Influences of Evaluative Contexts in Human-Robot Interaction and Relationships with Personal Traits*

Tatsuya Nomura, *Member, IEEE*, and Takayuki Kanda, *Member, IEEE*

***Abstract*—To investigate the effects of robots on human self-assessment under the more general contexts of learning, this research focused on the evaluative feedback by robots. A psychological experiment was conducted by using a human-sized humanoid robot. The results found that the evaluative context tended to increase the subjects' negative responses regarding disclosure of their skills to the robot and the robot evaluating the skills, and the internal locus of control and the fear of negative evaluation positively influenced the subjects' responses to disclosing their skills to the robot. This paper discusses the implications on the use of robots in self-monitoring for healthcare and education.**

I. Introduction

Information communication technologies (ICT) are expected to be applied in a variety of fields in daily life. In particular, those fields using artificial agents have the potential to be used for human self-assessment in healthcare and education. For example, Bickmore and Picard [1] validated the effects of an exercise adviser agent for health behavioral changes of human clients. Brave et al. [2] investigated the psychological effects of sympathy from an artificial agent on users and suggested its usefulness in applications such as teacher agents.

Moreover, robots with physical embodiment can have an impact on human self-assessment due to their physical presence. Kidd and Breazeal [3] conducted experiments to compare the use of software agents and robots, and found that robots were more suitable than agents for specific tasks such as pointing at objects in real surroundings. Wainer et al. [4] suggested some effects of a robot on autistic children's learning of collaborative behaviors with others.

The existing studies focused on behavioral expressions from agents and robots for interacting with humans. Related to human interactions, evaluation by others is considered to be one of the important factors in self-assessment based on critical thinking. Social psychological studies have found that people with higher social anxiety tend to fear a negative evaluation from others, and these people tend to avoid communication with others to lessen the amount of self-disclosure [5]. Moreover, Pertaub et al. [6] found that even in virtual reality settings, human speakers were affected by the attitudes of audience agents. These results imply that evaluative contexts with robots may influence human self-assessment. Therefore, it is important to explore human personal traits related to this influence.

In this research, a psychological experiment was conducted to investigate the impact of robot evaluations on humans under a learning context by using the following: a human-sized humanoid robot, psychological scales to measure human personal traits, and an open-ended questionnaire for human responses to the robot and the context. This paper reports the experimental results and then discusses their implications in self-monitoring for healthcare and education.

*The research was supported in part by the Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research No. 21118006.

Tatsuya Nomura is with Department of Media Informatics, Ryukoku University, Otsu, Shiga 520-2194, Japan (corresponding author to provide phone: +81-77-544-7136; fax: +81-77-544-7150; e-mail: nomura@rins.ryukoku.ac.jp).

Takayuki Kanda is with ATR Intelligent Robotics and Communication Laboratories, Keihannna Science City, Kyoto 619-0288, Japan (e-mail: kanda@atr.jp)..

II. Method

The experiment was conducted from February 2011 to January 2012. The subjects were 155 Japanese university students (male: 82, female: 75) with mean age 20.5 (*SD* = 1.8).

A. The Robot Used in the Experiment

In the experiment, a human-sized humanoid robot was used, as shown in Figure 1. The robot, named "Robovie-R2," was developed by ATR Intelligent Robotics and Communication Laboratories. This robot stands 110 cm tall, weighs approximately 57 kg, and has 3 degrees of freedom (DOFs) on its head and 4 DOFs on each arm. Two CCD cameras and a speaker are located on the head, and this robot has speech capability based on audio data through a built-in PC. Although the robot can move around by using the two wheels at the bottom of the body, this function was not used in the experiment.

The experiment was conducted under a Wizard-of-Oz setup. The equipment consisted of a PC for operation of the robot (CPU: Core2Duo 2.8 GHz, memory: 4 GB), another PC for operating a printer, and a digital video camera and microphone for recording scenes of the experimental sessions. The PC and the robot were connected via a wireless LAN.

B. Evaluation by the Robot under a Learning Context

The experiment assumed a situation in which humans self-monitored their current skills under the context of learning English conversation. By evaluating their conversation skills, the humanoid robot assisted the self-monitoring of subjects who had low self-confidence regarding English.

The behaviors of the robot were as follows:

Figure 1. Robovie-R2, used in the experiment

- The robot uttered some questions asking each subject to answer about her/his current skill of English conversation (e.g., “If you are asked for route directions by a foreign person speaking English, can you answer well?”).
- During the utterances, the robot moved its head and arms as if it was actively following the conversation with the subject.
- At the end of the interaction, the subject was provided with a document in which the contents of the interaction were summarized.

The above questions included the abovementioned one asking subjects to answer about their self-confidence regarding their English conversation skills. It was confirmed that no subjects showed high self-confidence in their English skills.

The experiment had the following between-subjects design. The differences between the conditions focused only on the instruction and feedback from the robot:

- **Evaluation Condition:** Subjects were instructed that the robot diagnosed their English skills. Negative evaluation results of the subject’s current English skill (by presenting some scores) were added to the document provided to each subject.
- **Non-Evaluation Condition:** Subjects were instructed that the robot summarized their utterances about their English skills. These subjects were not shown any evaluation results.

C. Measures

1) FNE and LOC:

As human personal traits related to evaluative contexts, the experiment focused on the fear of negative evaluation (FNE) [6] and the locus of control (LOC) [7].

As mentioned in the introduction, people with a strong fear of negative evaluation from others can possibly be influenced by the robot's evaluation. The experiment adopted the Japanese version of “Fear of Negative Evaluation Scale” [8]. This Lickert-type scale consists of thirty items (e.g., “I am often afraid that others find my faults.”), and the degree of the fear is calculated as the sum of all the item scores (Yes = 1, No = 0), including the reverse of some items (maximum score: 30, minimum score: 0).

The locus of control is a person’s degree of belief that she/he can control her/his own destiny. People with an external locus of control tend to think that their own successes are controlled by others and environments, and so they have a greater tendency to be influenced by social stimuli in their environment and to modify their behavior in accordance with the responses and evaluations of others. People with an internal locus of control tend to think that they control their successes. Rickenberg and Reeves [9] found that the LOC affected users’ evaluations of web sites when an animated

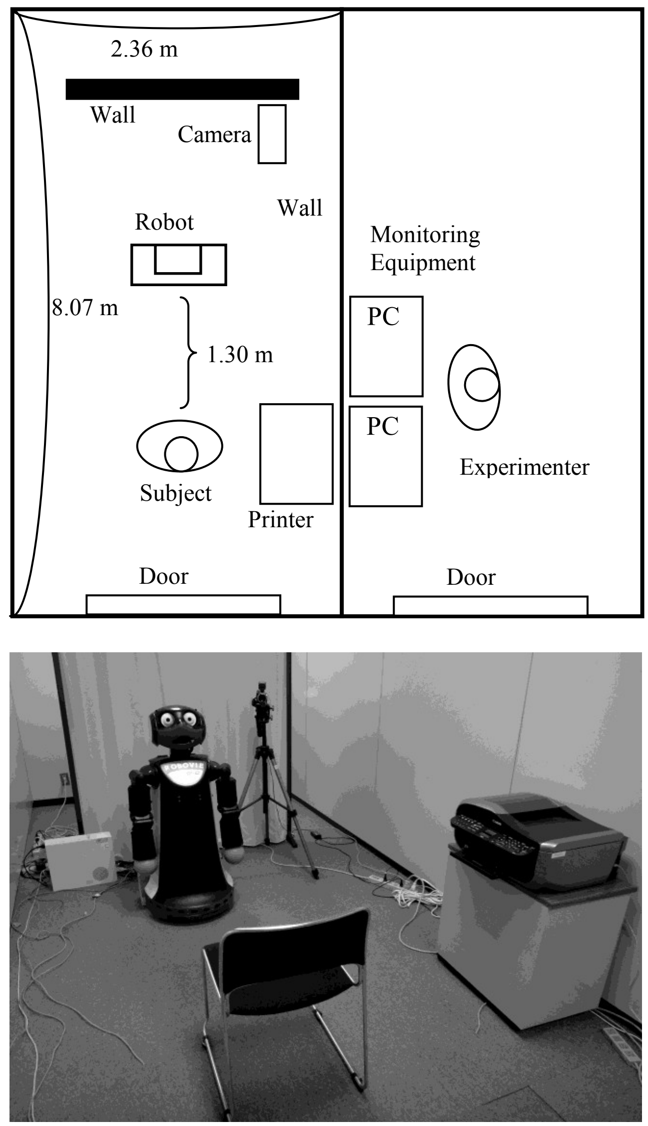

Figure 2. Illustration of the experimental room (top), and a photo of an experimental session (bottom).

character showed behaviors of monitoring the users.

To investigate the relationships between this personal trait and evaluation by the robot, the experiment adopted the Japanese version of the LOC scale [10]. This scale consists of eighteen items (e.g., "Do you think that you are deciding your own life by yourself?"), and the degree of locus of control was calculated as the sum of all the item scores (four levels per answer: 1. I think so – 4. I do not think so), including the reverse of some items. A higher scale score means a stronger internal locus of control (maximum score: 72, minimum score: 18).

These two psychological scales were conducted at the pre-stage of each experimental session.

2) Open-ended questionnaire:
To explore the subjects' opinions about the evaluation by the robot, they were asked to answer the following open-ended items after the experimental session:

- Q1: What do you think about disclosing your skill of English conversation to the robot?
- Q2: What do you think about the robot evaluating/deciding your skills of English conversation?
- Q3: What do you feel about the future spread of robots evaluating/deciding people's skills, as in the experiment, in society?

Subjects were asked to answer freely regardless of their positive or negative opinions. The responses were classified based on the contents after all the sessions.

D. Procedure

Each session in the experiment was conducted according to the following procedure:

- Before entering the experimental room shown in Figure 2, each subject responded to the pre-session questionnaire, which included demographic items, the FNE, and the LOC. Then, she/he was instructed to interact with the robot about her/his English communication. At this stage, a video explaining the robot behaviors was shown to the subject.
- The subject entered the room alone and sat down on the chair in front of the robot.
- The robot uttered three questions about the subject's skill of English communication, and she/he answered the questions. These answers were monitored by the experimenter.
- After the interaction, the experimenter manually summarized the subject's answers (with some negative evaluation results on the evaluation condition) and printed out a summary on the printer next to the robot. The involvement of the human experimenter was concealed from the subject.
- After the session, the subject exited the experimental room and responded to a post-session questionnaire that included the open-ended questions.

III. Results

Among a total of 155 subjects, 76 subjects (male: 41, female: 35) were assigned to the evaluation condition, and 79 subjects (male: 41, female: 38) were assigned to the non-evaluation condition.

A. Coding of Open-ended Responses

To quantitatively analyze the open-ended responses, the responses were manually classified into several categories based on the contents. The classifications were determined by two people.

First, coding categories were created for each question. Then, the two people independently categorized 135 responses (87%). The κ-coefficients showing the degrees of equality between the categorization results by the two people were calculated to validate the reliability of coding categories. As a result, the κ-coefficients for items Q1, Q2, and Q3 were .878, .916, and .905, respectively. These values showed sufficient reliability of coding categories. Furthermore, the two people discussed the contents of the responses and

TABLE 1. Categories Extracted from Open-Ended Questions and Examples of Answers

Question	Category Label	Example
Q1. Disclosing the skill of English conversation to the robot	Positive	"I could talk about my own skill with the robot straightforwardly."
	Negative	"I had no reason to disclose my English skill to the robot."
	Both positive and negative	"Summarization by the robot led me to easily understand. But I thought that nothing that I had already known was clarified."
	Neutral	"I have no specific thought."
Q2. The robot evaluating/deciding the skill of English conversation	Positive	"I thought that the decision by the robot was exact and great."
	Negative	"I thought that it was hard for the robot to evaluate the English skill."
	Both positive and negative	"The attempt is interesting, but I don't think that humanoid appearances are needed."
	Neutral	"I have no specific thought."
Q3. The future spread of robots evaluating/deciding people's skills in society	Positive	"I think that evaluation by robots is better than that by humans if they can exactly evaluate people's skills."
	Negative	"I think that it weakens the relationship between people in society."
	Both positive and negative	"I think that they are effective in the case of easy assessment, but may have no practical use if they cannot detail the evaluation contents."
	Neutral	"I don't feel the reality because I think that it is in the far future."

categories until they reached a consensus about each classification. Finally, each response was classified into one of the following categories.

Table 1 shows these categories and examples of sentences classified into the categories. For all of the questions, the responses were commonly classified into sentences consisting of one of the following: positive opinions, negative opinions, both positive and negative opinions, and neutral opinions.

B. Relationships between Response Categories and Robot Evaluation

Table 2 shows the frequency in the categories of responses about disclosing the skills of English conversation to the robot. Approximately half of the respondents were positive. The frequency of these categories was statistically analyzed by using the condition of robot evaluation/non-evaluation. The χ^2-test showed differences between the evaluation conditions in the category frequency ($\chi^2(3) = 9.144, p < .05$). The residual analysis with $\alpha = .05$ revealed the following differences: in the evaluation condition the frequency of positive responses (36.8%) was less than average (47.1%), and in the non-evaluation condition the frequency (57.0%) was more than average at a statistically significant level. Moreover, in the evaluation condition the frequency of negative responses (32.9%) was more than average (23.9%), and in the non-evaluation condition the frequency (15.2%) was less than average at a statistically significant level.

Table 3 shows the frequency in the categories of responses about the robot evaluating/deciding the skill of English conversation. More than 60% of the respondents were positive. The χ^2-test showed differences between the evaluation conditions in category frequency ($\chi^2(3) = 10.071$, $p < .05$). The residual analysis with $\alpha = .05$ revealed the following differences: in the evaluation condition the frequency of positive responses (56.6%) was less than average (65.8%), and in the non-evaluation condition the frequency (74.7%) was more than average at a statistically significant level. Moreover, in the evaluation condition the frequency of negative responses (26.3%) was more than average (16.8%), and in the non-evaluation condition the frequency (7.6%) was less than average at a statistically significant level.

Table 4 shows the frequency in the categories of responses about the future spread of robots evaluating/deciding people's skills in society. Approximately half of the respondents were positive. The χ^2-test showed no difference between the evaluation conditions in the category frequency ($\chi^2(3) = 1.175$, *n.s.*).

TABLE 2. FREQUENCIES OF RESPONSE CATEGORIES FOR Q1 (DISCLOSING THE SKILL OF ENGLISH CONVERSATION TO THE ROBOT)

	Evaluation		Non-evaluation		Total	
	N	%	*N*	%	*N*	%
Pos	*28*	*36.8%*	**45**	**57.0%**	73	47.1%
Neg	**25**	**32.9%**	*12*	*15.2%*	37	23.9%
PN	10	13.2%	7	8.9%	17	11.0%
Neut	13	17.1%	15	19.0%	28	18.1%
Total	76	100.0%	79	100.0%	155	100.0%

(Pos: Positive, Neg: Negative,
PN: Both positive and negative, Neut: Neutral,
Bold: larger than the expected frequency,
Italic: smaller than the expected frequency)

TABLE 3. FREQUENCIES OF RESPONSE CATEGORIES FOR Q2 (THE ROBOT EVALUATING/DECIDING SKILL OF ENGLISH CONVERSATION)

	Evaluation		Non-evaluation		Total	
	N	%	*N*	%	*N*	%
Pos	*43*	*56.6%*	**59**	**74.7%**	102	65.8%
Neg	**20**	**26.3%**	*6*	*7.6%*	26	16.8%
PN	7	9.2%	7	8.9%	14	9.0%
Neut	6	7.9%	7	8.9%	13	8.4%
Total	76	100.0%	79	100.0%	155	100.0%

TABLE 4. FREQUENCIES OF RESPONSE CATEGORIES FOR Q3 (FUTURE SPREAD OF ROBOTS EVALUATING/DECIDING PEOPLE'S SKILLS IN SOCIETY)

	Evaluation		Non-evaluation		Total	
	N	%	*N*	%	*N*	%
Pos	37	48.7%	38	48.1%	75	48.4%
Neg	11	14.5%	9	11.4%	20	12.9%
PN	25	32.9%	26	32.9%	51	32.9%
Neut	3	3.9%	6	7.6%	9	5.8%
Total	76	100.0%	79	100.0%	155	100.0%

No gender differences were found for these frequencies: ($\chi^2(3) = 1.035$, *n.s.*, for Q1; $\chi^2(3) = .182$, *n.s.*, for Q2; $\chi^2(3) = 1.226$, *n.s.*, for Q3).

C. Factors Influencing Positive Opinions

To investigate relationships between the open-ended responses and the subjects' personal traits, a correlation analysis and a logistic regression analysis were conducted.

TABLE 5. CORRELATION COEFFICIENTS BETWEEN THE SIMPLIFIED OPEN-ENDED RESPONSE CATEGORIES, FNE, AND LOC

		LOC	Q1	Q2	Q3
FNE	Complete Samples (*N* = 155)	-.218**	.161*	.009	-.038
	Evaluation (*N* = 76)	-.175	.153	-.085	-.028
	Non-evaluation (*N* = 79)	-.268*	.113	.046	-.046
LOC	Complete Samples (*N* = 155)	--	.249**	.093	.112
	Evaluation (*N* = 76)	--	.439**	.115	.157
	Non-evaluation (*N* = 79)	--	.083	.070	.069

Q1. Disclosing the skill of English conversation to the robot, Q2. The robot evaluating/deciding skill of English conversation,
Q3. Future spread of robots evaluating/deciding people's skills in society,
1: positive, 0: others
(*$p < .05$, **$p < .01$)

TABLE 6. RESULTS OF LOGISTIC REGRESSION ANALYSES FOR POSITIVITY OF OPEN-ENDED RESPONSES IN COMPLETE SAMPLES

	Q1. Disclosing the skill of English conversation to the robot		Q2. The robot evaluating/deciding skill of English conversation	
	Exp(B)	*Wald*	*Exp(B)*	*Wald*
FNE	1.069	6.046*		
LOC	1.121	12.324***		
Evaluation	.469	4.662*	.442	5.540*
-2 Log Likelihood	191.478		193.433	
Cox & Snell R^2	.137		.036	
Nagelkerke R^2	.183		.050	

(* $p < .05$, *** $p < .001$)

Cronbach's reliability coefficient α of the FNE and the LOC scales was .892 and .694, respectively. The mean score of FNE was 15.3, and the standard deviation was 6.9. No gender difference was found for the scale score (t = -1.625, *n.s.*, d = .275). The mean score of the LOC was 50.8, and the standard deviation was 5.9. Although the effect was small, a gender difference was found for the scale score (male: M = 51.9, SD = 5.8, female: M = 49.6, SD = 5.8; t = 2.358, $p < .05$, d = .381).

For the analyses, the open-ended response categories were simplified by re-coding responses as positive = 1 and others = 0. Then, the correlation coefficients between these response categories, the FNE, and the LOC were calculated independently for the "complete" set of all samples, the samples in the "evaluation" condition, and those in the "non-evaluation" condition. Table 5 shows these coefficients. Note that Pearson's correlation coefficients are equal to the point biserial correlation coefficients in this case. In the complete samples, statistically significant levels of positive correlations were found between Q1 and FNE, and between Q1 and LOC, although both these correlations were small. The correlation between Q1 and LOC was at a moderate level in the samples of the evaluation condition, although no correlation was found in those of the non-evaluation condition. A test of equality found that the correlation coefficient in the sample of the evaluation condition was higher than that of the non-evaluation condition at a statistically significant level (Z = 2.366, $p < .05$).

Logistic regression analyses were performed with the simplified open-ended response categories as dependent variables by using the backward elimination method. The independent variables were the scale scores of the FNE and the LOC, gender (male = 0, female = 1), and the condition of evaluation by the robot (evaluation condition = 1, non-evaluation condition = 0). Table 6 shows the models extracted for Q1 and Q2. No model was extracted for Q3. These models show the influences of the evaluation condition on the positivity of opinions about disclosing the skill of English conversation to the robot and the robot evaluating/deciding the skill of English conversation. Moreover, it was found that the FNE and the LOC influenced the positivity of opinions about disclosing the skill of English conversation to the robot.

The result of the correlation analysis suggests that the difference of the relationships between the LOC and the response positivity was dependent on the evaluation condition. Thus, logistic regression was re-performed for samples of the evaluation condition and those of the non-evaluation condition independently. As a result, a model was extracted only for Q1 in the samples of the evaluation condition, as shown in Table 7. The model was equal to that extracted in the complete sample. The good-of-fitness values of the model were higher than those in the complete samples.

IV. DISCUSSION

A. Findings

The analysis results found that the evaluative context in the human-robot interaction experiment tended to increase people's negative opinions about disclosing their skills to the robot and the robot evaluating/deciding the skills in comparison with opinions in the non-evaluative context. These results suggest that people do not prefer robotics applications providing evaluations in the current state of technologies.

In contrast, the opinions about the future spread of robots evaluating/deciding people's skills in society were not influenced by the context. The statistically estimated interval of the rate of positive opinions with 95% reliability was [40.3%, 56.5%], and so suggests that many people expect robotics applications to provide evaluations in the future, regardless of their experiences with the robot in the experiment.

Moreover, the evaluative context emphasized the relationship between human responses about disclosing their skills to the robot and the fear of negative evaluation and locus of control, although such a relationship was not found under the non-evaluative context. Only in the evaluative context, people having a stronger internal locus of control and a stronger fear of negative evaluation showed positive

TABLE 7. RESULTS OF LOGISTIC REGRESSION ANALYSES FOR POSITIVITY OF OPEN-ENDED RESPONSES IN SAMPLES OF EVALUATION CONDITION

	Q1. Disclosing the skill of English conversation to the robot	
	Exp(B)	*Wald*
FNE	1.097	4.792*
LOC	1.244	13.786***
-2 Log Likelihood	79.036	
Cox & Snell R^2	.241	
Nagelkerke R^2	.330	

(* $p < .05$, *** $p < .001$)

responses for disclosing their skills to the robot. This result suggests that designers of robotics applications with evaluation contexts should pay attention to the personal traits of positive users.

B. Implications

From the perspective of robotics applications in self-monitoring for healthcare and education, the above findings have some important implications.

Robotics applications with evaluative contexts may make people have negative impressions at the current technological level in comparison with applications with non-evaluative contexts. Nevertheless, applications with evaluative contexts are expected in society, and people having a strong internal locus of control, who believe that they can control their own destiny, may positively use this type of application.

An important experimental result is that the fear of negative evaluation has a positive effect on opinions about self-disclosure to the robot. Although the cause cannot be determined in this research, it is estimated that people with a fear of negative evaluation from others prefer an evaluation from robots over that from humans. This result may lead to an advantage of robotics applications in self-monitoring in learning contexts. In fact, Joinson [11] found that computer-mediated communication encouraged self-disclosure in comparison with face-to-face communication.

It is estimated that people with a fear of negative evaluation have a stronger external locus of control. In fact, a negative correlation was found between the LOC and the FNE in the experiment (Table 5). This result implies that two types of people, having traits contrary to each other, may positively use robotics applications with evaluative contexts. If so, robotics designers should also pay attention to the possibility that the system interface adapted for one type of user (e.g., with an internal locus of control) may not be suitable for another type of user (e.g., with an external locus of control).

C. Limitations

The difference between the evaluative and non-evaluative contexts basically lies on the experimental instruction, and the robot did not give any concrete assessment or advice on the English skills. Some subjects showed negative responses about this fact, and this methodological limitation may have influenced the subjects' responses.

Moreover, the current research used only open-ended responses as dependent variables. To investigate the relationships between personal traits and evaluative contexts, more psychological and behavioral indices should be measured, such as ease of use, intention to use, and utterances during the interaction.

Finally, the sampling of subjects in the experiment was limited to Japanese university students. Thus, influences of age and culture were not taken into account. These problems should be addressed in future research.

ACKNOWLEDGMENT

The authors deeply thank Dr. Michihiro Shimada and Ms. Sayaka Taniguchi in the ATR Intelligent Robotics and Communication Laboratories for their cooperation with the conduction of the experiment in this research.

REFERENCES

[1] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer interaction," *ACM Trans. Computer-Human Interaction*, vol.12, pp.293-327, 2005.

[2] S. Brave, C. Nass, and K. Hutchinson, "Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent," *Int. J. Human-Computer Studies*, vol.62, pp.161-178, 2005.

[3] C. Kidd and C. Breazeal, "Effect of a robot on user perceptions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2004, pp.3559-3564.

[4] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "Collaborating with Kaspar: Using an Autonomous Humanoid Robot to Foster Cooperative Dyadic Play among Children with Autism," in *Proc. Int. Conf. Humanoid Robots*, 2010.

[5] M. R. Leary, *Understanding social anxiety: Social, personality, and clinical perspectives*. Beverly Hills: Sage, 1983.

[6] J. B. Watson and R. Friend, "Measurement of social-evaluation anxiety," *J. Consulting and Clinical Psychology*, vol.33, pp.448-457, 1969.

[7] J. B. Rotter, "Generalized expectancies for internal versus external locus of control of reinforcement," *Psychological Monographs: General and Applied*, vol.80, no.1, pp.1-28, 1966.

[8] R. Ishikawa, K. Sasaki, and I. Hukui, "Standardization of Japanese Version of FNE and SADS," *Jap. J. Behavioral Therapy*, vol.18, no.1, pp.10-17, 1992.

[9] R. Rickenberg and B. Reeves, "The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of. User Interfaces," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2000, pp.49-56.

[10] M. Kanbara, K. Higuchi, and N. Shimizu, "New Locus of Control Scale: Its Reliability and Validity," *Jap. J. Educational Psychology*, vol.30, no.4, pp.302-307, 1982.

[11] A. N. Joinson, "Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity," *Euro. J. Social Psychology*, vol.31, pp.177-192, 2001.