Are Computers Gender-Neutral?

Gender Stereotypic Responses to Computers

Clifford Nass, Youngme Moon, and Nancy Green

Department of Communication

Stanford University

Stanford, CA 94305-2050

(415) 723-5499

e-mail: nass@leland.stanford.edu

Running head: ARE COMPUTERS GENDER-NEUTRAL?

Journal of Applied Social Psychology (in press)

Abstract

This study tested whether computers embedded with the most minimal gender cues will evoke sex-based stereotypic responses. Using an experimental paradigm (N=40) that involved computers with voice output, the study tested three sex-based stereotypes under conditions in which all suggestions of gender were removed, with the sole exception of vocal cues. In all three cases, gender stereotypic responses were obtained. Because the experimental manipulation involved no deception regarding the source of the voices, this study presents evidence that the tendency to gender stereotype is extremely powerful, extending even to inanimate machines.

Are Computers Gender-Neutral?

Gender Stereotypic Responses to Computers

The pervasiveness of gender stereotypes is well-documented in the psychological literature. The general finding is that men are perceived to possess more instrumental attributes (i.e., self-directing, goal-oriented characteristics such as independence, assertiveness, and decisiveness), and fewer expressive attributes (i.e., emotive qualities such as kindness, sensitivity, emotional responsiveness, and need for affiliation) than women (Rosenkrantz, Vogel, Bee, Broverman, & Broverman, 1968). Despite recent upward trends in female employment and education, these gender-based stereotypes have persisted over time (e.g., Romer & Cherry, 1980; Ruble, 1983; Spence, Helmreich, & Stapp, 1974).

The present study seeks to determine whether machines embedded with minimal gender cues will generate sex-based stereotypic responses. In an experimental manipulation, subjects are exposed to either a male- or female-voiced computer. The use of this computer-based manipulation allows all other gender cues – e.g., visual appearance, nonverbal communication (Briton & Hall, 1995) – to be effectively removed from the interaction. Subjects are not even led to believe that they are interacting with another human; rather, the voice is explicitly associated with the computer. The purpose of the study, then, is to determine if gender stereotypes are sufficiently powerful to not only influence responses under conditions in which gender cues are downplayed, but to influence responses under conditions in which gender cues are clearly irrelevant.

If this is so, then this finding would have profound theoretical and social implications. First, it would demonstrate that the tendency to gender-stereotype is not only deeply ingrained, but can be triggered by minimal gender cues, even when those cues are disembodied. It would also cast a new light on the use of voice technology in various machine interfaces. Indeed, if the predictions in the present study are met, it would suggest that – contrary to popular beliefs among designers and engineers – computers and other forms of technology are not gender-neutral in the eyes of users.

In particular, three gender stereotypes are tested. The first stereotype is, “evaluation from males is more valid than evaluation from females.” There is significant evidence that both men and women are prone to this stereotype. First, both men and women attend to male voices more intently than female voices (Robinson & McArthur, 1982); thus, evaluative comments delivered by male voices should resonate more powerfully than the same comments delivered by female voices. Second, “as agents of influence, men are regarded as more dominant and influential and as more effective leaders than women” (Eagly & Wood, 1982, p. 916).

A second stereotype concerns dominant behavior. Although most instrumental and expressive traits are regarded as desirable to some degree in both men and women, dominance and aggressiveness are regarded as undesirable in women but not in men (Costrich, Feinstein, Kidder, Maracek, & Pascale, 1975; Deutsch & Gilbert, 1976; McKee

& Sheriffs, 1959; Pleck, 1978; Spence *et al*., 1974). When males are placed in dominant roles, they tend to be perceived as being "assertive," or "independent." When females are placed in dominant roles, they tend to be perceived as being "pushy" or "bossy."

The third stereotype is that "women know more about subjects that are typically regarded as 'feminine,' whereas men know more about subjects that are typically regarded as 'masculine.'" Evidence for this stereotype is mixed. There is evidence that people do engage in sex-based stereotyping across subjects. Some occupations, for example, have been identified as either "masculine" or "feminine"; these stereotypes affect the entrance of men and women into these occupations (e.g., Heilman, 1979). On the other hand, at least one study examining the relative ratings of authors writing on "feminine" or "masculine" subjects found that papers attributed to male authors were rated more highly across *both* masculine and feminine subjects (Goldberg, 1968).

The present study tests the strength of all three of these stereotypes, using an experimental manipulation involving voice-based computers. All other suggestions of gender are removed. In so doing, this study informs the question, "Are machines gender-neutral?" Throughout the experiment, subjects are fully aware of the fact that the voice belongs to a computer rather than to another human; thus, to the extent that stereotypic responses occur under these conditions, it would appear that the tendency to gender stereotype is very strong indeed, extending even to inanimate machines.

Method

Respondents

Respondents were 40 undergraduate volunteers. All respondents had extensive word-processing experience. Respondents were told they would be participating in a study that involved the use of a computer tutor. All respondents signed informed consent forms, and were fully debriefed at the end of the experimental session.

Procedure

Subjects were randomly assigned to conditions in a 2 (subject gender) x 2 (tutor voice: male/female) x 2 (evaluator voice: male/female) x 2 (topic: computers, love-and-relationships) mixed design. The first three factors were between-subjects factors; the fourth factor (topic) was a within-subject factor. The overall experimental procedure simulated a familiar academic situation: preparing for, taking, and receiving an evaluation for a test.

Upon arrival, the subject was told that he or she would use computers for four distinct tasks: a practice session, a tutoring session, a testing session, and an evaluation session.

<u>Practice Session</u>. In the practice session, the subject was trained in how to use the computer. The subject was also told that the computers used during the tutoring and evaluation sessions would use voice output, and the voice output was demonstrated.

<u>Tutoring Session</u>. During the tutoring session, the "tutor" computer orally presented the subject with ten facts on each of two topics: computers (e.g., "The more wire a computer has, the more slowly it runs.") and love-and-relationships (e.g., "More flowers are ordered for Mother's Day than for any other holiday."). The topics were chosen to be stereotypically male and stereotypically female, respectively. After receiving each fact, the subject indicated on a three-point scale ("very familiar/somewhat familiar/not at all familiar") how much he or she knew about that fact. This information was, in reality,

irrelevant. Although subjects were told that the computer chose each fact based on their familiarity with the previous facts, in actuality, each subject received the same 20 facts.

Testing Session. Following the tutoring session, the subject was directed by the tutoring session voice to go to another computer, the “tester” computer, to begin the testing session. During this session, the tester computer administered a 12-item, five-alternative multiple choice test without voice accompaniment (i.e., text only). The instructions indicated that a total of 12 questions would be randomly chosen from a set of 2,500 possible questions. In fact, all subjects received the same 12 questions (e.g., "What percentage of married women wear wedding rings?"). Each question had an "objectively" correct answer.

Evaluation Session. Upon completing the testing session, the subject was directed by the testing computer to go to a third computer, the “evaluator” computer, to begin the evaluation session. There, the evaluator computer orally informed the subject that he or she had answered 6 of the 12 questions correctly (three out of six for both topics). The computer then reviewed each question separately. For each item, the computer indicated whether the respondent gave a correct answer and then evaluated the performance of the tutor computer in preparing the subject. The overall evaluation of the performance of the tutor computer was generally positive (e.g., “Your answer to this question was correct. The tutor computer chose useful facts for answering this question. Therefore, the tutor computer performed well.”).

In sum, not only was the evaluator computer given the role of rating the performance of the subject, but it was also given the role of rating the performance of the tutor computer. The stance of the evaluator computer was thus distinctly dominant (Strong, Hills, Kilmartin, DeVries, Lanier, Nelson, Strickland, & Meyer, 1988).

The evaluator computer then asked the subject to complete a pencil-and-paper questionnaire next to the computer, and to notify the experimenter when the subject was finished. Upon completing the questionnaire, the subject was debriefed, thanked, and asked not to discuss the experiment with anyone else.

The entire experiment, including all four sessions and the final questionnaire, lasted approximately 40 minutes.

Manipulation

Male and female subjects were randomly assigned to tutor and evaluator computers delivering either male or female voice output. This created four possible tutor-evaluator combinations: male-male, male-female, female-female, female-male. The tutor and evaluator computers used prerecorded CD-quality (sampling rate of 44.1 Khz) human voices to deliver standardized scripts for the tutoring and evaluation sessions. Four different voices (two male, two female) were used, so that no subject heard the same voice in both tutor and evaluator roles. In addition, all four voices were pretested to guard against the possibility of significant differences in paralinguistic characteristics. Pretests indicated that the four voices were not perceived differently with respect to intonation, pacing, or other paralinguistic cues.

Other visual characteristics of the computer interface were identical across conditions and machines. This ensured that subjects would perceive all three computers as being programmed by the same person (which, in fact, they were), thereby ensuring that individuals did not associate the voice with the programmer in a form of para-social interaction (Horton & Wohl, 1956; Nass & Sundar, 1996).

Measures

The paper-and-pencil questionnaire consisted of two sets of questions.

The first set of questions asked subjects for their assessment of the tutoring computer’s performance during the tutoring session (e.g., “How informative was the tutor computer?”). The second set of questions asked subjects for their assessment of the evaluator computer during the evaluation session (e.g., “How competent was the evaluator computer?”). All of the items were measured on 10-point Likert scales.

Based on factor analysis, three indices were created.

Friendliness was an index comprised of four items: affectionate, likable, sympathetic, and warm. The reliability for this index was high for the tutoring session (Cronbach's alpha = .82) and satisfactory for the evaluation session (Cronbach's alpha = .67).

Competence was an index comprised of four items from the tutoring session: competent, informative, knowledgeable, and the question, "How much did the computer improve your final score?" (Cronbach's alpha = .78).

Informative was an index comprised of three items: helpful, sophisticated, and the question, "How well did the tutor computer choose facts about the topic of xxx?" (Cronbach's alpha for computer topic = .60; Cronbach's alpha for love-and-relationships topic = .68).

Results

Stereotype 1: Evaluation from males is more valid than evaluation from females.

This first stereotype led to the prediction that positive evaluation from a male-voiced computer would have a greater influence on subjects than positive evaluation from a female-voiced computer. This prediction was, in fact, confirmed: Subjects in the male-voiced evaluator condition rated the tutor computer more positively with respect to friendliness and competence than subjects in the female-voiced evaluator condition (See Table 1).

TABLE 1

Means for Assessment of Tutor and Evaluator as a Function of Gender of Evaluator Voice

	Evaluator Voice		
	Male	**Female**	***F***
Tutor Computer:			
Friendliness	19.65	14.15	5.74*
Competence	28.41	24.26	5.10*
Evaluator Computer:			
Friendliness:	16.25	11.55	6.03*

*p<.05.

To conduct this analysis, a full-factorial analysis of variance was performed, using the three between-subjects factors: subject gender, tutor voice, and evaluator voice. With respect to friendliness and competence, there was a significant main effect for evaluator voice, such that when the evaluator computer had a male voice, the tutoring session was regarded as significantly more friendly, *F* (1, 32) = 5.74, *p* < .05, and more competent, *F* (1, 32) = 5.10, *p* < .05, compared to when the evaluator computer had a female voice (See Tables 2 and 3 for complete ANOVA tables). Consistent with the literature, there were no significant differences between male and female subjects. There were no other main effects, and no significant interactions with respect to these measures.

TABLE 2

Analysis of Variance Table: Tutor Computer Friendliness

Source of Variation	**Sum of Squares**	**df**	**Mean Square**	***F***
Main Effects				
Subject Gender	62.38	1	62.38	1.18
Tutor Voice	.10	1	.10	.00
Evaluator Voice	302.23	1	302.23	5.74*
2-Way Interactions				
Subject Gender x Tutor Voice	19.53	1	19.53	.37
Subject Gender x Evaluator Voice	1.62	1	1.62	.03
Tutor Voice x Evaluator Voice	129.42	1	129.42	2.46
3-Way Interactions				
Subj. Gender x Tutor Voice x Eval. Voice	4.87	1	4.87	.09

**p*<.05.

TABLE 3

Analysis of Variance Table: Tutor Computer Competence

Source of Variation	**Sum of Squares**	**df**	**Mean Square**	***F***
Main Effects				
Subject Gender	33.03	1	33.03	.98
Tutor Voice	82.23	1	82.23	2.43
Evaluator Voice	172.85	1	172.85	5.10*
2-Way Interactions				
Subject Gender x Tutor Voice	.33	1	.33	.01
Subject Gender x Evaluator Voice	12.27	1	12.27	.36
Tutor Voice x Evaluator Voice	15.44	1	15.44	.46
3-Way Interactions				
Subj. Gender x Tutor Voice x Eval. Voice	62.13	1	62.13	1.83

*p<.05.

Stereotype 2: Dominance in females is unbecoming.

This second stereotype led to the prediction that a female-voiced computer in a dominant role would be evaluated more negatively than a male-voiced computer in the same role.

A full-factorial analysis was performed to test this prediction, using the three between-subjects factors. Results supported the prediction: Subjects in the female-voiced evaluator condition rated the evaluator computer as being significantly less friendly, F (1, 32) = 6.03, p < .05, than subjects in the male-voiced evaluator condition (See Table 1). Again, there were no significant differences between male and female subjects. There were no other main effects, and no significant interactions with respect to this measure. (See Table 4 for complete ANOVA table.)

TABLE 4

Analysis of Variance Table: Evaluator Computer Friendliness

Source of Variation	**Sum of Squares**	**df**	**Mean Square**	***F***
Main Effects				
Subject Gender	16.90	1	16.90	.46
Tutor Voice	4.90	1	4.90	.13
Evaluator Voice	220.90	1	220.90	6.03*
2-Way Interactions				
Subject Gender x Tutor Voice	.40	1	.40	.01
Subject Gender x Evaluator Voice	10.00	1	10.00	.27
Tutor Voice x Evaluator Voice	32.40	1	32.40	.89
3-Way Interactions				
Subj. Gender x Tutor Voice x Eval. Voice	.10	1	.10	.00

*p<.05.

Stereotype 3: Women know more about “feminine” topics, whereas men know more about “masculine” topics.

This stereotype led to the prediction that a female-voiced tutor computer would be perceived as more informative with respect to facts relating to love-and-relationships, whereas a male-voiced tutor computer would be perceived as more informative with respect to facts relating to computers.

To test this prediction, a four-factor, full-factorial analysis of variance was performed, using the three between-subjects factors (subject gender, tutor voice, and evaluator voice), as well as the within-subject factor (topic). Consistent with the prediction, there was a significant two-way interaction between topic and gender of tutor voice, such that the male-voiced tutor computer was perceived as more informative about computers compared to the female-voiced tutor computer, whereas the female-voiced tutor computer was perceived as more informative about love-and-relationships compared to the male-voiced tutor computer, *F*(1, 32) = 7.85, *p* < .01 (See Table 5.)

TABLE 5

Means for Tutor Informativeness as a Function of Topic and Gender of Tutor Voice.

Tutor Voice:	**Female**	**Male**	***F* (interaction)**
Topic:			
Computers:	1.22	1.26	7.85**
Love-and-Relations:	1.22	1.12	

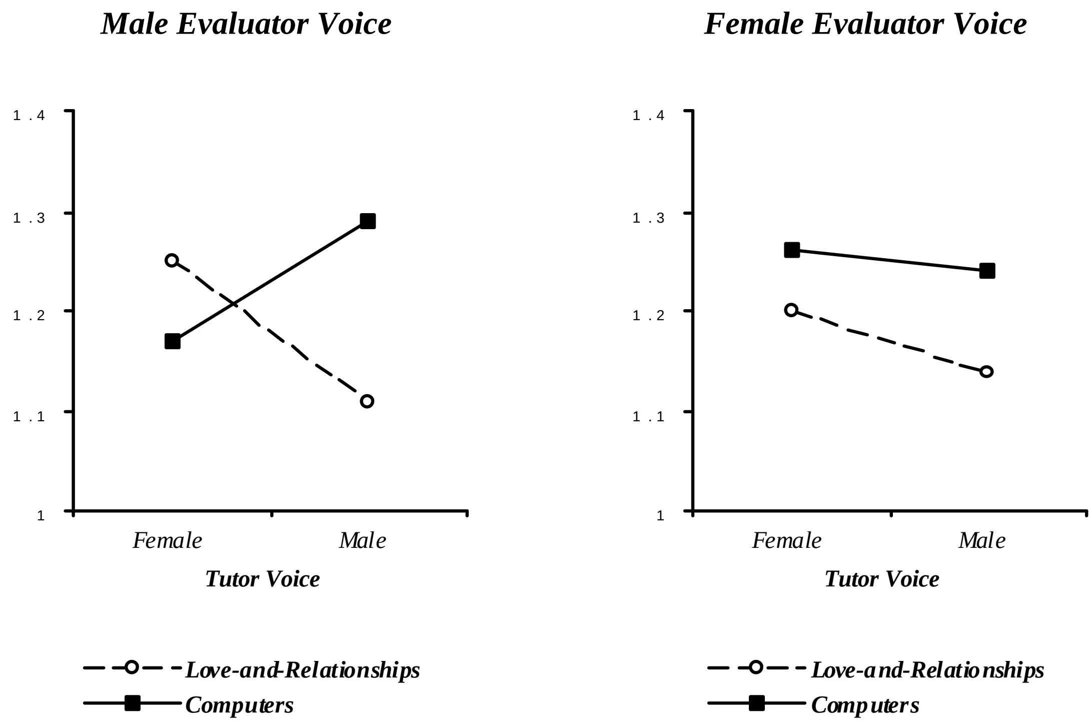
***p*<.01.

This significant two-way interaction, however, was actually an artifact of a significant three-way interaction between topic, gender of tutor voice, and gender of evaluator voice, *F*(1, 32) = 4.25, *p* < .05 (See Figure 1). For subjects in the male-voiced evaluator condition, the male-voiced tutor was perceived as being a better teacher of computers than love-and-relationships, whereas the female-voiced tutor was perceived as being a better teacher of love-and-relationships than computers, *F*(1, 16) = 11.14, *p* < .01. Thus, for this condition, Stereotype 3 was supported. For subjects in the female-voiced evaluator condition, however, both male- and female-voiced tutors were perceived as being better teachers of computers than love-and-relationships, *F*(1, 16) = .29. In other words, for this condition, Stereotype 3 was not supported.

There was also a significant main effect for topic (*F* (1, 32) = 6.79, *p* < .05), such that the tutor computer was perceived as being more informative about computers than love-and-relationships; this was likely an artifact of the relative quality of the facts in the two topics. There were no other main effects, and no other interaction effects.

FIGURE 1

Means for Tutor Informativeness as a Function of Topic, Gender of Tutor Voice, and Gender of Evaluator Voice.



Finally, in post-experimental debriefs, subjects were thoroughly questioned about the plausibility of the experimental procedure. Had subjects believed the evaluator computer's evaluation of the tutor computer? Had subjects believed that the tutor computer had tailored the choice of question to the subject's rating of familiarity with the previous item? And had subjects believed the evaluation given by the evaluator computer? With respect to all of these queries, subjects reported no problems in accepting the experimental cover story.

Furthermore, all subjects denied harboring stereotypes or being influenced by the gender of the computer voices. When asked, none of them said they thought the voices represented the programmers of the computers; in fact, most subjects thought the three computers used in the experiment were programmed by the same person (which they were), and that the person was male (which is ironic, since they said they did not harbor stereotypes).

Discussion

This study presents an experimental demonstration of the power of gender stereotypes. Using a manipulation that effectively removed all other gender cues, including physical appearance and nonverbal communication, from the interaction, this study provides evidence that vocal cues embedded in a machine are sufficient to evoke sex-based stereotypic responses. Specifically, subjects used the stereotypes "evaluation from males is more valid than evaluation from females," "dominance is more desirable in men than women," and (for male-voiced evaluators) "women know more about subjects that are typically regarded as 'feminine,' whereas men know more about subjects that are typically regarded as 'masculine.'" These stereotypic responses were generated in the absence of another *human* in the interaction; indeed, all subjects were explicitly informed that the interaction was with a computer. And again, in post-experimental debriefs, all subjects not only denied applying gender stereotypes in their evaluations of the computers, but agreed that to do so with respect to computers would be clearly inappropriate.

It thus appears that the tendency to gender stereotype is deeply ingrained in human psychology, extending even to inanimate machines. This finding has significant social implications. The key implication is that when voice technology is embedded in a machine interface, voice selection is highly consequential. Indeed, by choosing (or *casting*) a particular voice, a designer or engineer may trigger in the user's mind a whole set of expectations associated with that voice's gender. For designers and engineers to assume that any voice is neutral, is a mistake; a male voice brings with it a large set of expectations and responses based on stereotypes about males, whereas a female voice brings with it a large set of expectations and responses based on stereotypes about females.

This study also suggests that *any* suggestion of gender in a given technology, however minor, may trigger stereotypic responses. For example, visual representations of computer characters and even their language style may elicit gender-stereotypic responses.

Finally, this study suggests that voice technology may evoke stereotypic responses along dimensions other than gender. People may consciously or unconsciously assign an age, a social class, and a geographic location to a disembodied voice. This, in turn, will create expectations about how the technology will, or should, behave.

The decision to imbue a given technology with voice can therefore involve difficult choices. Designing technology that conforms to the user's gender stereotypes may be the simplest way to meet his or her expectations about the technology. On the other hand, technology that challenges these stereotypes may serve to change, in the long run, the deeply ingrained biases that underlie the findings in the present study.

References

Briton, N. J. , & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. Sex Roles, 32, 79-90.

Costrich, N., Feinstein, J., Kidder, L., Maracek, J., & Pascale, L. (1975). When stereotypes hurt: Three studies of penalties in sex-role reversals. Journal of Experimental Social Psychology, 11, 520-530.

Deutsch, C. J., & Gilbert, L. A. (1976). Sex role stereotypes: Effect on perceptions of self and others and on personal adjustment. Journal of Counseling Psychology, 23, 373-379.

Eagly A. H., & Wood, W. (1982). Inferred sex differences in status as a determinant of gender stereotypes about social influence. Journal of Personality and Social Psychology, 43, 915-928.

Goldberg, P. A. (1968). Are women prejudiced against women? Transaction, April, 28-30.

Heilman, M. E. (1979). High school students' occupational interest as a function of projected sex ratios in male-dominated occupations. Journal of Applied Psychology, 64, 275-279.

Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observation on intimacy at a distance. Psychiatry, *19*, 215-229.

McKee, J. P., & Sheriffs, A. C. (1959). Men's and women's beliefs, ideals, and self-concepts. American Journal of Sociology, 64, 356-363.

Nass, C. & Sundar, S. S. (1996). Is human-computer interaction social or parasocial? Unpublished manuscript.

Pleck, J. H. (1978). Males' traditional attitudes toward women: Conceptual issues in research. In J. A. Sherman & F. L. Denmark (Eds.), The psychology of women: Future directions in research. New York: Psychological Dimensions.

Robinson, J., & McArthur, L. Z. (1982). Impact of salient vocal qualities on causal attribution for a speaker's behavior. Journal of Personality and Social Psychology, 43, 236-247.

Romer, N., & Cherry, D. (1980). Ethnic and social class differences in children's sex-role concepts. Sex Roles, 6, 246-263.

Rosenkrantz, P. S., Vogel, S. R., Bee, H., Broverman, I. K., & Broverman, D. M. (1968). Sex role stereotypes and self concepts in college students. Journal of Consulting and Clinical Psychology, 32, 287-295.

Ruble, T. L. (1983). Sex stereotypes: Issues of change in the 1970's. Sex Roles, 9, 397-402.

Spence, J. T., Helmreich, R., & Stapp, J. (1974). The Personal Attributes Questionnaire: A measure of sex-role stereotypes and masculinity-femininity. JSAS Catalog of Selected Documents in Psychology, 4, 43.

Strong, S. R., Hills, H. I., Kilmartin, C. T., DeVries, H., Lanier, K., Nelson, B. N., Strickland, D., & Meyer, C. W. (1988). Journal of Personality and Social Psychology, 54, 798-810.