# Estimation of subjective evaluation of HRI performance based on objective behaviors of human and robots*

Yoshiaki Mizuchi[1][0000-0002-2830-815X] and Tetsunari Inamura[1,2][0000-0002-0028-6438]

[1] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
{mizuchi, inamura}@nii.ac.jp
[2] SOKENDAI(The Graduate University for Advanced Studies), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

**Abstract.** The conventional approach to the evaluation of the performance of human-robot interaction (HRI) is subjective evaluation, such as the application of questionnaires. As such subjective evaluation is time-consuming, an alternative automatic evaluation method based on only objectively observable factors (i.e., human reaction behavior) is required for autonomous learning by robots and for scoring in robot competitions. To this end, we aim to investigate the extent to which subjective evaluation results can be approximated using objective factors. As a case study, we designed and carried out a VR-based robot-competition task in which the robot was required to generate comprehensible and unambiguous natural language expressions and gestures to guide inexpert users in everyday environments. In the competition, both event data and human behavioral data (i.e., interaction histories) were observed and stored. Additionally, to acquire subjective evaluation results, we asked third-parties to evaluate the HRI performance by reviewing the stored interaction histories. From the analysis of the relationship between objective factors and subjective evaluation results, we demonstrate that the subjective evaluation of HRI can indeed be reasonably approximated on the basis of objective factors.

**Keywords:** Human-robot interaction · Natural language generation · RoboCup@Home · Virtual reality.

## 1 Introduction

The evaluation of human-robot interaction (HRI) is important to improve the social skills of interactive robots. Moreover, there is a particular need for automated evaluation; for example, for autonomous learning of human-interaction

* This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was also supported by JST CREST Grant Number JPMJCR15E3, Japan.

skills, robots need to be able to evaluate their interactions, responding appropriately to both positive and negative results to improve their HRI policies without specific interventions by developers. Due to the fact that the requirement exists in the context of robot competitions such as RoboCup@Home, these provide an appropriate and effective context for investigating techniques and approaches for the improvement of HRI.

The evaluation of HRI performance is generally subjective, with the standard method being the use of questionnaires for the evaluation of social and cognitive factors. However, subjective evaluation suffers from two drawbacks: first, the acquisition of answers from test subjects is time-consuming, and second, the number of samples is limited to the number of available human respondents. Accordingly, the conventional questionnaire-based approach to subjective evaluation is impractical for fair and efficient evaluation of HRI performance in contexts such as robot competitions.

One of the solutions to these problems is an alternative method that approximates the subjective evaluation results from only observable objective factors, namely, human reaction behaviors. In order to achieve this, we need to analyze the relationship between the subjective evaluation results and the objective factors. Therefore, it is necessary to observe and store data on all the events and human behaviors in question, in addition to acquiring subjective evaluation results. As experiments of this nature are time-consuming and costly, we chose to utilize immersive virtual reality (VR) techniques as a more time- and cost-effective solution to carry out HRI experiments, and thereby to acquire the requisite data. We have already proposed a VR-based software platform, SIGVerse [7], which enables human users to log in and make use of an avatar to interact with virtual robots and environments through immersive VR interfaces. A basic concept has been also proposed for the evaluation of the performance of the HRI in a VR-based robot competition task [3]; however, no evaluation method that approximates subjective evaluation results has been proposed or discussed to date.

The aim of this study is to investigate how to approximate subjective evaluation results of HRI performance on the basis of objective factors. In this paper, we present a case study of a robot competition in which the virtual robot had to guide non-expert human users to complete a certain task by means of verbal/nonverbal communication. Data on human reaction behaviors occurring in response to the instructions given by different robots were recorded and stored. Additionally, to acquire subjective evaluation results, we asked third-parties to evaluate the HRI performance by reviewing the stored interaction histories. From the the analysis of the relationship between objective factors and subjective evaluation results, we demonstrate that the subjective evaluation of HRI could be reasonably approximated using objective factors.

## 2 Related Works

With respect to the evaluation of the social and cognitive effects of robots, various measurement scales have been developed. In research fields relating to human-agent interaction and social robotics, developing psychological scales is a common focus area (e.g.,[4, 8]). Although Bartneck et al. [1] developed the 'godspeed questionnaire' as a standardized measurement tool for the measurement of the psychological effects of HRI, their study and similar works in the field have focused on psychological effects, using only questionnaire-based subjective evaluations.

In cases of targeting objectively measurable factors such as comfortable distances [9], the effects of HRI can be evaluated directly. However, the effects of most of HRI functions, particularly natural language instructions and gestures generated by robots, are not measurable directly. With respect to measurable factors to verify the effectiveness of HRI functions, the focus has been on metrics such as required time and success rates in a specific task (e.g., [6]). Although using those metrics is a reasonable approach to compare the performance of HRI, it is not clarified what metrics are appropriate for the evaluation and how to choose appropriate metrics.

With respect to experiments to evaluate the effects of HRI, a few research groups have carried out field studies in real-world environments (e.g., [2]). Although such field studies are important for the development of social robots, in terms of the efficient evaluation, especially for robot competitions, they offer somewhat infeasible solutions for evaluating parallel sessions, easily controlling experimental conditions, or reproducing interaction histories. The utilization of VR techniques offers a feasible solution to these problems. The Generating Instructions in Virtual Environments (GIVE) challenge [10] was proposed to evaluate natural language generation (NLG) systems which guide human users to perform a task in a virtual environment, but HRI is not a focus area in this challenge. In the absence of any similar challenges focusing on HRI, no forum exists in which the embodiment of both robots and human instruction-followers (i.e., gestures by a virtual robot and physical actions of an avatar to interact with objects) are not targeted.

Kanda et al.[5] focused on the analysis of the relationship between objective results (the body movements of subjects) and subjective impressions. To determine an evaluation method for HRI performance, they attempted to estimate the subjective evaluation results from the body movements with multiple linear regression analysis. Ideally, the scoring methods and rules for robot competitions should be developed through this sort of analysis. However, there is no such work on direct evaluation of HRI performance involving generated natural language sentences/utterances and gestures generated by robots, nor is there an appropriate platform available. Accordingly, as the primary step towards this ultimate goal, we focus on a case study of a robot competition to demonstrate the contribution our platform can make, and propose and validate an approach to determine the most effective evaluation method for HRI performance.

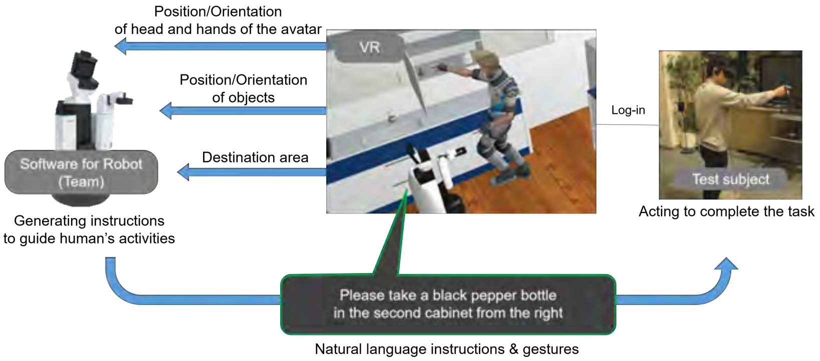
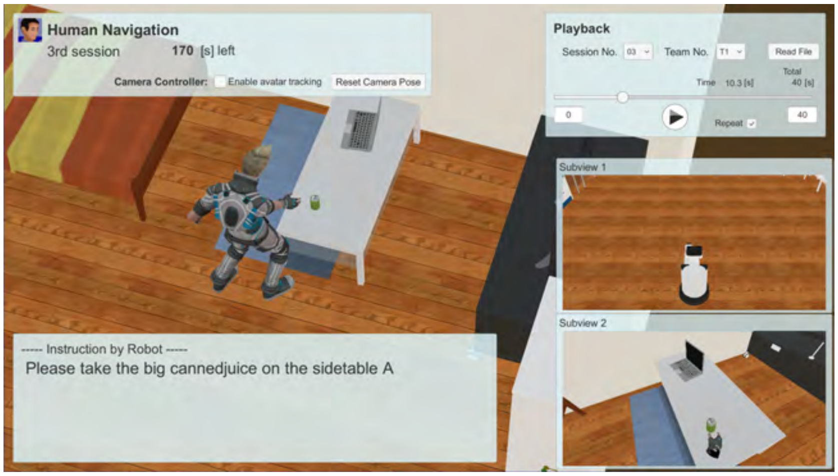

Position/Orientation
of head and hands of the avatar
Position/Orientation
of objects
Destination area
VR
Log-in
Software for Robot
(Team)
Generating instructions
to guide human's activities
Test subject
Acting to complete the task
Please take a black pepper bottle
in the second cabinet from the right
Natural language instructions & gestures

**Fig. 1.** Flow of the proposed competition task.

## 3 Case Study

We designed a robot competition task in which the virtual robot has to guide non-expert human users to achieve a certain task by verbal communication. Figure 1 shows the flow of the proposed competition task. At the start of the task, only the robot receives the pre-defined label name and position/orientation of the target object, the position/orientation/size of the destination area, and the position/orientation of other objects (i.e., all existing furniture objects and graspable objects except the target object). The position/orientation of the head and both hands of the avatar can be determined by the robot at any time. The role of the robot is to generate natural language instructions and gestures to properly control the navigation actions of a human instruction-follower (i.e., a test subject); in other words, the robot needs to convert positional information to natural language instructions and gestures. The generated natural language instructions are given to the avatar visually and audibly. The role of the test subject is to find/take a target object and then carry it to a destination by following the instructions from the robot. The test subjects is not told about room layouts, the target object, or the destination area. Accordingly, while the conventional robot competition tasks have focused on the understanding of requests from human operators by a robot, the roles of robot and human are reversed in this task.

We developed the competition system on the SIGVerse ver.3 platform [7]. Each team could develop controller software for the virtual robot by utilizing existing ROS-based software packages and libraries for real robots. The test subject logs in to a virtual avatar through the VR interface (e.g., Oculus Rift and Touch) and can interact with the virtual robot and objects in the virtual environment. With visual feedback and the tracking of head/hand poses, the test subject can manipulate virtual objects and the drawers/doors of furniture in a manner similar to that in which they would interact with the real physical environment. The test subject can move around the virtual environment by using



Human Navigation
3rd session 170 [s] left

Playback
Session No. 03 Team No. T1
Time 10.3 [s] Total 40 [s]
0 40
Repeat ☑

----- Instruction by Robot -----
Please take the big cannedjuice on the sidetable A

**Fig. 2.** Screenshot of the playback system for subjective evaluation by third-parties.

the joystick of a hand-held controller. Additionally, they can request instructions at any time by using a button on the hand-held controller.

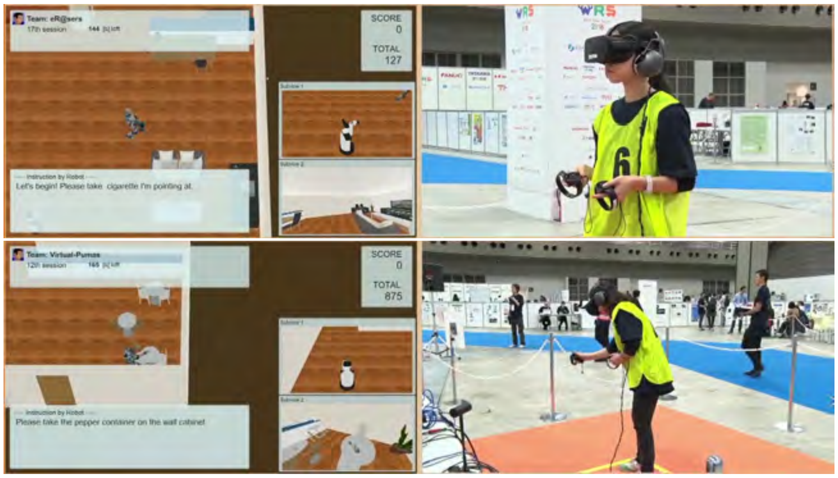
We organized a VR-based HRI competition at the World Robot Summit, Service Category, Partner Robot Challenge[3], held in October 2018 in Tokyo. This provided the context for the case study used in the research. The proposed competition task falls into the 'Human Navigation' in the rulebook[4]. We shared the competition software on GitHub[5] in advance, and only required participants to develop controller software (i.e., instruction generators) for the virtual robot. In the competition, the test subjects learned in advance, and practiced enough operational procedures to move, grasp objects, open/close doors and drawers, and request additional instructions from the robot in a test environment.

In addition to the robot competition, to analyze the relationship between objective factors and the result of subjective evaluation by humans, we asked third-parties to watch the recorded interaction histories from the competition and to subjectively evaluate these. Figure 2 shows the playback system for use in facilitating evaluation by third-parties. The evaluators are able to control the pose of a camera in the VR environment and observe all the events from an arbitrary point of view, and can watch a scene as many times as necessary. The behavior of the robot and the first-person user perspective are provided in sub-windows. The scores in the competition were hidden during the playback of the interaction histories, to prevent evaluator-bias. The evaluators rated the efficiency of the interaction between the robot and the test subject using a 5-point Likert scale questionnaire.

[3] `http://worldrobotsummit.org/en/wrc2018/service/`

[4] `http://worldrobotsummit.org/download/rulebook-en/rulebook-simulation_league_partner_robot_challnege.pdf`

[5] `https://github.com/PartnerRobotChallengeVirtual/`



— Instruction by Robot —
Let's begin! Please take cigarette I'm pointing at.

— Instruction by Robot —
Please take the pepper container on the wall cabinet

**Fig. 3.** Screen-shots of the Human Navigation task.

## 4 Results

Figure 3 shows screen-shots of the competition system and the behavior of a test subject. The recordings of all the competition sessions are available on a YouTube channel[6]. In the competition, we evaluated 7 robot-software controller submissions developed by 7 participating team. Each team competed in 28 sessions in parallel, making a total of 196 sessions. Each session lasted up to 180 seconds. Different room layouts were used in each session. We invited 16 test subjects to take part in the evaluation; each subject acted in the testing role up to twice for each team (either 7 or 14 sessions per tester). Consequently, 7.9 hours of human-robot interaction history data were collected. The history data include not only the sentences generated by the robots, but also all the test subject's embodied reactions, such as wandering motions due to a vague instructions from the robot, mistakes due to misunderstandings, and so on.

We evaluated interaction behaviors according to the competition rulebook[4]. As no criteria existed to directly evaluate the generated utterance, we formulated the scoring method empirically based on the following objectively measurable factors:

- Required time to complete the task.
- Required time to grasp the target object.
- Frequency of incorrect object-grasps.
- Number of instructions given by the robot.

The score distribution for each team is shown in Fig. 4. Each dot denotes the average score and the error bar denotes 95% confidence intervals. The average values and 95% confidence intervals of the objective measures are also shown in Fig 5–8. Table 1 compares the scores and the results of the objective measures. The time taken is assumed to be 180 seconds if a task remains incomplete then

[6] https://bit.ly/2QOjJAZ

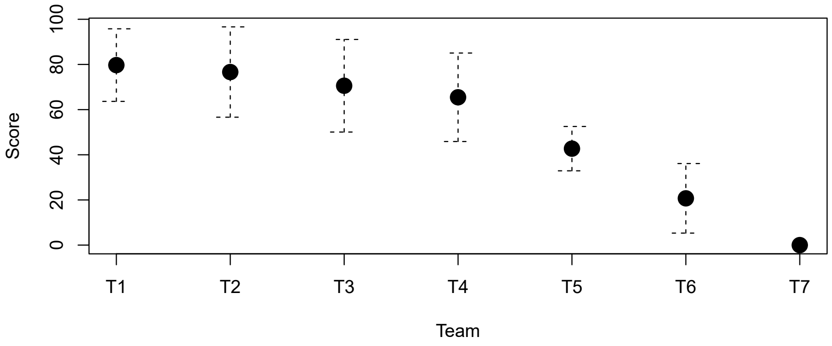


**Fig. 4.** Score distribution of the teams.

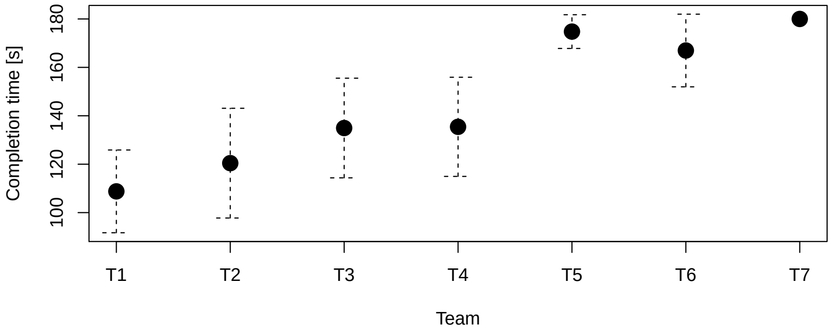


**Fig. 5.** Required time to complete the task for each session.

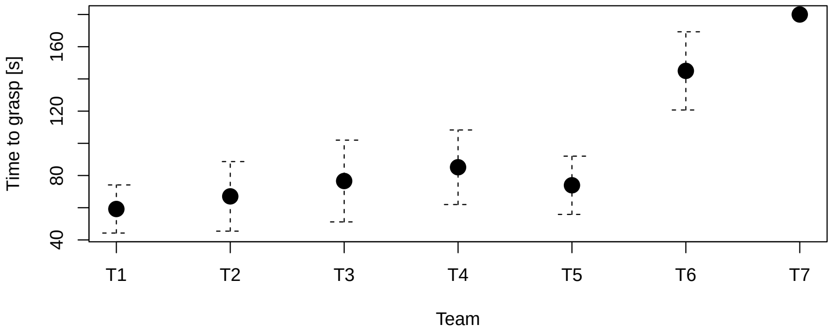


**Fig. 6.** Required time to grasp the target object for each session.

180 seconds have elapsed. The completion rate is the percentage of successfully completed sessions. The recognition rate for target objects is reported as the average of the percentage of the number of successful target-grasps out of the total number of object-grasps in each session; where test subjects were unable

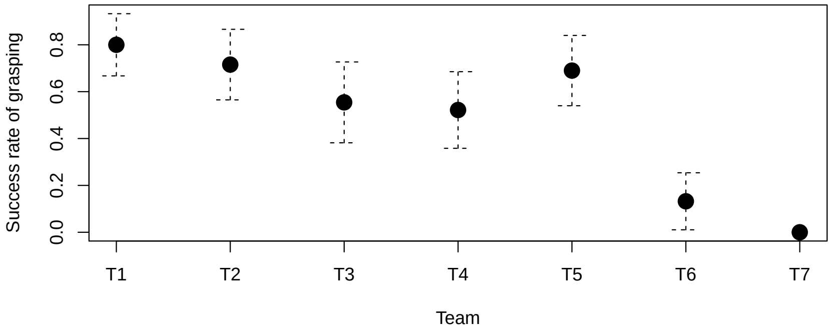


**Fig. 7.** Recognition rate of target object for each session.

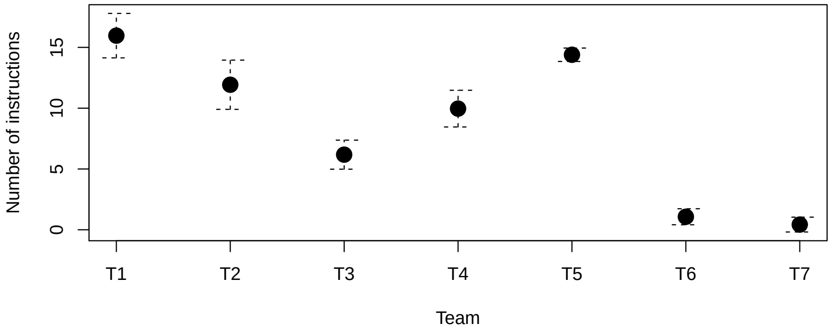


**Fig. 8.** Number of instructions given by robots in each session.

to grasp any object, the recognition rate is assumed to be 0%. For the other measures, average values of that measure in each session are shown. The letters denote groups of teams with no significant differences between them. If two teams do not have the same letter, the difference between those two teams is significant with $p < 0.05$. All teams were compared pairwise using the test for proportions and Steel-Dwass tests. These results indicated that the test subjects tended to fail the task more frequently and need a longer time to complete the task if the robot instructions were incomprehensible.

For a subjective evaluation, we invited 10 third-party evaluators. Each evaluator was asked to evaluate 168 sessions (7 teams × 24 sessions); we excluded 4 sessions which were affected by a log-file that was accidentally corrupted. Figure 9 shows the result of the subjective evaluation using a 5-point Likert scale questionnaire. The comparison of the subjective scores given in Table 2. The scores were compared pairwise using a post-hoc Steel-Dwass test. Some differences exist between the evaluated and competition scores, with a significant difference between T2 and T5 and in the order of T3 and T4; these can be attributed to the fact that we decided the scoring method subjectively.

**Table 1.** Comparison for objective measures.

| Team | T1 | T2 | T3 | T4 | T5 | T6 | T7 |
|---|---|---|---|---|---|---|---|
| Score | 79.7 | 76.6 | 70.6 | 65.5 | 42.7 | 20.7 | 0.0 |
| | A | A | A | A | | | |
| | | B | B | B | B | | |
| | | | | | | C | C |
| Completion rate [%] | 82.1 | 60.7 | 50.0 | 50.0 | 14.3 | 10.7 | 0.0 |
| | A | A | | | | | |
| | | B | B | B | | | |
| | | | | | C | C | C |
| Required time to complete the task [s] | 108.8 | 120.4 | 134.9 | 135.4 | 174.8 | 167.0 | 180.0 |
| | A | A | A | A | | | |
| | | | B | B | B | | |
| | | | | | C | C | C |
| Recognition rate of grasping the target [%] | 80.0 | 71.5 | 55.4 | 52.2 | 69.0 | 13.2 | 0.0 |
| | A | A | A | A | A | | |
| | | | | | | B | |
| | | | | | | | C |
| Required time to grasp the target [s] | 59.2 | 67.1 | 70.2 | 85.1 | 73.9 | 145.0 | 180.0 |
| | A | A | A | A | A | | |
| | | | | | | B | B |
| Number of instructions | 16.0 | 11.9 | 6.2 | 10.0 | 14.4 | 1.1 | 0.4 |
| | | | A | | | A | A |
| | | | B | | | | |
| | | C | | C | | | |
| | D | | | | D | | |

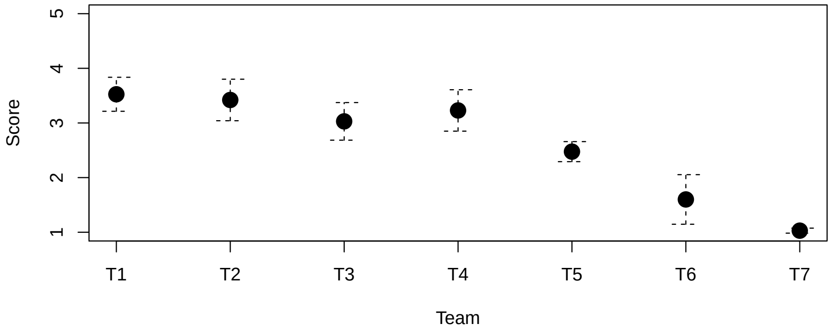
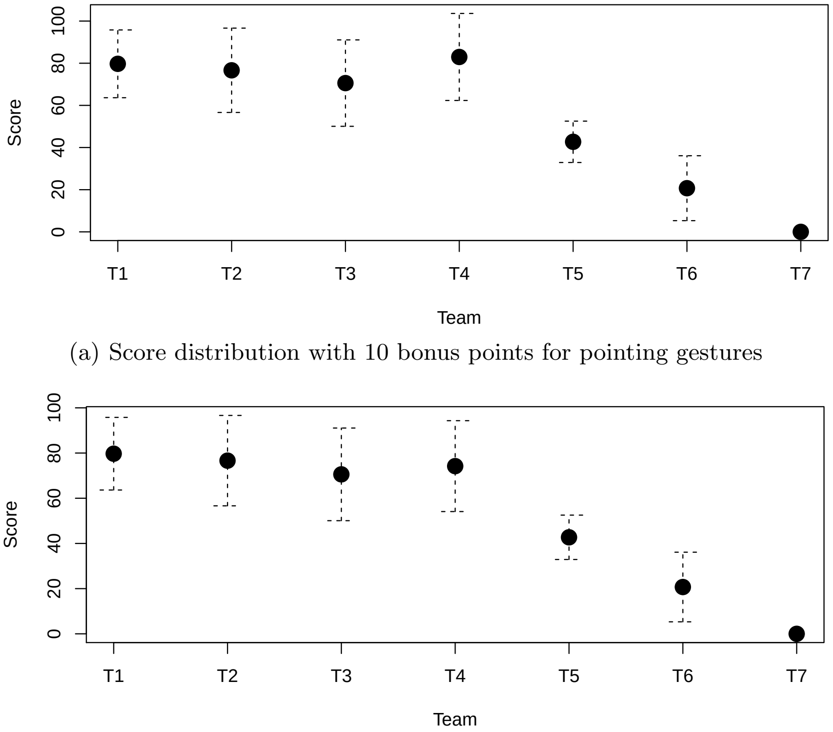


**Fig. 9.** Score of efficiency of interaction between each robot and test subjects subjectively evaluated by third-parties

To improve the validity of the scoring method, we analyzed the relationship with the subjective evaluation results and modified the scoring method used in the robot competition. Evaluator comments indicated that the key factor in

**Table 2.** Comparison result of subjective scores.

| Team | T1 | T2 | T3 | T4 | T5 | T6 | T7 |
|---|---|---|---|---|---|---|---|
| Efficiency of HRI<br>(Subjective result) | 3.5 | 3.4 | 3.0 | 3.2 | 2.5 | 1.6 | 1.0 |
| | A | A | A | A | | | |
| | | | B | | B | | |
| | | | | | | C | C |



(a) Score distribution with 10 bonus points for pointing gestures

(b) Score distribution with 5 bonus points for pointing gestures

**Fig. 10.** Modified score distributions that were added bonus points for pointing gestures.

the higher T4 score was the inclusion of robot pointing gestures. Accordingly, we awarded bonus points for pointing gestures; if the robot had been able to correctly point out a target object and destination, bonus points were added once for each of these achievements. Figures 10(a) and (b) show the modified scores when 5 or 10 points were added for each pointing gestures. From the results shown in Fig. 10(a), if 10 points were added, the score distribution would be distorted because T4's score would exceed those of T1 and T2. The order of score distribution in Fig. 10(b) corresponds to the subjective evaluation results. We were, thus, able to modify the evaluation method appropriately while continuing to utilize our system and the recorded interaction histories. The importance of such a revision process in the development of valid evaluation methods for HRI performance was, thereby, demonstrated effectively.

## 5 Discussions

With respect to system efficiency, the third-party evaluators required 15 hours on average (with a maximum of 30 hours) to review all the interaction histories and respond to the questionnaire. The time-consuming nature of the subjective evaluation of HRI performance is a weakness of this approach. Additionally, evaluation based on objective measures of human behavior, such as the time required to complete tasks, did not correspond to the subjective evaluation results. This supports our supposition that using only such objective measures is inadequate for the evaluation of HRI performance, and underscores the importance of creating an evaluation method that approximates the subjective evaluation result from objective factors.

In the robot competitions, we often face the similar problem in the process of defining scoring methods. Although organizers define the scoring method by discussion, often heated, the validity thereof has never been assessed to date. In other words, conventional evaluation criteria for the evaluation of HRI performance have only been subject to manual and empirical modification. Our proposed improvement process, modelled in this work, demonstrates its potential for improving evaluation methods without human intervention.

Another contribution of our study was the development of a VR system and the demonstration, through a case study in a robot competition, of its applicability in facilitating the observation of interaction histories. Playback of recorded events and human behaviors recorded by the system facilitates subjective evaluation by enabling the evaluator to observe the HRI they are evaluating form arbitrary points of view.

## 6 Conclusions

In this study, we aimed to devise a method for evaluating the performance of HRI that uses objective factors to approximate subjective evaluation. As a case study, we organized a VR-based HRI competition during the World Robot Summit held in October 2018. Analysis of the relationship between objective/observable factors and the results of subjective evaluation by third-parties was described.

One of the main contributions of this work was to demonstrate the necessity for a method of objective evaluation that can approximate subjective evaluation results from only objective factors. Another important contribution was the proposal and trial of an improvement process for the evaluation method that can function effectively without manual intervention.

In this study, we subjectively revised the evaluation method as a first step. However, numerical analysis methods such as multiple regression analysis should be used to more approximate the human subjectivity. Additionally, although our system can obtain embodied human behavior, such as changes to the head directions of avatars which can be assumed to present the frequency with which a test subject loses their way, such factors are not currently used for evaluation. In future work, we would like to focus on embodied behavior so as to determine a more effective evaluation method.

## References

1. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics **1**(1), 71–81 (2009)
2. Brščić, D., Kidokoro, H., Suehiro, Y., Kanda, T.: Escaping from Children's Abuse of Social Robots. In: Proc. of ACM/IEEE International Conference on Human-Robot Interaction. pp. 59–66 (2015)
3. Inamura, T., Mizuchi, Y.: Competition design to evaluate cognitive functions in human-robot interaction based on immersive VR. In: RoboCup 2017: Robot World Cup XXI, Lecture Notes in Artificial Intelligence. vol. 11175, pp. 84–94 (2017)
4. Kamide, H., Kawabe, K., Shigemi, S., Arai, T.: Anshin as a concept of subjective well-being between humans and robots in Japan. Advanced Robotics **29**(24), 1624–1636 (2015)
5. Kanda, T., Ishiguro, H., Imai, M., Ono, T.: Development and evaluation of interactive humanoid robots. Proc. of the IEEE **92**(11), 1839–1850 (2004)
6. Knepper, R.A., Tellex, S., Li, A., Roy, N., Rus, D.: Recovering from failure by asking for help. Autonomous Robots **39**(3), 347–362 (2015)
7. Mizuchi, Y., Inamura, T.: Cloud-based multimodal human-robot interaction simulator utilizing ROS and unity frameworks. In: IEEE/SICE Int'l Symp. on System Integration. pp. 948–955 (2017)
8. Nomura, T., Kanda, T.: RapportExpectation with a Robot Scale. International Journal of Social Robotics **8**(1), 21–30 (2016)
9. Rossi, S., Staffa, M., Bove, L., Capasso, R., Ercolano, G.: User's Personality and Activity Influence on HRI Comfortable Distances. In: Lecture Notes in Computer Science. vol. 10652, pp. 167–177 (2017)
10. Striegnitz, K., Denis, A., Gargett, A., Garouf, K., Koller, A., Theune, M.: Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In: Proc. of European Workshop on Natural Language Generation. pp. 270–279 (2011)