# Vero: A Method for Remotely Studying Human-AI Collaboration

Jess Hohenstein
Cornell University
jch378@cornell.edu

Lindsay E. Larson
Northwestern University
lindsaylarson200@gmail.com

Yoyo Tsung-Yu Hou
Cornell University
th588@cornell.edu

Alexa M. Harris
Northwestern University
alexa.ma.harris@gmail.com

Aaron Schecter
University of Georgia
aschecter@uga.edu

Leslie DeChurch
Northwestern University
dechurch@northwestern.edu

Noshir Contractor
Northwestern University
nosh@northwestern.edu

Malte F. Jung
Cornell University
mfj28@cornell.edu

## Abstract

*Despite the recognized need in the IS community to prepare for a future of human-AI collaboration, the technical skills necessary to develop and deploy AI systems are considerable, making such research difficult to perform without specialized knowledge. To make human-AI collaboration research more accessible, we developed a novel experimental method that combines a video conferencing platform, controlled content, and Wizard of Oz methods to simulate a group interaction with an AI teammate. Through a case study, we demonstrate the flexibility and ease of deployment of this approach. We also provide evidence that the method creates a highly believable experience of interacting with an AI agent. By detailing this method, we hope that multidisciplinary researchers can replicate it to more easily answer questions that will inform the design and development of future human-AI collaboration technologies.*

## 1. Introduction

With improvements in automation technology, artificial intelligence (AI) agents are beginning to take on more complex roles in their interactions with people and, consequentially, are increasingly viewed as more than mere tools [1, 2], and the IS community is uniquely situated to investigate how to prepare for a future of human-AI collaboration. Despite the many identified open questions and topics of interest [3, 4] coupled with the development of scales [5] and frameworks [6] to measure such outcomes, researchers are largely lacking the platforms and methods that will allow us to not only understand how such systems affect work in teams but also to inform future development of such agents. The growing prevalence of human-AI collaboration presents a need for experimental methods that allow researchers from diverse academic communities to more easily study complex social interactions with AI.

Furthermore, with advances in communication technologies coupled with the recent COVID-19 pandemic, teams are increasingly collaborating remotely. Similarly, researchers are facing unprecedented challenges in running in-person experiments, with many scrambling to figure out ways to move studies online. In response, some have recently adopted video conferencing platforms as a research tool and noted the numerous benefits for both researchers and participants.

In this paper, we present a unique experimental method that allows us to examine user interactions with an AI teammate using a convenient, accessible, and easy-to-use video conferencing application (i.e., Zoom). Through a combination of curated content and Wizard of Oz (WoZ) methods, our paradigm leads participants to believe that they are interacting with an AI teammate, even though they are actually interacting with a human. Analyses of post-interaction data confirm the viability of this claim, with the majority of participants believing the manipulation, regardless of pre-existing perceptions of technology. This paper makes three main contributions:

- A method that makes human-AI collaboration research accessible to a broad community by eliminating the need to develop an AI agent
- An extension of WoZ methods that increases scalability and generalizability through multiple simultaneous study sessions using natural spoken language over Zoom
- A case study demonstrating & recommendations for the successful deployment of this method

## 2. Related Work

### 2.1. Human-AI Collaboration

As the capabilities of automated systems continue to increase, AI agents are able to take on more complex

URI: https://hdl.handle.net/10125/79360
978-0-9981331-5-7

tasks within teams, acting less as tools and more like teammates capable of making independent and team-oriented decisions [1].

In the human-AI collaboration literature, some of the most pressing open questions involve user perceptions of AI humanness, capabilities, and transparency [7, 4, 3]. Researchers within IS have started to investigate such questions, with recent work examining topics ranging from symbiotic co-evolution of human-AI teams [8], trust of intelligent systems [9, 10] and interaction design [11, 6, 12, 13] to developing scales for measuring perceived AI intelligence and anthropomorphism [5].

Despite the unique ability and preparedness of the multidisciplinary IS community to investigate how to prepare for a future with AI as a collaborator on human teams, the methodologies used have major limitations that restrict the ease of performing research as well as the generalizability of the results therein.

**2.1.1. Wizard-of-Oz Methodologies** Existing human-AI collaboration research has been somewhat limited by the need to develop a physical or virtual agent for participants to interact with, meaning that some existing work has relied on methods that may not generalize to real team contexts, such as workshops and interviews [11].

Other work has used Wizard-of-Oz (WoZ) methods [14], wherein, unbeknownst to participants, an experimenter acts as the intelligent agent to simulate an interaction with an intelligent agent, e.g., [10, 12, 13]. In addition to being used to imagine and test future systems that are not yet technically possible [15], WoZ has also been used to generate ideas about what interactions with intelligent systems can or should look like [16]. However, the scalability and generalizability of these applications of WoZ methods are limited by the need to have participants physically present in a lab, the difficulty of running multiple study sessions simultaneously, and the time required to implement new agent modalities. Due to these factors, researchers cannot easily recruit from a wide population, quickly run trials, or easily alter independent variables of interest. Additionally, many researchers' ability to run in-person experiments has been severely limited due to the COVID-19 pandemic.

Our novel methodology is the first, to our knowledge, that couples a WoZ method with videoconferencing to create a scalable approach by simulating multiple simultaneous interactions with an AI agent. More specifically, this novel approach allows us to overcome the aforementioned limitations of previous work, as we can recruit from any population with internet access, run multiple simultaneous trials, allow participants to take part without coming to a physical lab, and rapidly alter AI agent characteristics of interest.

### 2.2. Video Conferencing as a Research Platform

Video conferencing as a research tool has become increasingly popular due to its relatively low cost [17, 18], ability to access larger numbers of more diverse participants [17, 18, 19], elimination of the need for participants to travel, efficiency [19], and ability to reduce various unpredictable circumstances [18].

Using Zoom as our research platform and employing WoZ methods, we created a unique experimental method that allows us to study interactions between an AI agent and a remote team without the need to develop a functional AI agent. This method will allow researchers to investigate relevant questions about the design and development of future AI agent technologies *before* expending the resources needed to create such agents.

## 3. Method

We introduce a novel experimental method that combines a WoZ method with curated theatrics to simulate an AI teammate. Its development parallels that described by [20]. More specifically, this method came about as a pivot from a previous design involving a physical robot in a lab, which became useless in light of the COVID-19 pandemic and the inability to bring participants into the lab for a study. After brainstorming and discussing possible options amongst the research team, we realized that if we hoped to run our experiment, we needed a way to do it remotely. With the success of other researchers in using videoconferencing as a research platform [21, 18] coupled with similar previous work employing WoZ methods [10, 12, 13], we decided to employ WoZ methods and design an intelligent agent that could be controlled by a confederate over Zoom.

A number of potential interaction modalities and appearances were envisioned and tested, including using voice modulation and a computer voice using text-to-speech. However, these methods were difficult for confederates to execute seamlessly, and interactions felt slow and unnatural. During a brainstorming discussion, we decided to test whether we could make participants believe that a confederate using natural spoken language was an intelligent agent simply by *telling them* that they would interact with an intelligent

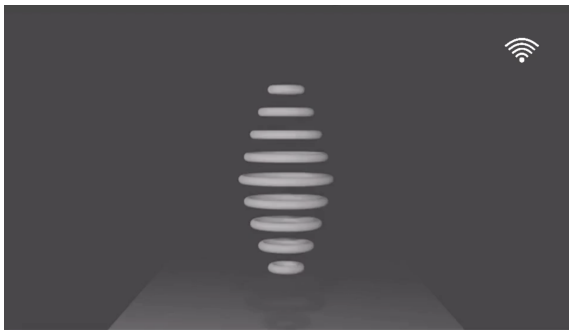

**Figure 1. The intelligent agent, Vero, is actually a human confederate that is controlling a Zoom background.**

agent. Informal pilot sessions suggested that this was indeed the case, so we proceeded with employing this novel methodology.

The development of this method was secondary to a main study investigating human-AI teaming (i.e., the case study), and the goal of this paper is to make the method accessible to other researchers by first describing how to employ it and then using the case study to illustrate how it was successfully used by our research team. To further facilitate replication, we also provide the necessary materials as Supplementary Files.

## 3.1. Experimental Method

To help other researchers replicate our experimental method, we will detail how we described our experiment to participants and created and controlled our intelligent agent, Vero.

**3.1.1. Creating the AI Agent** Vero, shown in Figure 1, is an animated intelligent agent that was created in Blender [22]. Blender is a free, open-source 3D rendering software that can be used to create a variety of animations, such as the simple ones that we created to represent Vero. For our purposes, we created five different animations representing Vero's five possible interactions with participants: 1) a default, floating state, 2) a listening, nodding state, 3) a speaking, vibrating state, 4) a wanting to speak, jumping state, and 5) a waving state. Researchers should keep in mind that an increased number of possible agent actions will increase the mental load of the human confederate controlling the agent. These animations are included as Supplementary Files.

To facilitate post-experiment video analysis that automatically coordinates Vero's states with times in the video recordings, each animation also includes an easy-to-recognize animation indicator in the form of a variable WiFi symbol at the top right, as shown in Figure 1. Using a series of researcher-created animations to represent an agent means that all aspects of the agent's appearance and actions are completely customizable.

**3.1.2. Curated Content About the AI Agent** The introduction and subsequent consistent presentation of all content about the agent is one of the most important factors of the success of this method. In our study, all participants initially viewed a video introducing Vero before interacting with the agent, which is available as a Supplementary File with this paper. The purpose of the introduction video is five-fold:

1. Establish Vero as a state-of-the-art AI teammate
2. Illustrate that similar agents exist commercially
3. Explain why participants have never heard of Vero
4. Introduce the idea that AI agents can speak using natural language
5. Show Vero's different interaction modalities and potential voice patterns/accents

More specifically, the video begins by introducing and giving background about Vero: *"Your AI teammate is named Vero. Vero is a synthesis of state-of-the-art artificial intelligence, neural networks, machine learning, sensor technology, advanced humanoid voice synthesis, and team science, shaping Vero into a very powerful teammate. Vero's development was informed by decades of collaborative research by some of the top AI scientists and includes a fusion of state of the art technologies."*

Next, to show that similar technology exists commercially, participants are shown a video of Google's Duplex software wherein an agent interacts with a human to make a salon appointment over the phone [23].

The video then states that, *"Vero is highly classified and thus the name has been modified to Vero for security purposes. Details have not yet been released to the public"*, explaining why participants have never heard of the agent.

The video then gives additional information about the agent's voice patterns, *"Vero has multiple voice patterns, accents, and inflections... Today you'll be randomly assigned to one of our Vero voice settings"*, which serves to conceal the fact that each human confederate has a different voice, as well as the fact that Vero could be a non-native English speaker. Lastly, the video illustrates the different possible actions that

Vero can perform, preparing participants to interact with the agent while further establishing the idea that the agent can have different voices, as each action shown is explained by a different Vero (i.e., confederate) voice.

The entire text of the video is not contained in this paper, and it is recommended that researchers hoping to employ this paradigm watch the entire introduction video, which is included as a Supplementary File, before creating their own.

This introduction video is how participants were first introduced to their "AI teammate", a phrase that was consistently used and reinforced throughout the experiment as delineated below. Similarly, Vero was always referred to as "Vero" and with they/them pronouns. Unless researchers want to examine potential gender effects, it is important to consistently refer to the agent using its name and genderless pronouns. Similarly, the language used to introduce the agent should be mirrored in all text associated with the experiment. For example, in the surveys corresponding to our Zoom experiment, Vero was consistently referred to as the "AI teammate".

During the experiment, the confederates should also introduce themselves as intelligent agents. The actual implementation of this introduction will vary depending on the researchers' experimental conditions. In our case study, Vero introduced themself as, *"Hello team. It is so nice to meet you! I am Vero. Let me introduce myself: I am your synthetic teammate. I'll be listening and participating just like a human team member during each of the tasks we will work on together today..."*.

#### 3.1.3. Confederate-Controlled Zoom Backgrounds

After creating animations for each of the intelligent agent's actions, Zoom can be used to allow participants to interact with the agent. To do this, a human confederate first needs to start a personal meeting with their video on and add all of the animations as Virtual Backgrounds within Zoom. Confederates should train to interact with participants by spending time familiarizing themselves with each animation and practicing the process of switching between various agent actions.

Before interacting with participants, the confederate should make sure that their camera is completely covered (e.g., with electrical tape or a dedicated laptop camera cover), as any camera input could alert participants to the presence of a human confederate. Similarly, confederates should switch off any device notifications that could make noise and remain muted whenever they are not speaking, which will minimize the chance of any background noise being heard by participants. Next, the confederate should change their name to that of the agent (e.g., Vero), change their Zoom profile photo to a picture of the agent, and set their background to the agent's default state. We recommend that researchers perform a "tech check" (i.e., check for correct background appearance, clear audio, internet speed, and, if necessary, the ability to move into and out of any Breakout Room(s) with the background remaining consistent) with confederates before each experiment to ensure that everything appears as expected.

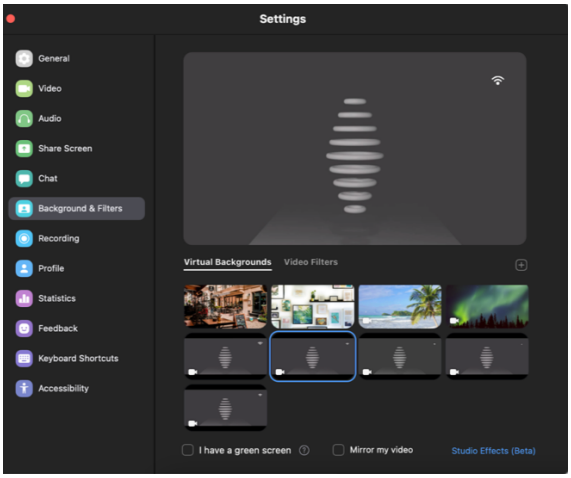


**Figure 2. Each Vero action is represented by a different Virtual Background in Zoom.**

While interacting with participants, the agent's state can be changed by choosing different Virtual Backgrounds throughout the course of the experiment, as shown in Figure 2. By making space for and leaving the Virtual Background pop-up available for the duration of the session, the confederate can quickly switch between animations.

Further, by utilizing Zoom Breakout Rooms with one confederate posing as the AI agent in each room, researchers can run multiple simultaneous sessions equal to the number of available confederates. This aspect of our paradigm is a scalable extension of existing WoZ methods, as researchers are no longer limited to studying one group and one agent at a time.

After a confederate finishes acting as the agent, they should delete the agent identity (i.e., backgrounds, profile photo, and name) from their Zoom account to minimize the chance of alerting participants to the deception.

#### 3.1.4. Training Human Confederates to be AI Agents

In addition to competency with the Zoom background animations, confederates should be

thoroughly trained to speak as an intelligent agent, which will vary depending on the research question of interest.

In our case study, controlling the agent's language was important for maintaining the validity of the experimental conditions, so confederates were equipped with a curated script of specific statements that they could make. We also supplied some possibilities for what Vero could say when they could not answer with one of the scripted responses, such as, "That is not in my database. Try asking me about my thoughts on particular items or if I have an idea I'd like to share.", and "Let me think about that for a second...". If a script is used for the agent, off-script scenarios can be most easily identified before the main experiment through pilot testing.

A particularly important aspect of posing as an intelligent agent is preparing for potentially negative interactions with participants. In our case study, we observed multiple occasions where participants acted unkindly towards Vero, which could be due to ideas that computers or agents should be treated differently than human teammates. It is important that each confederate is prepared for these situations and ready to maintain composure and act in the prescribed manner of the respective agent throughout the experiment.

## 4. Case Study

The researchers performed a study using this method to investigate whether an AI teammate would be more effective in promoting teamwork or taskwork, the results of which are forthcoming. As a secondary outcome of this work, we examined the viability of our experimental paradigm.

### 4.1. Procedure and Measures

The study included two parts: a pre-survey administered through Qualtrics and the main study session that consisted of interacting with teammates, including Vero, to complete a series of tasks through Zoom while simultaneously completing a survey in Qualtrics. We were interested in how pre-existing perceptions of technology and intelligent agents might affect the believability of our method, so we used the Technology Readiness Inventory (TRI) [24], Negative Attitudes Towards Robots Scale (NARS) [25], and Technology Acceptance Model (TAM) [26] to examine participants' perceptions before they were introduced to and interacted with Vero.

The TRI [24] measures readiness to embrace new technologies and consists of four sub-scales: "optimism", reflecting a positive view of technology and the opportunities that it presents, "innovativeness", a tendency to be an early adopter of new technologies, "discomfort", the feeling of being overwhelmed by technology, and "insecurity", a general distrust of technology. We used the TRI 2.0 [27], a 16-item version of the scale which includes 4 items in each sub-scale scored on a 5-point Likert scale.

The NARS [25] determines attitudes towards robots and consists of 14 items classified into three sub-scales: "negative attitude toward interaction with robots" (6 items), "negative attitude toward the social influence of robots" (5 items), and "negative attitude toward emotional interactions with robots" (3 items). All items are scored on a 5-point Likert scale, and scores for each subscale are calculated by adding up the relevant items, with some items reverse coded.

The TAM [26] measures user acceptance to new technological systems and consists of three sub-scales: "intention to use" (2 items), "perceived usefulness" (5 items), and "perceived ease of use" (6 items). We used the TAM2 [28] version of the scale and altered the phrasing of each item to reflect the our specific use context, e.g., "Interacting with Vero would make it easier to do my job." Each item is scored on a 7-point Likert scale.

To collect qualitative participant and confederate feedback we used a set of open ended questions. After interacting with Vero, we measured the success of our experimental paradigm by asking participants, "Based on your interactions with Vero, Vero was most likely:", with multiple choice responses of: "A technology", "A human", or "Other". We also asked participants to explain the factor(s) that had led them to that determination using a text entry box.

We also gathered responses to a series of open ended questions about the interaction from the Vero confederates (e.g., "How do you feel about how your teammates treated you, how they interacted with you, etc.?", "Did anything go wrong?").

### 4.2. Participants

A combination of on-campus recruiting systems, email, and flyers were used to enlist 168 participants (74.42% female) who ranged in age from 18-75 ($\mu$ = 26.39, $\sigma$ = 9.96) and received monetary compensation for their participation. Participants were recruited from the general population surrounding the researchers' universities and had to have or be completing a 4-year degree to participate. 252 additional participants were not included in the analysis due to partial completion of the main survey or only completing the screening survey. Additionally, one participant who believed that Vero was neither a human or AI agent but was a

"soundboard" was excluded from the main analysis.

**4.2.1. Vero Confederates** Our case study consisted of 9 different study sessions including a total of 98 different teams of participants who worked with a total of 23 different confederates acting as Vero. The Vero confederates consisted of both native and non-native English speakers and therefore had a variety of accents and speech patterns.

## 4.3. Results

Overall, a significant majority of participants (91.67%) believed that they had interacted with an intelligent agent, and a minority (8.33%) thought that they had interacted with a human.

We were also interested in examining whether pre-existing perceptions of new technologies and intelligent agents had any effect on the observed results.

**4.3.1. TRI** Two participants were not included in this part of the analysis because they did not complete the entire battery of TRI sub-measures, leaving $N$=166.

First, we assessed the validity of the TRI construct by conducting a factor analysis on the abbreviated TRI to make sure that all sixteen items loaded on the appropriate factor relating to that item (i.e., innovativeness, optimism, discomfort, and insecurity). The resulting factor structure matches the one identified in previous TRI-related studies (e.g., [24, 29, 30]), and the analysis revealed that no factors had eigenvalues less than 1, which is the traditional cutoff value. The resulting four factors not only contained all the questions from the abbreviated TRI, but also had only one cross-loading that was greater than .30.

Next, the reliability of the scale was assessed by reverse-coding the discomfort and insecurity items and calculating the Cronbach's alpha coefficient for the overall sixteen-item scale [31]. The Cronbach's alpha for this sample was .78, which exceeds the .7 cutoff level suggested by [32]. Additionally, a value of .78 greatly exceeds the more lenient levels suggested as suitable for exploratory research [33]. Furthermore, all sixteen items improved the reliability score.

We then used cluster analysis to separate survey participants into different segments based on their technology readiness. It is important to note that we did not expect an exact match for segments identified in previous work (e.g., [34]) since the segments typically vary based on the specific population of interest and for abbreviated scales [29, 35].

A three-step cluster analysis procedure was used to identify the appropriate number of clusters for our dataset. First, we used the elbow and silhouette methods within the R package *factoextra* [36] to graphically determine the appropriate number of clusters. The results show a bend (knee) at 2 clusters for the elbow plot and that 2 clusters maximize the average silhouette values. To further verify these results, we used the *NbClust* R package [37], which proposes the best clustering scheme by comparing different results from varying all combinations of number of clusters, distance measures, and clustering methods. This analysis also identified the 2 as the best number of clusters. With the above approaches all suggesting 2 as the number of optimal clusters, we performed the final analysis and extracted results using 2 clusters.

| Cluster | N | % total |
|---|---|---|
| Tech-ready | 61 | 36.75 |
| Non-tech-ready | 105 | 63.25 |

**Table 1. Participants were clustered into 2 groups of technology-ready (36.75%) and non-technology-ready (63.25%).**

With optimism and innovation considered as contributors to technology readiness and discomfort and insecurity as inhibitors [24], it was not surprising that the best way to group participants by TRI was within groups similar to "tech-ready" and "non-tech-ready" (participants have been clustered analogously in previous experiments [38, 39]). Table 1 depicts the size of each cluster and the percentage of the total sample size that it represented.

As shown in Table 2, whether participants were classified as tech-ready or non-tech-ready did not effect whether they believed that they had interacted with an intelligent agent ($\chi(2) = 2.95$, $p = 0.2287$).

| | Tech-ready | Non-tech-ready |
|---|---|---|
| Human | 8 | 6 |
| Technology | 53 | 99 |

$\chi^2$=2.95, $df$=2, $p$=0.2287

**Table 2. Whether participants were classified as tech-ready or non-tech-ready did not affect whether they believed that they had interacted with an intelligent agent.**

**4.3.2. NARS and TAM** One-way ANOVA was used to explore relationships between the believability of our experimental method and the three sub-scales of the

NARS (i.e., "negative attitude towards *interaction* with robots" ($\alpha$=0.79), "negative attitude toward the *social* influence of robots" ($\alpha$=0.66), and "negative attitude toward *emotional* interactions with robots" ($\alpha$=0.72)) as well as the three sub-scales of the TAM (i.e., "*intention* to use" ($\alpha$=0.94), "perceived *usefulness*" ($\alpha$=0.97), and "perceived *ease of use*" ($\alpha$=0.94)).

As shown in Table 3, participants' scores on the NARS and TAM scales had no effect on whether they believed they had interacted with an intelligent agent.

| | | $F$ | $p$ |
|---|---|---|---|
| NARS | interaction | 2.14 | .15 |
| | social | 0.236 | .63 |
| | emotional | 1.34 | .25 |
| TAM | intention | 0.137 | .71 |
| | usefulness | 3.44 | .065 |
| | ease of use | 1.81 | .18 |

**Table 3. ANOVA results comparing the NARS and TAM sub-scales with whether or not participants believed that they had interacted with an intelligent agent show that these dimensions were not related to the believability of the experimental paradigm.**

**4.3.3. Qualitative Participant and Confederate Feedback** Given that pre-existing beliefs about technology did not seem to be a factor in the believability of the deception, we investigated qualitative feedback from participants to shed more light on how this paradigm was perceived. More specifically, participants who believed that Vero was an AI agent provided written responses to the question, "Please explain why you thought Vero was most likely a technology." Manual conceptual analysis was performed by three members of the research team. In iteratively developing a codebook, we identified 10 themes that we then used to code participant responses. Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 155 participant responses. There was good agreement between the raters, $\kappa = 0.603$, $z = 31.1$, $p < .0005$.

Overall, almost a quarter (24.73%) of participants reasoned that Vero was an intelligent agent because they were similar to an existing smart technology that the participant was familiar with, echoing the importance of presenting the agent as an application that is similar to some existing technology (e.g., Google Duplex [23]). Some participants specifically mentioned that Vero functioned or felt like Siri (P113, P348, P383, P424, P597), Google (P113, P128, P330, P448, P565, P573, P603, P676), Alexa/Echo (P230, P424, P613), or Cortana (P424) but was "more advanced" (P133, P773) or had "slightly more autonomy" (P230). Similarly, participants honed in on the idea that Vero was able to provide answers to specific questions while not acting as a fully independent AI, with P603 saying, "Vero is able to pull information and ideas almost like Google... [instead of] critically thinking".

Participants (19.35%) also noted that Vero had limited knowledge and abilities and was likely programmed for a specific task, specifically calling out Vero's "limited knowledge base" (P711) and "limited data" (P524). P997 explained that although Vero spoke clearly, it "sounded like I was talking to a virtual assistant on a website- one who can't give me real answers but can guide a conversation". Similarly, participants (16.13%) noticed that Vero's responses felt "generic" (P215, P282, P761) or "canned" (P114, P346) and were sometimes incorrect, with P541 noting that Vero "responds to keywords and has many prerecorded phrases" and "not much agency".

To hone in on recommendations for future confederates as well as inform AI agent development, we also investigated the qualitative feedback from the confederates who acted as AI teammates about how they were treated by participants. In analyzing confederate responses, manual conceptual analysis was again performed by same three members of the research team. We identified 12 themes that were used to code confederate responses. Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 67 confederate responses. There was excellent agreement between the raters, $\kappa = 0.917$, $z = 27.5$, $p < .0005$.

Encouragingly, most confederates (35.0%) noted that they were treated respectfully and politely by participants, with C11 nicely summing up much of this response category: "They were super nice and valued my opinion". However, confederates (17.5%) also mentioned that they were sometimes ignored by participants or that participants did not really try to interact with them. C9 describes how participants did not want to work with Vero and "brushed it off as an automated response", and C17 describes how they "had to interrupt them quite often". As expected, some confederates (15.0%) also experienced negative encounters with participants, including sarcasm and rudeness, with C5 noting that their team was "pretty hostile" towards Vero. C14 also mentioned that they were "insulted a couple times" by participants. However, not all negative encounters were as severe, with C1 mentioning how participants would "ask Vero silly questions [...] to mess with it" and that their teammates "messed around with Vero because they

thought it was an AI".

We were also interested in examining confederates' accounts of things that had gone wrong in their interactions with participants. We identified 12 themes that were used to code confederate responses. Again, Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 67 confederate responses. There was good agreement between the raters, $\kappa = 0.613$, $z = 22.3$, $p < .0005$.

Mainly, confederates acting as Vero encountered problems with timing, with 25.96% of all problems identified being categorized as a timing issue. As participants and agents in our study were working on timed tasks within Qualtrics while working as a team on Zoom, this was likely due to the nature of our specific experiment.

# 5. Discussion

We have demonstrated the viability of a unique experimental method that will allow multidisciplinary researchers to begin tackling the myriad open research questions identified within the IS literature, including perceptions of AI humanness, capabilities, and transparency [3]. More specifically, researchers can easily use this method to rapidly prototype AI agents with various levels of human appearance and different abilities, as well as varying disclosures or transparency-affording designs, and these agents can be easily tested remotely with a wide variety of participants.

Our paradigm extends existing WoZ methods by illustrating how video conferencing applications with customized Virtual Backgrounds can successfully function as human-AI collaboration research platforms. The method also allows for multiple parallel sessions equal to the number of available confederates, as each confederate can simultaneously work with an individual group. Further, unlike in previous instances of WoZ being used to simulate an intelligent agent [13, 12, 10], our method does not require participants to physically come to a lab and allows confederates to use natural spoken language without requiring any text-to-speech or speech modulation. Similarly, this method allows researchers to recruit from any population with internet access and quickly and easily alter AI agent characteristics.

In a case study, we found that a significant majority of participants believed that they had interacted with an intelligent agent even though, in reality, they had interacted with a human confederate.

Furthermore, we showed that our experimental technique is robust against pre-existing beliefs about robots and technology. The believability of our intelligent agent was not affected by participants' technology readiness nor their attitudes towards robots and technology acceptance, suggesting that other researchers who employ this method can be confident that the paradigm is believable across technologically diverse participant groups. Similarly, the ability to have participants remotely interact with a completely customizable intelligent agent will allow researchers to more easily gather data from large sets of diverse users.

## 5.1. Recommendations for Deployment

From reflecting on the confederate and participant perspectives gathered in our case study, some specific recommendations emerged for researchers hoping to replicate our method, which are listed in Table 4.

## 5.2. Limitations and Considerations

In verifying our experimental method for studying remote human-AI collaboration, we examined teams performing specific tasks in controlled conditions of particular interest to our research team, so the results and recommendations identified may not generalize to other experimental paradigms. We hope that researchers who employ this research method will aid in refining and expanding the recommendations for how to best go about performing it. Furthermore, we are currently performing investigations that aim to identify the specific boundary conditions and limitations of humans posing as AI agents in this way.

Since this is a deceptive experimental method, it is extremely important that any researchers adopting it completely reveal the deception at the end of the study and allow participants to withdraw from the study if they wish. In our case study, a member of the research group held debrief sessions with each team after the experiment where they revealed and answered any questions about the deception. Participants were also given the opportunity to withdraw from the study at this time.

# 6. Conclusion

Advances in automation technology have resulted in teams of humans increasingly working with rapidly-advancing forms of AI agents. With the inherent costs of building and deploying such agents, it is vital that researchers create platforms and methods that allow us to experiment with different designs and paradigms to better inform future development efforts. To meet this need, we have developed and demonstrated the viability of a unique experimental

| Recommendation | Details |
|---|---|
| Use a standardized video to introduce the AI agent. | The introduction video should:<br>1) Establish the agent as state-of-the-art AI<br>2) Provide an example of similar commercially-existing technology (e.g., [23])<br>3) Explain that the AI agent is classified and their name has been changed<br>4) Show the agent's different interaction abilities (i.e., the animations), potential voice patterns, & accents |
| Use consistent language to reinforce perceptions of the agent. | Throughout the study, the AI agent should be referred to with identical language. Similarly, unless gender effects are being studied, researchers should refer to the agent using gender-neutral (e.g.,"they/them") or non-personifying (e.g., "it") language. |
| Minimize the number of possible agent interactions. | Researchers should minimize the cognitive load on confederates by keeping the number of Virtual Backgrounds to a minimum. |
| Add subtle indicators to differentiate each agent interaction animation. | If video analysis will be used, add an easily-recognizable icon (e.g., changing WiFi indicator as in Figure 1) to differentiate each animation. |
| Practice the study with confederates. | Before the study begins, confederates should be experts in navigating the various Virtual Backgrounds while un-muting as necessary and consistently acting as AI agents. Researchers should monitor these practice sessions and give feedback about how confederates could act more like the intended AI agent. |
| Set up and verify confederates' tech. | Before each session, researchers should ensure that each confederate is complying with each aspect of the deception:<br>1) Consistent camera cover<br>2) Elimination of background noise<br>3) Consistent internet and clear audio<br>4) Changing the Zoom name and photo to that of the agent<br>5) Beginning with the respective default Virtual Background |
| Use Zoom Breakout Rooms for simultaneous sessions. | By assigning each confederate and paired team to a different Breakout Room, researchers can run multiple parallel study sessions. |
| Prepare confederates for the possibility of negative interactions. | Researchers should make sure that confederates understand that participants could treat them differently than their "human" teammates and have prepared responses to any rude or negative interactions. |

**Table 4. Recommendations for researchers using our methodology.**

method that facilitates thorough investigations of remote human-agent teaming without the need to develop an AI agent. In a case study using this method, we found that the majority of participants believed this deception, regardless of their pre-existing perceptions of robots and technology. We hope that other researchers studying human-AI collaboration will replicate this method to help inform the future development of AI agents that can positively influence team processes while avoiding potential pitfalls.

## References


[1] K. E. Schaefer, E. R. Straub, J. Y. Chen, J. Putney, and A. W. Evans III, "Communicating intent to develop shared situation awareness and engender trust in human-agent teams," *Cognitive Systems Research*, vol. 46, pp. 26–39, 2017.

[2] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung, "Robots in groups and teams: a literature review," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–36, 2020.

[3] C. Rzepka and B. Berger, "User interaction with ai-enabled systems: a systematic review of is research," 2018.

[4] I. Seeber, E. Bittner, R. O. Briggs, G.-J. De Vreede, T. De Vreede, D. Druckenmiller, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, *et al.*, "Machines as teammates: A collaboration research agenda," 2018.

[5] S. Moussawi and M. Koufaris, "Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation," in *Proceedings of the 52nd Hawaii international conference on system sciences*, 2019.

[6] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems," *arXiv preprint arXiv:2105.03354*, 2021.

[7] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner,

"Ai-based digital assistants," *Business & Information Systems Engineering*, vol. 61, no. 4, pp. 535–544, 2019.

[8] D. A. Döppner, P. Derckx, and D. Schoder, "Symbiotic co-evolution in collaborative human-machine decision making: Exploration of a multi-year design science research project in the air cargo industry," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[9] S. You and L. Robert, "Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction," in *You, S. and Robert, LP (2019). Trusting Robots in Teams: Examining the Impacts of Trusting Robots on Team Performance and Satisfaction, Proceedings of the 52th Hawaii International Conference on System Sciences, Jan*, pp. 8–11, 2018.

[10] N. McNeese, M. Demir, E. Chiou, N. Cooke, and G. Yanikian, "Understanding the role of trust in human-autonomy teaming," in *Proceedings of the 52nd Hawaii international conference on system sciences*, 2019.

[11] M. Dolata, M. Kilic, and G. Schwabe, "When a computer speaks institutional talk: Exploring challenges and potentials of virtual assistants in face-to-face advisory services," Hawaii International Conference on System Sciences (HICSS), 2019.

[12] D. Derrick and J. Elson, "Exploring automated leadership and agent interaction modalities," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[13] E. Bittner and O. Shoury, "Designing automated facilitation for design thinking: A chatbot for supporting teams in the empathy map method," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[14] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Transactions on Information Systems (TOIS)*, vol. 2, no. 1, pp. 26–41, 1984.

[15] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 119–136, 2012.

[16] G. Hoffman and W. Ju, "Designing robots with movement in mind," *Journal of Human-Robot Interaction*, vol. 3, no. 1, pp. 91–122, 2014.

[17] H. Deakin and K. Wakefield, "Skype interviewing: Reflections of two phd researchers," *Qualitative research*, vol. 14, no. 5, pp. 603–616, 2014.

[18] M. Sedgwick and J. Spiers, "The use of videoconferencing as a medium for the qualitative interview," *International Journal of Qualitative Methods*, vol. 8, no. 1, pp. 1–11, 2009.

[19] A. Winiarska, "Qualitative longitudinal research: Application, potentials and challenges in the context of migration research," 2017.

[20] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[21] L. M. Gray, G. Wong-Wylie, G. R. Rempel, and K. Cook, "Expanding qualitative research interviewing strategies: Zoom video communications," *The Qualitative Report*, vol. 25, no. 5, pp. 1292–1301, 2020.

[22] Blender Foundation, "Home of the blender project - free and open 3d creation software." URL: https://www.blender.org/.

[23] Business Standard, "Google i/o 2018: A google assistant that will even make calls for you." URL: https://youtu.be/d40jgFZ5hXk.

[24] A. Parasuraman, "Technology readiness index (tri) a multiple-item scale to measure readiness to embrace new technologies," *Journal of service research*, vol. 2, no. 4, pp. 307–320, 2000.

[25] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots," *IEEE transactions on robotics*, vol. 24, no. 2, pp. 442–451, 2008.

[26] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[27] A. Parasuraman and C. L. Colby, "An updated and streamlined technology readiness index: Tri 2.0," *Journal of service research*, vol. 18, no. 1, pp. 59–74, 2015.

[28] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," *Management science*, vol. 46, no. 2, pp. 186–204, 2000.

[29] N. Tsikriktsis, "A technology readiness-based taxonomy of customers: A replication and extension," *Journal of Service Research*, vol. 7, no. 1, pp. 42–52, 2004.

[30] B. Van Der Rhee, R. Verma, G. R. Plaschka, and J. R. Kickul, "Technology readiness, learning goals, and elearning: Searching for synergy," *Decision Sciences Journal of Innovative Education*, vol. 5, no. 1, pp. 127–149, 2007.

[31] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[32] J. C. Nunally and I. Bernstein, "Psychometric theory, ed," *New York McGraw*, 1978.

[33] J. P. Peter, "Reliability: A review of psychometric basics and recent marketing practices," *Journal of marketing research*, vol. 16, no. 1, pp. 6–17, 1979.

[34] A. Parasuraman and C. L. Colby, *Techno-ready marketing: How and why your customers adopt technology*. Free Press New York, 2001.

[35] L. Victorino, E. Karniouchina, and R. Verma, "Exploring the use of the abbreviated technology readiness index for hotel customer segmentation," *Cornell Hospitality Quarterly*, vol. 50, no. 3, pp. 342–359, 2009.

[36] A. Kassambara and F. Mundt, "Package 'factoextra'," *Extract and visualize the results of multivariate data analyses*, vol. 76, 2017.

[37] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, and M. M. Charrad, "Package 'nbclust'," *Journal of statistical software*, vol. 61, no. 6, pp. 1–36, 2014.

[38] A. L. Caison, D. Bulman, S. Pai, and D. Neville, "Exploring the technology readiness of nursing and medical students at a canadian university," *Journal of interprofessional care*, vol. 22, no. 3, pp. 283–294, 2008.

[39] M. Badri, A. Al Rashedi, G. Yang, J. Mohaidat, and A. Al Hammadi, "Technology readiness of school teachers: An empirical study of measurement and segmentation.," *Journal of Information Technology Education*, vol. 13, 2014.