25th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
August 26-31, 2016. Columbia University, NY, USA

Assessment of Robot to Human Instruction Conveyance Modalities Across Virtual, Remote and Physical Robot Presence

Paul Robinette[1], Alan R. Wagner[1], Ayanna M. Howard[2]
[1]Georgia Tech Research Institute
[2]School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA, USA
paul.robinette@gtri.gatech.edu, alan.wagner@gtri.gatech.edu, ayanna.howard@ece.gatech.edu

Abstract— **Most Human-Robot Interaction (HRI) experiments are costly and time consuming because they involve deploying a physical robot in a physical space. Experiments using virtual environments can be easier and less expensive, but it is difficult to ensure that the results will be valid in the physical domain. To begin to answer this concern, we have performed an evaluation comparing participants' understanding of robotic guidance instructions using robots that were virtually, remotely, or physically present for the experiment. All but one set of experimental conditions gave similar results across the three presence levels. Further, we find that qualitative responses about the robots were largely the same regardless of presence level.**

I. Introduction

New simulation tools allow for easier creation and deployment of virtual environments than ever before. A variety of free and/or open-source toolkits now exist to create 3D models and 3D interactive environments. Human-robot interaction experiments can now be created in the same way as video games, with virtual agents in a virtual setting. These experiments can then be deployed online to take advantage of crowdsourcing. Thus, virtual experiments can be quickly and inexpensively created and completed with the help of participants on the internet. Results from the virtual experiments can be used in an iterative fashion to improve a robot for later, potentially more complicated experiments.

Currently, human-robot interaction experiments are usually performed in a laboratory or real-world setting where robots physically interact with human participants. Two alternatives to the traditional physical presence experiment paradigm now exist: a remote presence paradigm where the robot is located elsewhere so interaction occurs through video streaming and a virtual presence paradigm where the participant interacts with a simulation of a robot in a virtual environment. Some interactions, such as those involving emergency situations, are difficult to perform in a laboratory setting and can be impossible to perform in a real-world setting. In this work, we determine the extent to which virtual, remote, or physical environments can be used to evaluate human understanding of instructions given by robots. We compare our previous results using virtual robots [1] to results from two new experiments presented in this paper.

II. Related Work

Other researchers have explored the difference between various robot-presence paradigms in different research areas. Related work has found that participants are generally unlikely to follow "unusual requests," such as throwing textbooks in the trash, given by a remote presence robot as compared to a physically present robot [2]. Other work has found that physically present robots are rated better than their virtual or remote counterparts at coaching tasks [3]. In [4], robots were more effective in influencing human participants for 3D (real-world) tasks, but virtual agents were more effective for 2D tasks (i.e. tasks on a computer screen). Another study found that virtual agents and physical agents both had their benefits and problems when conducting discussions with participants about health topics [5]. These prior efforts give evidence that robots must be present to have a social effect on human participants in real-world tasks. Our work differs from these domains in that it is focused on participant understanding of robot instructions. We specifically test the ability for a human participant to understand robot instructions in emergency scenarios, but this work also applies in other domains where robots would have to convey information to nearby people. We do not ask participants to act on a virtual robot's guidance in a real-world situation, we simply ask participants what they think the robot is trying to communicate. We expect that participants will be able to understand instructions from a robot in each of the domains we test (virtual, remote and physical), even though related work suggests that participants will be more likely to follow these instructions from a physical robot than from other paradigms. A similar study that compared responses to videos of a robot approaching a human actor with real-world responses to a robot approaching a participant found little difference in the methodologies for that domain [6].

In the emergency domain, simulated emotions have been tested to improve human responses when a robot instructs a human to leave a room due to an unexpected emergency [7]. This work began similar to our previous work in [1] by using videos posted online to determine if participants could understand the emotions displayed by the robot [8]. The robot gave clear, verbal instructions aided by emotional actions, so participants were only tested on their ability to understand the robot's emotional actions and comply with its requests. In the past, we have used virtual environments to evaluate prototypes of our emergency guidance robot [1], [9], [10].

III. Virtual, Remote and Physical Presence Experiments

There are many different factors that will influence whether a virtual, remote or physical experiment is best for a particular human-robot interaction study. In this section, we briefly discuss the advantages and disadvantages of each.

A. Physical Presence Experiments

A physical human-robot interaction experiment requires the use of an actual robot (as opposed to a virtual robot) and thus typically requires physical space to perform the experiment. The physical space is most often a laboratory, but can also be a house, a public place or a workspace such as a factory or office. Regardless, the participant and/or the robot, along with any other necessary equipment, must be transported to the location of the experiment. Transporting a robot can be expensive and prone to errors. Robots used in experiments are typically under active development and thus are often unsuited for locations far from the laboratory. Convincing participants to come to a laboratory to perform an experiment can also be costly and results in self-selection: only those who have spare time and means of transportation are likely to participate. For this reason, many HRI experiments performed in university laboratories utilize students of the university as participants.

Many HRI experiments are appropriate to administer in a laboratory setting. For example, learning by demonstration typically requires participants to touch or closely observe a physical robot and does not depend on its surroundings in any particular way. Other HRI experiments, such as those involving search and rescue robots, present problems for experimenters. It is difficult to transform a laboratory into a believable disaster area. Previous work has presented experiments with a selection of props and a written scenario [7]. Others use specially built areas, such as the Disaster City at Texas A&M (see [11] and [12] for examples), but such areas are rare and expensive to create.

Most laboratory robots are under active development and thus are not completely free of errors. An error made by a robot during an experiment can potentially injure a participant and will almost certainly affect the response of the participant. Another potentially confounding factor is noise and other distractions from nearby laboratories as an experiment is in progress. Even the presence of the experimenter can affect the outcome of the experiment. Controlling these factors in a laboratory setting requires considerable effort.

B. Virtual Presence Experiments

We define a virtual human-robot interaction experiment as an experiment where participants observe and interact with a simulation of a robot through a computer. The robot must be entirely simulated and the interaction must take place in some sort of a virtual environment, similar to interactions in video games. This paradigm is attractive because most scenarios that are difficult to create in a laboratory are fairly easy to create using modern three-dimensional modeling software and game engines. It is possible to create the exact scenario that the experimenter would like to test in a virtual environment.

Another benefit of virtual experiments is that they can be deployed to participants anywhere in the world via the internet. Most game engines have an option to create a web-based game that can be loaded by a web browser plugin. Even participants with no video game experience can then interact with a virtual robot in the environment chosen by the experimenter. Recently, many experiments have used this technique to increase the number of participants who experience their robot [13], [8], [10]. Crowdsourcing an experiment on the internet using services such as Amazon's Mechanical Turk allows for a larger participant population base than is typically available for physical experiments. Other studies have found that Mechanical Turk provides a more diverse participant base than traditional human studies performed with university students [14], [15], [16], [17]. These studies found that the Mechanical Turk user base is generally younger in age but otherwise demographically similar to the general population of the United States (at the time of those studies, Mechanical Turk was only available in USA). Crowdsourcing also allows the experiment to be performed in parallel with typically much faster results than physical experiments. A virtual experiment that requires one hundred participants to each spend one minute interacting with a robot can have final results in minutes or hours, rather than the days or weeks necessary to recruit, assemble, and supervise such a population for a physical experiment. Thus, virtual experiments are most useful in situations where the experimenter wishes to iterate through prototypes or pilot studies rapidly.

The behavior of a simulated, virtual robot can be controlled easier than a physical robot. This is not applicable for user studies or other studies where the quirks of the robot are being examined, but can be helpful when exploring the behaviors a robot should perform to effectively influence a human participant. As an example, in our previous virtual studies, we measured the loss of trust a participant experienced in a robot after the robot performed specific errors [10]. If the robot performed other errors than those specified in the experimental case, or malfunctioned during the control case, then we would have to discard results and recruit additional participants. By using a simulation environment, we could tightly control the behavior of the robot and ensure that no unintentional errors were committed.

Virtual experiments are not without their problems. Participants must volunteer for the experiment, thus there is still self-selection bias in the participant population. This is balanced by allowing a much larger body of participants to volunteer through the use of the internet. While virtual experiments remove the possibility of noise and other distractions from nearby laboratories, they lose the ability to tightly control the environment in which a participant performs the experiment. A participant may choose to perform the experiment while watching television or listening to music and thus miss an important component. This can be mitigated by asking participants to explain their responses, thus ensuring that a thoughtful process was used in their actions, and by asking participants questions which check their understanding of the experiment. An additional issue is that other studies have found that social interactions between humans and robots are not always well represented through non-physical presence [2], [3].

C. Remote Presence Experiments

The use of video streaming technology for remote presence experiments allows for a happy medium between virtual and physical experiments. In remote experiments, participants view a video of a robot (either prerecorded or live) and complete their tasks through a web interface. Remote experiments allow participants to observe the actual robot hardware as it performs its experimental tasks, but do not allow participants to touch the robot. The use of prerecorded videos allows experimenters to gather participant feedback on designs that are still under active development and might not perform perfectly in every trial. Additionally, videos can be recorded or streamed from a laboratory setting, which allows participants to be involved in the experiment even if they cannot physically travel to the laboratory. Remote experiments can often be crowdsourced, similar to virtual experiments. Videos can be placed on a service like Amazon's Mechanical Turk and be available to a larger participant population than physical experiments. These experiments have similar drawbacks to virtual experiments, with the one major improvement being the use of the actual robot in the experiment to remove any effect simulation artifacts would have on participant responses. As mentioned above, the remote presence paradigm has been previously tested in [2], [3].

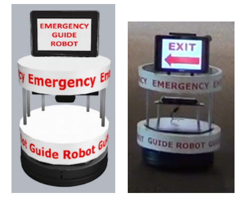
IV. Experimental Setup

Three experiments were performed and compared to determine the effect of the virtual, remote, and physical presence paradigms on participant responses. The first experiment asked participants on Mechanical Turk to rate videos of virtual robots performing gestures (complete results from this experiment alone can be found in [1]). The second experiment tested the remote paradigm by recording videos of real robots performing gestures and again using Mechanical Turk to gather data. The third experiment tested the physical paradigm by gathering participants in a laboratory environment and measuring their response to a real robot performing instructions in front of them. Each experiment used the platforms or a subset of the platforms discussed below.

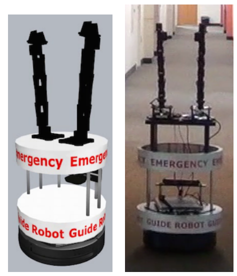
A. Robot to Human Information Conveyance Modalities

In [1], we defined and evaluated methods for a robot to instruct a participant to 1) proceed to the left or right, 2) proceed forward, 3) turn around, or 4) stay in place. A brief description of these methods follows, but a full description can be found in [1].

(a) Baseline Platform



(b) Dynamic Sign Platform



(c) Multi-Arm Gesture Platform

(d) Multi-Arm Gesture with Dynamic Sign Platform

Fig. 1: The virtual and remote/physical robots used in this study. Virtual platforms are shown on the left and remote/physical platforms on the right for each platform.

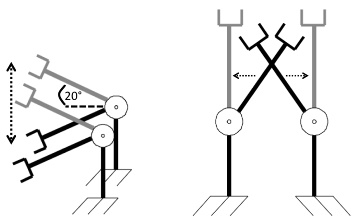
1) Baseline Platform: The robot platforms used in these experiments were designed based on the Willow Garage Turtlebot 2 due to its ease of use and general availability. The baseline platform (Figure 1a) used only its motion to convey instructions. For directional guidance, it turned in the direction the participant should proceed (e.g. left, backward or forward) and oscillated about that direction by 30 degrees to its left and right. To convey the stay in place instruction, it spun in place.

2) Dynamic Sign: The dynamic sign platform (Figure 1b) coupled the baseline platform with written instructions provided to the participant. Written instructions and symbols such as arrows were deployed through a mounted tablet interface (11.1" Samsung Galaxy Tab). The instruction screens can be seen in Figure 2. Additional, the robot carried signs that declared the robot's purpose as an emergency guidance aid. The signs were in two cylindrical components: one on the top of the Turtlebot and one just above the base. The top sign displayed "Emergency" in each of the four cardinal directions around the cylinder and the bottom sign displayed "Guide Robot" in the same manner.

3) Multi-Arm Gesture Platform: The Multi-Arm Gesture platform combined the baseline platform with a gestural interface for providing instructions. Gestures were enabled by mounting two robot arms (PhantomX Pincher AX-12) onto the base. For directional instructions, the whole platform turns to face the direction it wishes the human to proceed and the arms point forward (Figure 3a). The arms then oscillate

(a) Left (b) Turn Around (c) Forward (d) Wait

Fig. 2: Dynamic Signs Text and Symbols



(a) Example of arms pointing. In this figure, the arms are gesturing to the left. (b) Example of two arms crossing.

Fig. 3: Examples of Arm Gestures. In each case, the arm moves from the solid black position to the solid gray position and back in the direction of the dotted arrows.

slightly along the vertical axis to "wave" the participant in the required direction. For the stay in place instruction, the robot faces the participant and alternates between both arms straight up and arms crossed (Figure 3b). This robot also carried the Emergency Guide Robot signs.

4) Multi-Arm Gesture with Dynamic Sign: Our final platform combined the Multi-Arm Gesture platform with the Dynamic Sign platform (Figure 1d). The robot performed the same motions and arm gestures as in the Multi-Arm Gesture condition. Participants could also see the tablet as it approached them, but in the left and forward instruction conditions the robot turned to point in a particular direction and the tablet was obscured for the rest of the condition.

B. Virtual and Remote Presence Experiments

To evaluate human understanding of the virtual and remote robot guidance modalities we utilized two between-subjects experiments. The experimental setup was similar for each paradigm. Participants were recruited and the studies conducted using Amazon's Mechanical Turk service. This service allows an experimenter to recruit a large number of participants from across the United States and several other countries. The original virtual study contained additional variables not presented here and required 192 participants. Of those, 64 participants rated the same robots as in the remote paradigm and those results are presented below to contrast with the other paradigms. See [1] for additional details about the virtual study. A total of 64 participants performed this survey for the remote paradigm.

Participants began each study by reading and acknowledging a consent form. Next, they completed a demographic survey collecting information about gender, age, nationality, occupation, and education. Then, the participants were presented with videos of one particular robot performing each of the four instructions (one instruction for each video). Each participant was only shown the videos for one robot. Participants were asked "What is this robot asking you to do?" after each demonstration and given the following options to choose from:

- Go to the left
- Go to the right
- Go forward
- Turn around
- Stay in place
- Follow robot
- I do not know

They were then asked to give a confidence value for their answer (1-7) and asked to explain their answer in paragraph form. Several instructions were given as multiple choice answers for each video, including some that never appeared in the test so that participants could not use process of elimination to give an answer. The dependent variable being measured was their answer to the multiple choice question and the comments were used to understand that answer. The order of the videos was randomized.

In the virtual study, videos were each approximately 15 seconds long and 800 x 600 pixels in size. In the remote study, videos were each approximately 40 seconds long and 1280 x 720 pixels in size. Participants were paid $0.50 for completing the virtual study and $1.00 for completing the remote study. Participants were only allowed to take one study, not both. IRB approval was obtained before either study began.

C. Physical Presence Experiment

To evaluate our robot in a physical presence experiment we again used a between-subjects study. A total of 48 participants were recruited by posting fliers around the Georgia Tech campus and by emailing students. Only three conditions were tested in this study: the Baseline, Multi-Arm Gesture and Multi-Arm Gesture with Dynamic Sign platforms. The Dynamic Sign was not tested because it was unable to be seen at a distance in prior experiments and thus would not be suitable for real-world deployment.

The experiment began with participants reading and signing a consent form. Participants then lined up along a wall facing the robot's demonstration point in an office environment. The experiment location was in the same building as the videos for the remote experiment were recorded. A hallway in front gave the impression that participants could travel to the left, right, or forward. Multiple participants observed the robot's demonstrations in each trial; however, participants were instructed not to communicate with each other during the procedure and an experimenter was present to supervise. Participants observed a single platform perform all four instructions and answered survey questions about each. The survey questions were identical to those in the virtual and remote experiments except on paper instead of a webpage. In our prior experiments we failed to find an ordering effect, so in this experiment we did not randomize

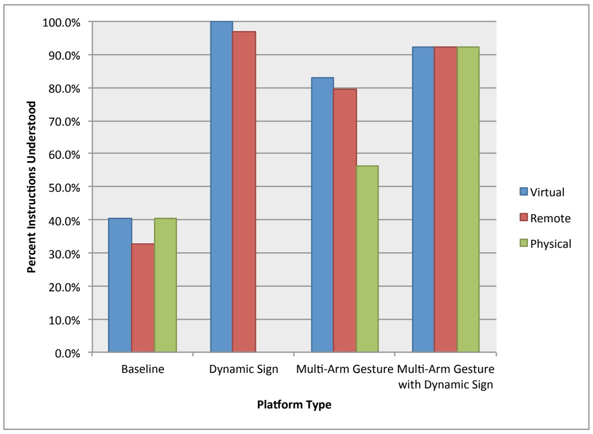


Fig. 4: Results for the physical experiment compared with corresponding platforms in the virtual and remote experiments. Note that the Dynamic Sign platform was not tested in the physical experiment. Multi-Arm Gesture was the only platform with statistically significant ($p < 0.05$) results between presence levels. This discrepancy is discussed further in Section V-C.2.

the order of the instructions for each session. Each session observed first the backward, then the left, then the forward and finally the stay instruction. Participants were allowed to revise previous answers as long as the experiment was in progress. After the demonstrations, participants completed a demographic survey and were allowed to ask any questions they might have about the robot or the experiment. Participants were compensated $10.00 for their time. IRB approval was obtained before the study began. All robot demonstrations were automated. The experimenter controlled which demonstration would be presented at which time using a laptop interface.

V. Results

Results from all three experiments can be seen in Figure 4. A summary of the results for virtual platforms is presented below, followed by our findings for remote and physical platforms.

A. Virtual Presence Experiment

In general, the virtual presence experiment found that the baseline platform was difficult to understand but that platforms with dynamic signs and robot arms could be understood. Combining the dynamic sign and multi-arm gesture platforms produced the best results. These results were promising, but required validation in the remote and physical paradigms. See [1] for a complete discussion of the virtual results. Results relevant to our comparison of presence levels are shown below.

B. Remote Presence Experiment

A total of 64 participants (denoted as N below, 39% female, mean age of 34.1 years old, 97% from USA) responded to a total of 256 questions (denoted as R below) in this study. Their answers were compared to answers ($N = 64, R = 256$) about the same platforms in the virtual study.

1) Baseline: Overall, 32.8% of instructions were correctly understood when presented by the remote Baseline platform. This is 7.8% worse than the virtual platform results (see Section V-C.1 for statistical analysis). As in the virtual study, the left instruction was generally understood but the other instructions were not. Participants interpreted the oscillating motion in the forward and backward instructions as the robot shaking its "head" to indicate "no." They generally interpreted this "no" to mean that they should stay in place and not proceed in any direction. They understood the spinning motion that we intended to mean stay in place as an indication that they should turn around. Comments indicated that participants were unsure about their answers even though they could clearly see the robot.

2) Dynamic Sign: Again, a small difference was found between the remote and virtual platforms for the dynamic sign. Overall, 96.9% of responses indicated participants understood the instructions. We expected 100% of responses to indicate understanding, as in the virtual case, but one person answered that it was telling him to go right when it was actually indicating left (although the explanation given indicates the participant understood the intention) and another participant reported that one of the four videos would not play for technical reasons. The difference was not statistically significant ($\chi^2(1, N = 32, R = 128) = 0.642$, $p = 0.154$). Thus we can conclude that the dynamic sign is understandable when experienced at this distance.

3) Multi-Arm Gesture: The remote Multi-Arm Gesture platform performed about the same as the virtual platform. Overall, 79.7% of instructions were understood for the remote platform compared to 82.8% for the virtual platform (see Section V-C.2 for statistical analysis). Even though there was some confusion as to whether the forward and backward instructions actually meant to follow the robot, a large majority of participants understood the instructions. Five participants were unable to understand the stay instruction. One thought it was indicating "no," and thus to turn around and go backwards, by crossing its arms. The others answered "unknown" and indicated they had no guess. One reported that he thought the robot was panicking.

4) Multi-Arm Gesture with Dynamic Sign: The remote Multi-Arm Gesture with Dynamic Sign performed exactly the same as its virtual counterpart. Comment responses were also very similar to the virtual case. When the dynamic sign was visible throughout the entire video (the backwards and stay instructions) participants answered exactly as we expected. When it was obscured for a portion of the time (the left and forward cases) a small number of participants were unable to understand the instructions (one in the left case, four in the forward case).

C. Physical Presence Experiment

A total of 48 participants (denoted as N below, 30% female, mean age of 24.7 years old) responded to a total of 192 questions (denoted as R below) for this experiment over fourteen sessions. The number of participants in each session ranged from one to seven. The results were broken up by platform type and compared to corresponding platform types in the virtual and remote experiments. Results can be seen in Figure 4. Participants viewed an actual robot performing live demonstrations, so there were occasional robot failures. Of the 56 total gestures performed, three were repeated. One due to the arms losing sync during the wave procedure (one arm was up and the other was down instead of moving together), one due to operator error (the wrong gesture was chosen) and one due to a participant arriving late. In each case, participants were instructed to ignore the failed demonstration and only answer survey questions about the correct one.

1) Baseline: The Baseline platform showed no difference in the physical experiment when compared to the virtual experiment (40.6% of participants understood the instructions) and a 7.8% greater understandability when compared with the remote experiment results ($\chi^2(2, N = 48, R = 192) = 0.191$, $p = 0.575$ across all three presence levels). No surprises were found in the comments, either. As in the previous experiments, participant comments generally indicated confusion rather than understanding for this platform.

2) Multi-Arm Gesture: The Multi-Arm Gesture platform did show a significant difference based on presence level ($\chi^2(2, N = 48, R = 192) = 0.393$, $p = 0.001$). Over all instructions, the physical platform was 26.5% less understandable than in the virtual experiment and 23.4% less understandable than in the remote experiment. This was a curious result, so further analysis was warranted. Comparing responses to individual instructions across the three presence levels revealed that the biggest difference was in the understandability of the backward instruction (Figure 5). The other three instructions ranged from a 6.2% to a 18.7% difference between presence conditions (left: $\chi^2(2, N = 48, R = 48) = 0.274$, $p = 0.360$, forward: $\chi^2(2, N = 48, R = 48) = 0.254$, $p = 0.413$, stay: $\chi^2(2, N = 48, R = 48) = 0.142$, $p = 0.695$), but the backward instruction had a 62.5% difference between the remote and physical conditions and a 68.7% difference between virtual and physical ($\chi^2(2, N = 48, R = 48) = 0.393$, $p < 0.001$). We believe that this is entirely due to our experiment location. Recall that participants were lined up along a wall to observe the robot and that we had hoped the doors in the wall would provide a believable route in the backwards direction. Instead, participants indicated in the comments that the robot was pointing at them but that no route was available behind them and thus the robot must be telling them something else. Five believed that the robot wanted them to follow it, five decided on stay in place and four thought the robot actually wanted them to move forward, believing that this was a beckoning gesture.

3) Multi-Arm Gesture with Dynamic Sign: The final platform performed exactly the same over all four instructions as it did in both the virtual and remote conditions. For each instruction, only one or two participants did not understand the direction correctly. One participant indicated that the robot

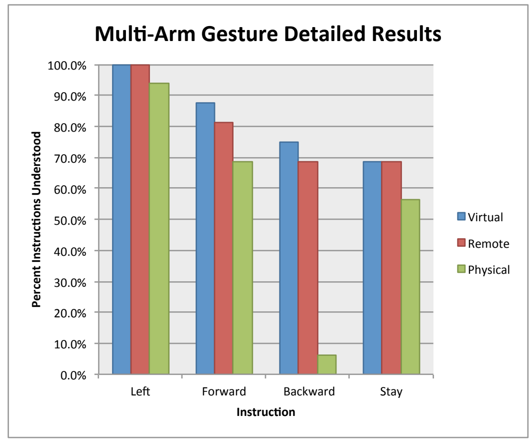


Fig. 5: Detailed results of the Multi-Arm Gesture platform across all three presence levels. The only major anomaly is participants' inability to understand the "backward" instruction in the physical experiment.

gave two different directions during the backward instruction and was recorded as an unknown in our results because they could not distinguish between the two directions. Both participants who did not understand the forward instruction indicated that they thought the robot wanted them to follow it, which is similar to the remote and virtual experiments.

VI. Discussion

All of the results in the remote experiment and the physical Baseline and Multi-Arm Gesture with Dynamic Sign show little difference between the virtual, remote, and physical experiments. These platforms ranged from a 0% difference to a 7.8% difference over all instructions. None of these results were significant at a $p = 0.05$ level. The only anomalous platform condition, Multi-Arm Gesture in the physical experiment, only had a significant difference in a single instruction. As explained above, we believe that is because the location of the experiment did not have an obvious exit route in the direction the robot indicated, and thus this result is spurious. Qualitatively, participants gave similar explanations for their interpretations of the instructions in all three experiments.

As in our previous study [1], participants attempted to understand the robot's instructions with any information that they had. All participants gave explanations for their response. One participant tried to help our design process by suggesting we use colored lights and loud sounds to aid people with cognitive disabilities or people taking narcotic medication in understanding the robot's instructions. Many participants in the remote experiment observed that there is a green light on the back of the Turtlebot base. This light shows brightly in the video but is dim in person, so we had not considered it as a potential feature of the robot. Participants who noticed the light interpreted the green light as a signal to mean go forward or follow the robot. This was only observed in cases where no intelligible guidance instructions could be seen by participants, such as the Baseline platform. Any participants who could see arm movements, for example, ignored the green light and focused on the intended gestures. The same effect was seen when the robot would tend slightly to the left or to the right at the end of its path: participants would interpret this minor deviation as an indication that they should proceed in that direction. We can infer that if participants cannot understand the instructions of a robot they will attempt to glean knowledge out of any feature visible, no matter how insignificant or unintentional that feature was to the robot designer.

A valid study should have as diverse a population of participants as the expected population of future users. When the demographics of our remote study population were compared with our physical study population, we found that the physical study population was much younger (average of 9.4 years younger) and had a somewhat lower female to male ratio. Moreover, 46 of 48 participants in the physical study indicated that they were students. This is not surprising given that recruitment for the study was mainly confined to the Georgia Tech campus, but many other studies use a similar recruitment strategy without attempting to gather a wider audience. For our study, this did not matter as participants in the physical and remote studies gave almost identical answers, but other studies may not be so fortunate.

VII. Conclusion

Our results in this study show that there is little difference between virtual, remote and physical robot presence in HRI experiments that focus on understanding instructions given by robots. Only one platform had a significant difference in responses between the presence levels tested and we have concluded that was an anomaly related to our testing setup. We do not generalize this result to mean that all HRI experiments can be performed in a virtual setting, but rather that this is one experiment in the subset of all experiments that can be performed in a virtual setting as accurately as in a physical setting. We feel confident that other experiments which rely on a participant's ability to understand instructions conveyed by a robot would be valid using a virtual setting.

Performing this experiment first in a virtual setting, then in a remote setting, and finally validating in a physical setting allowed us to generate seven virtual robots, prune these to four useful physical designs and then use crowdsourcing to again prune our platforms and experimental conditions before performing a costly and time-intensive physical presence experiment with three robots. Early tests with a large number of participants (192) conducted in the virtual domain allowed us to focus on platforms that were worthwhile in later, more costly and time consuming tests with 48 participants. We believe other design processes can benefit from a similar process when developing new robots for HRI tasks.

Future work includes determining the extent to which experiments in the emergency domain can be evaluated in virtual environments. Experiments in the emergency domain must be simulated: in a physical setting they are simulated with props and narratives, in a virtual setting they are simulated with computer graphics. In future work, we will present experiments that determine the mapping between results taken from the virtual domain and those from the physical domain.

VIII. Acknowledgements

Support for this research was provided by the Motorola Foundation Professorship and Air Force Office of Scientific Research contract FA9550-13-1-0169. The authors would like to acknowledge Sergio Garcia-Vergara and Paul Drews for their help during video recordings.

References

[1] P. Robinette, A. Wagner, and A. Howard, "Assessment of robot guidance modalities conveying instructions to humans in emergency situations," in *RO-MAN, 2014 IEEE*, IEEE, in press 2014.

[2] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[3] J. Wainer, D. J. Feil-Seifer, D. Shell, M. J. Mataric, *et al.*, "Embodiment and human-robot interaction: A task-based perspective," in *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pp. 872–877, IEEE, 2007.

[4] K. Shinozawa, F. Naya, J. Yamato, and K. Kogure, "Differences in effect of robot and screen agent recommendations on human decision-making," *International Journal of Human-Computer Studies*, vol. 62, no. 2, pp. 267–279, 2005.

[5] A. Powers, S. Kiesler, S. Fussell, and C. Torrey, "Comparing a computer agent with a humanoid robot," in *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*, pp. 145–152, IEEE, 2007.

[6] S. N. Woods, M. L. Walters, K. L. Koay, and K. Dautenhahn, "Methodological issues in hri: A comparison of live and video-based methods in robot to human approach direction trials," in *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pp. 51–58, IEEE, 2006.

[7] L. Moshkina, "Improving request compliance through robot affect.," in *AAAI*, 2012.

[8] S. Park, L. Moshkina, and R. C. Arkina, "Recognizing nonverbal affective behavior in humanoid robots," *Intelligent Autonomous Systems 11: Ias-11*, p. 12, 2010.

[9] P. Robinette and A. Howard, "Trust in emergency evacuation robots," in *10th IEEE International Symposium on Safety Security and Rescue Robotics (SSRR 2012)*, 2012.

[10] P. Robinette, A. R. Wagner, and A. M. Howard, "The effect of robot performance on human-robot trust in time-critical situations," Tech. Rep. GT-IRIM-HumAns-2015-001, Georgia Institute of Technology. Institute for Robotics and Intelligent Machines, January 2015.

[11] B. A. Duncan and R. R. Murphy, "Comfortable approach distance with small unmanned aerial vehicles," in *RO-MAN, 2013 IEEE*, pp. 786–792, IEEE, 2013.

[12] R. R. Murphy, "Human-robot interaction in rescue robotics," *IEEE Transactions on Systems, Man, and Cybernetics, Part C Applications and Reviews*, vol. 34, no. 2, pp. 138–153, 2004.

[13] J. Orkin and D. Roy, "The restaurant game: Learning social behavior and language from thousands of players online," *Journal of Game Development (JOGD)*, vol. 3, no. 1, pp. 39–60, 2007.

[14] G. Paolacci, J. Chandler, and P. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[15] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?," *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[16] A. J. Berinsky, G. A. Huber, and G. S. Lenz, "Evaluating online labor markets for experimental research: Amazon. com's mechanical turk," *Political Analysis*, vol. 20, no. 3, pp. 351–368, 2012.

[17] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce*, pp. 209–218, ACM, 2010.