05 (2.48)	2.66 (1.47)	3.42 (1.14)	6.55 (2.29)
Amount of Disclosure 1	17.60 (1.20)	10.19 (3.79)	8.87 (3.52)	10.92 (5.80)	17.73 (3.74)	12.19 (5.52)	9.15 (3.27)	11.65 (7.94)
Amount of Disclosure 2	3.53 (.63)	2.98 (.45)	2.67 (.44)	3.06 (.70)	3.58 (.60)	2.94 (.42)	2.84 (.49)	3.04 (.91)
Intimacy of disclosure	3.74 (.59)	3.48 (.58)	3.48 (.57)	3.56 (.86)	4.05 (.61)	3.59 (.45)	3.47 (.49)	3.62 (.81)
Trust of agent	4.83 (.44)	4.55 (.46)	4.00 (.58)	5.06 (.94)	5.09 (1.04)	4.15 (.95)	4.59 (.82)	4.85 (.74)
Hesitancy to disclose	4.07 (.94)	4.53 (1.85)	4.56 (1.96)	5.63 (.86)	4.00 (1.29)	5.51 (.66)	5.08 (1.46)	4.20 (2.10)
Negativity of agent	3.60 (1.33)	4.29 (1.09)	3.47 (2.06)	3.85 (1.17)	2.71 (1.51)	6.04 (1.81)	5.24 (1.48)	4.48 (1.29)
Strangeness of agent	4.20 (1.76)	5.15 (.92)	4.76 (.51)	5.59 (1.68)	4.33 (1.76)	6.07 (1.55)	6.30 (1.43)	4.75 (1.42)
Liking of the face	3.97 (.55)	3.42 (.69)	4.13 (.80)	3.76 (.92)	3.74 (.51)	2.96 (.53)	2.28 (.99)	3.36 (.54)
Liking of the voice	4.47 (.76)	2.73 (1.06)	4.40 (.91)	3.08 (1.03)	4.44 (.50)	1.97 (.71)	2.92 (1.35)	2.33 (1.11)
Judging time	4.32 (.91)	5.12 (1.01)	5.49 (1.42)	4.55 (.87)	4.16 (.61)	5.15 (.67)	4.68 (1.39)	3.82 (.85)

Note: Amount of Disclosure 1 was the word-count measure. Amount of Disclosure 2 was the coders' ratings. The standard deviations are in the parentheses. "Hnoid" is an abbreviation for "humanoid." Judging time was in seconds.

Table 2 Analysis of Variance Interaction Results for Dependent Measures in Study 1

	Males		Females		All Participants			
	Face × Voice		Face × Voice		Face × Voice		Three-Way Interaction	
Dependent Variables	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
Consistency	36.72***	.51	38.08***	.51	74.76***	.51	.00	.00
Amount of Disclosure 1	14.47***	.29	5.48*	.13	17.02***	.19	.11	.00
Amount of Disclosure 2	6.83**	.16	4.43*	.11	10.96***	.13	.03	.00
Intimacy of disclosure	.64	.02	2.59	.07	2.82†	.04	.25	.00
Trust of agent	11.05**	.24	4.45*	.11	13.27***	.16	.04	.00
Hesitancy to disclose	.42	.01	6.60*	.16	1.8	.02	5.13*	.07
Negativity of agent	.11	.00	17.69***	.33	10.71**	.13	7.93**	.10
Strangeness of agent	.02	.00	11.23**	.24	6.96**	.09	6.03*	.08
Liking of the face	.13	.00	18.88***	.34	10.03**	.12	6.96**	.09
Liking of the voice	.47	.01	9.37**	.21	7.13**	.09	2.96	.04
Judging time	6.58*	.16	9.82**	.21	15.94***	.18	.02	.00

Note: Degrees of freedom was (1, 36) for males and females and was (1, 72) for all participants. †$p = .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

Indicated by the objective word count and the subjective coders' ratings, both male and female participants disclosed more about themselves when the human face was paired with the human voice compared to the humanoid voice and when the humanoid face was paired with the humanoid voice compared to the human voice. The face–voice interaction remained significant in the three-way ANOVA. The three-way interaction was not significant. Although both measures of the amount of self-disclosure showed the same pattern of results, the word-count measure appeared to yield even more significant results, particularly among males. This might be due to the larger range of the word-count data.

For the rated intimacy level of the self-disclosure, the face–voice interaction was not significant for males ($p = .43$) but approached significance for females ($p = .12$). In the three-way ANOVA with all participants, the face–voice interaction approached significance at $p = .10$. One plausible explanation for the lack of significance at the conventional level for the intimacy data is that all nine self-disclosure questions are quite intimate (e.g., about guilt, death, or disliked things about best friend). Thus, if substantive responses were given, they were likely to have a relatively high baseline for intimacy. This argument receives some support from examining the means, which were all above 3.4 on a 1–7 rating scale. In comparison, participants were free to write as much or as little in their disclosure, leaving more room for variance in the amount of self-disclosure. The apparent difference in the *p* level for males and females might suggest greater sensitivity of females in the intimacy of self-disclosure as a reaction to the face–voice pairing of the agent. However, the

three-way interaction including gender was not significant. The gender difference became elucidated in the attitudinal measures.

In addition, self-reported trust of the talking-face agent yielded significant face–voice interaction effects supporting the consistency hypothesis. Participants of both genders reported greater trust of the talking-face agent when it was the human face with the human voice compared to the humanoid voice and when it was the humanoid face with the humanoid voice compared to the human voice. The face-voice interaction in support of consistency remained significant in the three-way ANOVA. The three-way interaction was not significant.

Female participants felt more hesitant to disclose their personal information when the face and the voice of the talking-face agent were human–humanoid mixed than when they were not. The face–voice interaction was not significant for males. This gender difference caused a significant three-way interaction effect. The face–voice interaction was not significant in the three-way ANOVA.

In summary, Hypothesis 1 was clearly supported by the results. Participants of both genders disclosed less personal information to the talking-face agent and found the agent less trustworthy when the face and voice were inconsistent. The same pattern was yielded in females' reported hesitance to disclose and approached significance for the intimacy level of the self-disclosure, particularly among females.

Testing H2: inconsistency effect on negative attitudes

A significant face–voice interaction effect showed that female participants perceived the talking-face agent more negatively when the human face was paired with the humanoid voice compared to the human voice and when the humanoid face was paired with the human voice compared to the humanoid voice. The face–voice interaction was not significant for males. This gender difference caused a significant three-way interaction effect. The face–voice interaction supporting the consistency hypothesis was significant in the three-way ANOVA.

The same pattern of findings emerged for the perceived strangeness of the talking-face agent. Female participants perceived the talking-face agent to be stranger when inconsistent than when consistent. Male participants showed no significant face–voice interaction effect. This gender difference caused a significant three-way interaction effect. The face–voice interaction effect was significant in the three-way ANOVA.

Female participants showed significant face–voice interaction effects in terms of their liking of the face alone and their liking of the voice alone. Females liked the human face more when it was paired with the human voice compared to the humanoid voice and liked the humanoid face more when it was paired with the humanoid voice compared to the human voice. Males did not show a significant face–voice interaction effect. This gender difference caused a significant three-way interaction effect. The face–voice interaction was significant in the three-way ANOVA. Similarly, female participants liked the human voice more when it was paired with the human face compared to the humanoid face and liked the humanoid voice more when it was paired with the humanoid face compared to the human face.

The face–voice interaction was not significant for males. The three-way interaction was not significant for liking of the voice. The face–voice interaction was significant in the three-way ANOVA.

Hence, Hypothesis 2 was supported for female participants but not for male participants. Females perceived the talking-face agent to be more negative and stranger when it was the human face talking with the humanoid voice compared to the human voice and when it was the humanoid face talking with the human voice compared to the humanoid voice. This inconsistency effect on negative attitudes was not found among males. Although the face and the voice were the same across all four pairing combinations, females liked the same face or the same voice more when it was paired with a consistent counterpart than with an inconsistent counterpart. This result provides further evidence that females seemed to be attitudinally sensitive to the face–voice consistency in a talking face.

Testing H3: effect of inconsistency on judging time

Time that the participants spent in forming judgment about the talking-face agent also showed significant face–voice interaction effects in support of the consistency hypothesis for both genders. Both male and female participants spent more time judging the agent when the human face was paired with the humanoid voice compared to the human voice and when the humanoid face was paired with the human voice compared to the humanoid voice. The face–voice interaction remained significant in the three-way ANOVA. The three-way interaction was not significant. Thus, Hypothesis 3 was supported, in accordance with the literature that inconsistency causes longer time and more effort in processing (Fiske & Neuberg, 1990; Schul et al., 1983).

Testing H4: gender difference in the inconsistency effects

The results showed that the participants of both genders showed consistency preference responses in terms of the amount of self-disclosure, trust, and time in forming judgment of the talking face. But female participants also exhibited significant consistency preference responses in all other attitudinal measures, whereas males did not. Females showed significant consistency preference responses in their perceived negativity and strangeness of the talking face, hesitancy to disclose, and liking of the face and the voice. Males did not show significant effects for these measures. The differences in the responses to these measures between the two gender groups caused significant three-way interaction effects for all these measures except for liking of the voice. Using a sign test (Siegel & Castellan, 1988), the number of measures for which females showed significant consistency preference effects (five) was significantly greater than the number for males (zero), $p < .05$. The N for the Sign test is the number of the measures for which the compared groups differed. It is five in this case. The measures for which both gender groups showed the same effects were excluded because they are considered ties and are not included in a sign test. Therefore, Hypothesis 4 received support showing more negative effects of inconsistency on females than on males in terms of negative attitudes toward the talking face.

Discussion

If computer-generated humanoid faces and voices are perceived categorically different from natural human faces and voices, a computer agent with a human face and a humanoid voice or a humanoid face and a human voice will be perceived as inconsistent. Literatures on perception, person perception, and verbal and nonverbal communication processing have shown that inconsistency leads to more negative attitudes, specifically hurts trust, and causes longer processing time. The results of Study 1 overall replicated these inconsistency effects. Both male and female participants took longer time in making judgment about the talking-face agent when it was the human face talking with the humanoid voice compared to the human voice or when it was the humanoid face talking with the human voice compared to the humanoid voice. This result aligns with the research that inconsistency causes longer processing time (Schul et al., 1983) and fits with the theory that inconsistency follows a piece-processing model and requires more processing effort (Fiske & Neuberg, 1990).

Both male and female participants behaviorally disclosed less personal information about themselves to the talking-face agent and reported less trust of the agent when the face and the voice of the agent were human and humanoid mixed than when they were not. Research in interpersonal communication has shown that trust is an important factor accounting for self-disclosure (Wheeless & Grotz, 1977). Thus, the results lend support to the hypothesized inconsistency effect on trust and trust-related behavior for talking faces. Furthermore, the explicit measure asking the participants to judge the consistency between the face and the voice of the talking face confirmed that both males and females thought that the human–humanoid mixed talking faces were more inconsistent than the nonmixed talking faces. All these results demonstrate inconsistency effects when the face and voice of a talking face are mixed in the human versus humanoid dimension and suggest that human and humanoid are perceived as two separate categories.

In addition, female participants exhibited significant consistency preference in their attitudinal responses toward the talking face, whereas males did not. Females reported more hesitancy to disclose to the agent, perceived the agent more negatively, and found it stranger when the agent had inconsistent face–voice pairings than when it had consistent pairings. Female users also liked the face or the voice more when it was paired with a counterpart in the same category. Therefore, females seem to be more sensitive to the face–voice inconsistency and have a stronger attitudinal preference for consistency than males. This gender difference may be explained by the general pattern that females are overall more attentive and sensitive to nonverbal cues and more susceptible to the influences of nonverbal cues than males (Argyle et al., 1970; Mazanec & McCall, 1976; Zahn, 1973).

Thus, it seems that both males and females hold consistency preference as reflected in their automatic response in processing time and in their trust-related responses, but females' consistency preference is overall more evident and stronger than males in terms of overt attitudinal judgment.

The obtained inconsistency effects might be limited by the nature of the experimental task in this study. The task, a one-time interaction with a zero-history interviewing agent asking intimate personal questions, might heighten the trust factor and the tendency to scrutinize. Driven by scrutiny and vigilance, people might be more attentive to any inconsistencies and unusual features of the talking-face agent. As a result, they might be more guarded and less trustful when the face and the voice appear inconsistent. A further test of the consistency thesis should involve an experimental context that does not heighten trust, but instead, stresses the demand of technological modular excellence, particularly in the voice modality. The one-sentence disclosure questions were not difficult to understand in any of the face–voice combinations in this study. But would the voice excellence factor become more influential when the spoken content poses greater comprehension demand? This reasoning motivated Study 2.

Study 2

Humans have an innate desire to understand the speech of other humans when they hear it (Nass & Gong, 2000; Slobin, 1979). Research has shown that people are much less tolerant of auditory degradation than visual degradation (Reeves, Detenber, & Steuer, 1993). Computer-generated humanoid speech might pose greater difficulty for processing than computer-generated humanoid faces. If a humanoid voice is hardly comprehensible, even if it is consistently matched with an obviously synthetic humanoid face, people might not be able to undertake any meaningful interaction with the agent beyond the first impression and understanding short messages. Instead, people should still be able to see the features of a humanoid face and its movements and expressions, despite its artificial appearance.

Existing research on processing computer-synthesized speech shows that synthetic speech is challenging for comprehending long messages but not so much for short messages, such as phrases and one or two sentences. Lai, Cheng, Green, and Tsimhoni (2001), for example, reported that users had more than 90% accuracy in understanding phrases spoken by either synthetic speech or natural human speech. When verbal messages became long, the synthetic speech became much more difficult to understand than the natural speech. Therefore, when speech comprehension is demanding, it is unsure which of the two, the face–voice consistency preference or the speech single-modality excellence, will prevail. Thus, Study 2 incorporated an experimental task heightening the demand for speech comprehension and assessed comprehension-related responses. It also replicated assessments of negative attitudes and time in making judgment to be consistent with Study 1 and the literature. Because lack of prior research on comparing face–voice consistency concern and voice modality excellence for talking-face agents, a research question was proposed:

RQ1: When comprehension of speech is demanding, will the preference for face–voice consistency or the need for voice modality excellence influence people's

comprehension-related responses, negative attitudes toward the talking face, and time in judging the talking face?

Method

The same 2 × 2 × 2 experimental design was used in Study 2. The same four combinations of talking faces were used.

Participants

Another sample of 80 undergraduate students (40 men and 40 women) enrolled in a large communication course at the same university were recruited in a subsequent academic term. Selection of English native speakers, random assignment of 10 males and 10 females to each type of talking face, and the reward with extra course credit were the same in Study 2 as in Study 1.

Procedure

Study 2 situated the human–agent interaction in a task involving listening to book reviews orally delivered by the talking face. The talking face was framed as a book-presenting agent on a prototype book Web site. The agent embodied in one of the four face–voice combinations delivered five book reviews orally, one at a time. The text of the book reviews was not displayed on the computer screen. Two sample books were "Windfall" by James Magnuson and "Last Things" by Jenny Offill. The length of the book reviews ranged from 134 to 193 words. The participants were told in the instruction that they could only listen to the book reviews once. No repeat function was offered to stress the comprehension demand of the task. After they heard a book review, they clicked a "continue" button to proceed to a new screen and answered a series of questions assessing comprehension-related responses.

The 9.60- × 7.15-cm video frame of the talking face was placed on the left center of the screen, with a 3.10- × 4.30-cm picture of the book cover displayed on the right side of the screen. The first book was treated as a practice round. After going through the five book reviews, the participants filled out a posttask questionnaire on the computer. We chose five fiction books that were unlikely to have been read by most undergraduate students in that university. There were questions after each book regarding whether the participants had read the book or other book(s) by the same author. No one had. The book reviews were modified from the reviews of the books at Amazon.com.

Measures

Comprehension-related responses assessed after each book review were perceived understanding of the book review, evaluation of the book review, and behavioral intent about the book. Perceived understanding of the book review was measured with the question: "How well did you understand the book review?" on a 10-point Likert scale (1 = *very poorly*, 10 = *very well*). The responses to this question for the

four books were averaged; the index was very reliable ($\alpha = .88$). Evaluation of the book review consisted of two questions: "How would you rate the quality of the book review?" (on a 10-point Likert scale, 1 = *very low quality*, 10 = *very high quality*) and "How much did you like the book review?" (on a 10-point Likert scale, 1 = *not at all*, 10 = *very much*). The averaged responses to these questions across the four books formed the index. The average reliability of this index across the books was very high ($\alpha = .90$). Behavioral intent about the book included three items: "How likely would you be to enjoy reading this book?," "How likely would you be to recommend this book to your friends?," and "How likely would you be to buy this book?" These three questions were answered on 10-point Likert scales (1 = *very unlikely*, 10 = *very likely*). The averaged responses to these questions across the four books formed this index. The average reliability among the three items was very high ($\alpha = .91$).

The posttask questionnaire included the same measures of perceived negativity and strangeness of the agent as in Study 1. Also as in Study 1, time spent per item in the questionnaire constituted the measure of time spent in judging the talking-face agent.

Results

The same analysis procedure was followed as in Study 1. Table 3 presents the means and standard deviations of all the dependent measures. Table 4 presents the ANOVA results for testing the interaction effects between the face and the voice factors. In the case of lack of higher-order interaction effects, the lower-order effects are presented in the text.

The interaction between the face and the voice was not significant for either gender group for the three comprehension-related measures, nor was it significant in the three-way ANOVA. The three-way interaction was also not significant. Instead, significant voice modality main effects emerged. Both female and male participants thought they understood the book reviews better when the reviews were delivered by the human voice than by the humanoid voice regardless of the face (females: $F(1, 36) = 59.91$, $p < .001$, $\eta^2 = .63$; males: $F(1, 36) = 5.75$, $p < .05$, $\eta^2 = .14$). The voice main effect was also significant in the three-way ANOVA, $F(1, 72) = 48.67$, $p < .001$, $\eta^2 = .40$. Additionally, a significant gender-voice interaction effect emerged in the three-way ANOVA. Both genders preferred the human voice to the humanoid voice, but females showed stronger preference, $F(1, 72) = 11.74$, $p < .001$, $\eta^2 = .14$.

Similarly, the book reviews were evaluated more positively by both females and males when the reviews were delivered by the human voice than by the humanoid voice regardless of the face (females: $F(1, 36) = 45.23$, $p < .001$, $\eta^2 = .56$; males: $F(1, 36) = 9.92$, $p < .01$, $\eta^2 = .22$; overall: $F(1, 72) = 50.96$, $p < .001$, $\eta^2 = .41$). There was a significant gender-voice interaction effect indicating strong preference for the human voice to the humanoid voice among females, $F(1, 72) = 8.98$, $p < .01$, $\eta^2 = .11$.

The behavioral intent about the books were also more positive when the reviews of the books were delivered by the human voice than by the humanoid voice

Table 3 Means and Standard Deviations of Dependent Measures in Study 2

	Males				Females			
	Human Face		Humanoid Face		Human Face		Humanoid Face	
	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice	Human Voice	Hnoid Voice
Perceived understanding	8.08 (.96)	6.93 (.65)	7.31 (1.83)	5.85 (2.67)	7.94 (1.22)	3.85 (1.89)	7.81 (1.06)	4.27 (1.88)
Evaluation of reviews	6.37 (1.46)	4.95 (1.22)	5.72 (.85)	4.47 (1.68)	6.49 (1.44)	3.90 (2.01)	7.02 (1.13)	3.08 (1.42)
Intent about books	5.03 (.46)	4.40 (.84)	4.62 (.82)	3.88 (1.23)	4.44 (.91)	3.56 (1.34)	4.47 (.85)	2.60 (.44)
Negativity of agent	3.43 (1.31)	4.14 (2.22)	3.22 (1.46)	4.47 (2.80)	2.90 (1.12)	7.22 (1.65)	6.01 (1.85)	4.96 (1.82)
Strangeness of agent	4.15 (1.60)	5.13 (1.93)	4.23 (2.12)	5.37 (2.76)	4.30 (1.12)	6.28 (1.26)	6.24 (1.10)	4.20 (2.05)
Judging time	4.52 (1.12)	5.13 (1.08)	5.24 (1.09)	4.18 (.61)	4.34 (1.03)	5.18 (1.13)	4.98 (.42)	4.02 (.79)

Note: The standard deviations are in the parentheses. "Hnoid" is an abbreviation for "humanoid." Judging time was in seconds.

Table 4 Analysis of Variance Interaction Results in Study 2

	Males		Females		All Participants			
	Face × Voice		Face × Voice		Face × Voice		Three-Way Interaction	
Dependent Variables	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
Perceived understanding	.08	.00	.31	.01	.03	.00	.34	.01
Evaluation of reviews	.04	.00	1.94	.05	.84	.01	1.40	.02
Intent about books	.04	.00	2.74	.07	1.82	.03	1.14	.02
Negativity of agent	.21	.01	26.90***	.43	9.58**	.12	14.31***	.17
Strangeness of agent	.01	.00	19.64***	.35	5.63*	.07	6.53**	.08
Judging time	7.04**	.16	10.16**	.22	16.82***	.19	.02	.00

Note: Degrees of freedom was (1, 36) for males and females and was (1, 72) for all participants. $^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

regardless of the face (females: $F(1, 36) = 21.49$, $p < .001$, $\eta^2 = .37$; males: $F(1, 36) = 6.09$, $p < .05$, $\eta^2 = .15$; overall: $F(1, 72) = 25.71$, $p < .001$, $\eta^2 = .26$). There were no other significant effects for these three comprehension-related measures.[2] Therefore, the results indicate that for perceived comprehension and responses based on comprehension, the superiority of the human voice prevailed. There were no face–voice interaction effects.

However, significant face–voice interaction effects did emerge for time in judging the talking-face agent for both genders and perceived negativity and strangeness of the agent among females. Both male and female participants took longer time in forming judgment of the talking face when it was the human face talking with the humanoid voice compared to the human voice or when it was the humanoid face talking with the human voice compared to the humanoid voice. Female participants perceived the talking-face agent more negatively and found it stranger when it was the human face talking with the humanoid voice compared to the human voice or when it was the humanoid face talking with the human voice compared to the humanoid voice. There were no face–voice interaction effects or other significant effects for these two measures of negative attitudes among males. The significant face–voice effects due to females remained significant in the three-way ANOVAs and also caused significant three-way interaction effects.

Discussion

Interestingly, when outcomes hinge on comprehending long, spoken verbal messages, being able to understand the speech seems a stronger desire than the preference for face–voice categorical consistency. When understanding book reviews of more than 100 words in length posed a substantial demand, especially in the case of synthetic speech, users did not exhibit consistency preference in their comprehension-related responses. Instead, voice superiority prevailed. Users reported better understanding

of the reviews, evaluated the reviews more positively, and showed more positive behavioral intent about the books when the reviews were delivered by the human voice than by the humanoid voice regardless of the face. It makes sense that if users could not comprehend the review of a book, their judgment of the book and the review would most likely be more negative than if it were more comprehensible.

However, inconsistency effects were still obtained for judging time among all participants and for negative attitudes among females, a pattern found in Study 1. Therefore, it seems only for comprehension-related outcomes that consistency preference is overwhelmed by preference for voice excellence. But even after a task that deliberately stressed speech comprehension demand, the participants' processing time and the females' negative attitudes still showed inconsistency effects rather than voice main effects. Therefore, the factor of voice modality excellence seems to only affect outcomes directly affected by voice clarity. The inseparability of face and voice in holistically perceiving a talking face still underlies preference for consistency between the face and the voice.

General discussion

The results of Study 1 and Study 2 showed inconsistency effects when mixing a human face with a humanoid voice or mixing a humanoid face with a human voice. In both studies, inconsistently paired talking faces incurred longer time for forming judgment about them than consistently paired talking faces, suggesting more processing efforts (Fiske & Neuberg, 1990; Schul et al., 1983). In both studies, female participants reported negative attitudes toward inconsistently paired talking faces, in accordance with the literature in terms of both the negative attitudes caused by inconsistency (Hendrick, 1972) and the females' sensitivity to nonverbal information (Hall, 1978, 1984). The self-disclosure task in Study 1 revealed negative effects that inconsistent pairings of human versus humanoid faces and voices had on trust and the amount of personal information disclosed to the talking-face agent. This result accords with the literature, which shows that inconsistency cues deception (Ekman & Friesen, 1969, 1974). The comprehension-heightened task in Study 2 highlighted the demand for voice excellence, which probably overshadowed the face–voice consistency concern but only for comprehension-related outcomes. In terms of processing time and negative attitudes, inconsistency effects from mixing human versus humanoid faces and voices remained robust.

All these results strongly indicate that mixing a human face with a humanoid voice or mixing a humanoid face with a human voice incurs inconsistency. We can confidently attribute the inconsistency effects to the distinction between the categories human and humanoid.

The human–humanoid difference might seem obvious in the faces and voices used in this study. But there are two subtle distinctions to be made here. The first is to distinguish apparent differences between any two exemplar faces or voices and the underlying difference of the identity categories that the faces or the voices belong

to. If the face and the voice of two male Caucasians of the similar age group were mixed with perfect lip synchronization, no inconsistency would be revealed, suggesting no underlying identity differences. The second distinction concerns the divergence between the approach of modular technological excellence and the framework of consistency preference. As in the examples of Massaro (1998), Ritter et al. (1999), Kiesler et al. (1996), and Sproull et al. (1996), a humanoid face was paired with a human voice or a human face was paired with a humanoid voice, driven by the need for modular excellence. Human versus humanoid categorical inconsistency, which was revealed by this research, instead, was overlooked in these applications or studies. The results of this research should not be modified by the bimodality of talking faces if comparison is to be made to voice-only communication, because both Nass and Gong (1999) and Kiesler et al. found the consistency effect when comparing talking face to voice only. The face–voice interaction effect for liking of voice among female participants in this research further confirmed this pattern.

The human–humanoid inconsistency effects indicate that humanoid is perceived to be different from human. But what is exactly humanoid? The computer-synthesized face and speech used in this research do typify many computer-synthesized faces and voices. For example, a similar-looking synthetic face was used in the study by Burgoon et al. (2000). The synthetic speech engine used by Burgoon et al. (2000) is the same as the one in this research. It is similar to the one used by Kiesler et al. (1996) and Sproull et al. (1996). But, humanoid, which is defined by this research as being anthropomorphic but clearly artificial and synthetic, is likely a continuum, if not a multidimensional construct.

Computer-synthesized faces and speech are increasingly varying in terms of how much they resemble faces and speech of real humans. Visual appearance of computer-synthesized faces can range from being very humanlike with all the nuances in lighting and hair to being only slightly humanlike. Computer-synthesized speech has also recently advanced to take on a larger variation by introducing the concatenation technique, which involves recording a large corpus of individual phonemes of natural voices and then combining them in real time to generate speech. This new technique results in computerized speech, which is natural sounding in its pronunciation but still somewhat synthetic sounding in its prosody (samples can be heard at http://www.naturalvoices.att.com/demos).

The field of communication is only starting to examine the degree of anthropomorphism (Nowak, 2004) and has not yet developed theoretical frameworks for it. The field needs systematic research to understand how people identify and categorize entities that are humanlike to different degrees. Do people distinguish all computer-generated, humanlike entities from natural human ones? Or, do people associate those that highly resemble natural humans with humans and treat those that do not as humanoid?

Certainly, human faces and voices that may be drawn to be compared to humanoid faces and voices are infinite. What is more theoretically interesting is to compare

prominent social identity categories such as race and gender to human versus humanoid. Social identity theory (Tajfel & Turner, 1986) highlights the power of group identity. It will be a new valuable venue of research to compare the power of social identity and human versus humanoid identity on people's preferences and responses. Specifically, would in-group preference be guided by a social identity such as race and gender regardless of the human or humanoid identity? Or, would natural human renditions prevail over humanoid ones subsuming social identity categories? In other words, social identities do apply to humanoid entities according to the Computers Are Social Actors paradigm (Reeves & Nass, 1996), but when human and humanoid entities are directly compared, it awaits future research to test which dimension of identity is more influential.

Another valuable venue of future research is comparing digital entities framed as agents versus avatars. For example, E.-J. Lee and Nass (2002) compared computer character representations that were framed for human–computer interaction versus computer-mediated communication. Addition of real or ostensible people to the equation adds a third perspective for judging identity categorical consistency. Despite the complexity, it adds rich opportunities for contrasting the visual immediacy from the computer representations and the ontological power of the humans being represented.

Another area that begs further research is the nature of context in which computer-generated characters interact with people. This research highlighted the importance of context on context-specific outcomes but nonetheless showed consistency as a general preference. Another context of particular importance is entertainment. The animation entertainment industry almost uniformly uses recorded human voices for computer-generated characters. For example, the popular American computer animation film "Shrek" used the recordings of human actress Cameron Diaz for the voice of the computer-animated character Princess Fiona (http://www.imdb.com/title/tt0298148). With computer-generated speech being available, why do film animators not seem to worry about the consistency concern? And why do moviegoers or film critics not seem to be bothered by the possible inconsistency?

There seem three explanations. The first explanation follows from the conceptual discussion of the degree of anthropomorphism and human identity. Princess Fiona clearly is more humanlike than the Baldi face used in this study. If people do not treat all computer-generated characters the same and differentiate them by their degrees of human resemblance, a voice that sounds more natural than the one used in this research would be needed to match with the face of Princess Fiona. If such a voice is feasible with computer synthesis, would it be preferable to Diaz's voice?

The second explanation concerns character design and expressions. Characters in animation films often have vivid and exaggerated expressions. The technology of computer graphics and animation has advanced to the extent of displaying vivid and exaggerated facial expressions in characters but computer-generated speech has not. Professional voice actors are trained to modify their voices to resemble high-pitched anthropomorphic mice or low-pitched ducks, for example. In the entertainment

context featuring caricatures and exaggeration, expressiveness may be the most salient factor and criterion.

The third explanation speculates that people may follow different cognitive processing routes and activate different factors in judging a character in an entertainment versus a nonentertainment context. Leveraging the Elaboration Likelihood Model (Petty & Cacioppo, 1986), Slater and Rouner (2002) pointed out that entertainment persuasive narratives are less likely to activate central processing than nonentertainment ones. It is plausible that people may be less likely to scrutinize the issue of identity consistency in the face and voice of an entertainment digital character, compared to characters that solicit personal information or deliver book reviews.

In addition to the domain of the context, other attributes of the context and interaction should be theorized, particularly leveraging interpersonal communication theories (Berger, 2005). This research employed one-time interactions. But digital communicators are increasingly used for repeated or long-term interactions, for example, as preferred virtual news anchor or personal assistant. Would face–voice consistency be more prominent in interactions expected to last? Or, would any cues of natural humans help comfort users (e.g., the elderly) for relational use of digital communicators? Expectation is also a valuable factor to incorporate in future theorization and examination, particularly leveraging the Expectancy Violations Theory (Burgoon & Le Poire, 1993). First, baseline of expectations needs to be assessed and compared for natural human representations and digital humanoid ones. Then, the examination of how judgment will be adjusted based on gauging performance and outcomes against expectations will illuminate the processes underlying cognition of and responses to human versus humanoid entities.

Future research should also need to examine people's experiences with digital communicators and environments. The participants in this research reported no exposure to the faces and the voices used. But they supposedly have much more experiences with human faces and voices in general. Interestingly, this experiential difference did not predict a main effect for human face or voice. The humanoid-face/humanoid-voice pairing did not receive the most negative responses, either. The interaction effect indicating consistency illustrates strong preference for categorical matching. When face and voice were matched, the human-face/human-voice combination clearly elicited more positive responses. Nonetheless, assessment of participants' prior exposure to various types of digital communicators including game avatars, computer agents, and animated characters will be still valuable. Specifically, would familiarity with digital characters predict even stronger consistency preference and also boost evaluations for the humanoid-face/humanoid-voice pairing? Considering the relative newness of digital communicators, exploratory research such as focus group and interview might be useful to help uncover issues and processes among novel and experienced users. Leveraging the Uncertainty Reduction Theory (Berger, 1986; Berger & Calabrese, 1975), which particularly highlights the prominence of uncertain reduction in initial interactions, exploratory research also has

value for identifying aspects receiving most uncertainty reduction efforts for humanoid entities and human entities. Research can also explore different uncertain reduction strategies used for collecting information for these types of identities and when entities incorporate mismatched elements.

In conclusion, this two-study research demonstrated that consistency preference is a general pattern underscored by the human versus humanoid categorical difference. Context-specific outcomes are primarily influenced by the quality of the component most pertinent to the judgment criterion of the context such as trust or voice clarity. As computer-generated anthropomorphic entities become more diverse and widely spread in digital media, the fundamental question of what is humanoid *and* what is human will become more tractable through systematic variations of the degree of anthropomorphism, social identities, and the nature of the interaction context.

Notes

1 The CSLU Toolkit can be downloaded for free for noncommercial research purpose from http://cslu.cse.ogi.edu/toolkit. The Toolkit is made by the CSLU at Oregon Graduate Institute.
2 Repeated-measures ANOVAs were also conducted to compare responses to the shortest book review and the longest one to discern any possible difference due to review length. No difference emerged due to review length.

References

Argyle, M., Alkema, F., & Gilmour, R. (1971). The communication of friendly and hostile attitudes by verbal and non-verbal signals. *European Journal of Social Psychology*, **1**, 385–402.

Argyle, M., Salter, V., Nicholson, H., Williams, M., & Burgess, P. (1970). The communication of inferior and superior attitudes by verbal and nonverbal signals. *British Journal of Social and Clinical Psychology*, **9**, 222–231.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, **41**, 1230–1240.

Baylor, A. L., & Kim, Y. (2003). *The role of gender and ethnicity in pedagogical agent perception*. Paper presented at the E-Learn (World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education), Phoenix, AZ. Retrieved July 8, 2004, http://pals.fsu.edu/publications.html.

Berger, C. R. (1986). Uncertain outcome values in predicted relationships: Uncertainty reduction theory then and now. *Human Communication Research*, **13**, 34–38.

Berger, C. R. (2005). Interpersonal communication: Theoretical perspectives, future prospects. *Journal of Communication*, **55**, 415–447.

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research*, **1**, 99–112.

Bickmore, T., & Cassell, J. (2001). Relational agents: A model and implementation of building user trust. In J. A. Jacko, A. Sears, M. Beaudouin-Lafon & R. J. K. Jacob (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'01)* (pp. 396–403). Seattle, WA: ACM Press.

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, **3**. Retrieved November 28, 2004, http://www.ascusc.org/jcmc/vol2003/issue2002/biocca2002.html.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human-computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior*, **16**, 553–574.

Burgoon, J. K., & Le Poire, B. A. (1993). Effects of communication expectancies, actual communication, and expectancy disconfirmation on evaluation of communicators and their communication behavior. *Human Communication Research*, **20**, 67–96.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Dindia, K., Fitzpatrick, M. A., & Kenny, D. A. (1997). Self-disclosure in spouse and stranger interaction: A social relations analysis. *Human Communication Research*, **23**, 388–412.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, **32**, 88–95.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality & Social Psychology*, **29**, 288–298.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic Press.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. New York: McGraw-Hill.

Gibbs, J. L., Ellison, N. B., & Heino, R. D. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in internet dating. *Communication Research*, **33**, 152–177.

Golinkoff, R. M., & Rosinski, R. R. (1976). Decoding, semantic processing, and reading comprehension skill. *Child Development*, **47**, 252–258.

Gong, L., & Lai, J. (2003). To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. *International Journal of Speech Technology*, **6**, 123–131.

Grabner-Kräuter, S., & Kaluscha, E. A. (2003). Empirical research in on-line trust: A review and critical assessment. *International Journal of Human-Computer Studies*, **58**, 783–812.

Graves, J. R., & Robinson, J. D. (1976). Proxemic behavior as a function of inconsistent verbal and nonverbal messages. *Journal of Counseling Psychology*, **23**, 333–338.

Green, E. J., & Barber, P. J. (1981). An auditory stroop effect with judgments of speaker gender. *Perception and Psychophysics*, **30**, 459–466.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, **85**, 845–857.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore: Johns Hopkins University Press.

Hamers, J. F., & Lambert, W. E. (1972). Bilingual interdependencies in auditory perception. *Journal of Verbal Learning and Verbal Behavior*, **11**, 303–310.

Hendrick, C. (1972). Effects of salience of stimulus inconsistency on impression formation. *Journal of Personality & Social Psychology*, **22**, 219–222.

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, **53**, 251–267.

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality & Social Psychology*, **70**(1), 47–65.

Lai, J., Cheng, K., Green, P., & Tsimhoni, O. (2001). On the road and on the Web? Comprehension of synthetic and human speech while driving. In J. A. Jacko, A. Sears, M. Beaudouin-Lafon & R. J. K. Jacob (Eds.), *Proceedings of the Conference on Human Factors in Computing Systems (CHI'01)* (pp. 206–212). Seattle, WA: ACM Press.

Lee, E.-J. (2004). Effects of visual representation on social influence in computer-mediated communication: Experimental tests of the social identity model of deindividuation effects. *Human Communication Research*, **30**, 234–259.

Lee, E.-J., & Nass, C. (2002). Experimental tests of normative group influence and representation effects in computer-mediated communication: When interacting via computers differs from interacting with computers. *Human Communication Research*, **28**, 349–381.

Lester, J. C., Stone, B. A., Converse, S. A., Kahler, S. E., & Barlow, S. T. (1997). Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical investigation. In B. D. Boulay & R. Mizoguchi (Eds.), *Proceedings of the 8th World Conference on Artificial Intelligence in Education* (pp. 23–30). Kobe, Japan: IOS Press.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin*, **109**, 163–203.

Maes, P. (1994). Agents that reduce work and information overload. *Communication of the ACM*, **37**(7), 31–40, 146.

Massaro, D. M. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Mazanec, N., & McCall, G. J. (1976). Sex factors and allocation of attention in observing persons. *Journal of psychology*, **93**, 175–180.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Mehrabian, A. (1971). When are feelings communicated inconsistently? *Journal of Experimental Research in Personality*, **4**(3), 198–212.

Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research*, **26**, 323–339.

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

Nass, C., & Gong, L. (1999). Maximized modality or constrained consistency? In D.W. Massaro (Ed.), *Proceedings of the International Auditory-Visual Speech Processing Conference* (pp. 1–5). Santa Cruz, CA: AVSP Press.

Nass, C., & Gong, L. (2000). Social aspects of speech interfaces from an evolutionary perspective: Experimental research and design implications. *Communications of the ACM*, **43**(9), 36–43.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, **7**(3), 171–181.

Nowak, K. L. (2004). The influence of anthropomorphism and agency on social judgment in virtual environments. *Journal of Computer-Mediated Communication*, **9**. Retrieved September 7, 2005, from http://jcmc.indiana.edu/vol2009/issue2002/nowak.html.

Olive, J. P. (1997). "The talking computer": Text-to-speech synthesis. In D. G. Stork (Ed.), *Hal's legacy*: *2001's computer as dream and reality* (pp. 101–131). Cambridge, MA: MIT Press.

Parke, F. I., & Waters, K. (1996). *Computer facial animation*. Wellesley, MA: A K Peters.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Pomerantz, J. R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, vol. 2: Cognitive processes and performance* (pp. 36–46). New York: John Wiley.

Reeves, B., Detenber, B., & Steuer, J. (1993, May). *New televisions: The effects of big pictures and big sound on viewer responses to the screen*. Paper presented at the Information Systems Division of the International Communication Association, Chicago.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.

Ritter, M., Meier, U., Yang, J., & Waibel, A. (1999). Face translation: A multimodal translation agent. In D.W. Massaro (Ed.), *Proceedings of the International Auditory-Visual Speech Processing Conference* (pp. 163–167). Santa Cruz, CA: AVSP Press.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception and Psychophysics*, **59**, 347–357.

Roy, L., & Sawyers, J. K. (1990). Interpreting subtle inconsistency and consistency: A developmental-clinical perspective. *Journal of Genetic Psychology*, **151**, 515–521.

Schul, Y., Burnstein, E., & Martinez, J. (1983). The informational basis of social judgments: Under what conditions are inconsistent trait descriptions processed as easily as consistent ones? *European Journal of Social Psychology*, **13**, 143–151.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences*. Boston: McGraw-Hill.

Slater, M. D., & Rouner, D. (2002). Entertainment-education and elaboration likelihood: Understanding the processing of narrative persuasion. *Communication Theory*, **1**, 173–191.

Slobin, D. I. (1979). *Psycholinguistics* (2nd ed.). Glenview, IL: Scott, Foresman.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, **11**, 97–124.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, **18**, 643–663.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago: Nelson-Hall.

Tanaka, W. (2000, October 15). Face to face: Talking images to add human touch to e-shops. *San Jose Mercury News*, pp. 3F–4F.

van Santen, J., Macon, M., Cronk, A., Hosom, P., Kain, A., Pagel, V, et al. (2000). When will synthetic speech sound human: Role of rules and data. In B. Yuan (Ed.), *Proceedings of*

International Conference of Spoken Language Processing (ICSLP 2000) (pp. 402–409). Beijing, China: China Military Friendship Publishing.

Volkmar, F. R., & Siegel, A. E. (1979). Young children's responses to discrepant social communications. *Journal of Child Psychology and Psychiatry*, **20**, 139–149.

Wheeless, L. R. (1978). A follow-up study of the relationships among trust, disclosure, and interpersonal solidarity. *Human Communication Research*, **4**, 143–157.

Wheeless, L. R., & Grotz, J. (1977). The measurement of trust and its relationship to self-disclosure. *Human Communication Research*, **3**, 250–257.

Zahn, G. L. (1973). Cognitive integration of verbal and vocal information in spoken sentences. *Journal of Experimental Social Psychology*, **9**, 320–334.