Dialogues For One: Single-User Content Creation Using Immersive Record and Replay

Klara Brandstätter
k.brandstatter@ucl.ac.uk
University College London
United Kingdom

Anthony Steed
a.steed@ucl.ac.uk
University College London
United Kingdom

Figure 1: Left: Additions to the existing record and replay tool; audio indicators display each recorded avatar's audio waveform. The yellow highlights indicate when the avatars are grabbing the object. Right: One actor performing a dialogue with their previously recorded self in a virtual kitchen environment.

ABSTRACT

Non-player characters are an essential element of many 3D and virtual reality experiences. They can make the experiences feel more lively and populated. Animation for non-player characters is often motion-captured using expensive hardware and the post-processing steps are time-consuming, especially when capturing multiple people at once. Using record and replay techniques in virtual reality can offer cheaper and easier ways of motion capture since the user is already tracked. We use immersive record and replay to enable a single user to create stacked recordings of themselves. We provide tools to help the user interact with their previous recorded self and in doing so allow them to create believable interactive scenarios with multiple characters that can be used to populate virtual environments. We create a small dialogue dataset with two amateur actors who used our tool to record dialogues alone and together in virtual reality. To evaluate whether stacked recordings are qualitatively comparable to conventional multi-user recordings and whether people could tell the difference between the two, we conducted two user studies, one online and one in virtual reality with 89 participants in total. We found that participants could not tell the difference and even slightly preferred stacked recordings.

VRST 2023, October 09–11, 2023, Christchurch, New Zealand

ACM ISBN 979-8-4007-0328-7/23/10.
https://doi.org/10.1145/3611659.3615695

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **User studies**; **Interaction techniques**.

KEYWORDS

virtual reality, record and replay, content creation

ACM Reference Format:
Klara Brandstätter and Anthony Steed. 2023. Dialogues For One: Single-User Content Creation Using Immersive Record and Replay. In *29th ACM Symposium on Virtual Reality Software and Technology (VRST 2023), October 09–11, 2023, Christchurch, New Zealand.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3611659.3615695

1 INTRODUCTION

Background characters or non-player characters (NPCs) are an integral part of many 3D and virtual reality (VR) applications. In games, NPCs populate the virtual world and make it appear more lively. They can give the players a sense of purpose, be it as villains players have to defeat, supporters for the players along their journey, or just fellow participants in the virtual world.

Liveliness also plays an important role in social virtual reality (SVR) experiences. SVR encompasses all applications where users can come together in a virtual world, hang out, chat, or play games with each other. A number of SVR experiences have emerged in the past years, including Rec Room [22], VRChat [23], Horizon Worlds [37], or Roblox [8]. In SVR, NPCs can make the virtual environment appear more populated, especially when fewer users are online, and could make it possibly more attractive to users [3]. However, creating NPCs that behave naturally can be a painstaking and expensive process. Animation for NPCs is often motion-captured using optical marker-based pose tracking and thus the movements

are inherently natural and human. However the post-processing necessary to clean up the data and resolve marker occlusions can be time-consuming and increases when dealing with capturing multiple people at once [18, 28].

VR devices already have some motion-capture capabilities. Their ease of use can potentially make the process of recording motion data, either of a single person or multiple people, significantly easier. Additionally, motion-capturing can already be done in the designated virtual environment together with all the necessary props. The broad concept of recording and replaying users or even whole virtual environments in VR for content creation has been explored in a number of previous works [14, 55, 57].

We use stacked recordings (recordings of replays) in VR to enable a single user to create content with multiple characters. This can be used to populate virtual environments with NPCs that move and behave naturally in a given environment. A drawback of single-user stacked recordings is that early recordings normally cannot interact with later recordings. Our contributions are presented as follows:

- First, we improve an existing record and replay tool to give a single user the ability to interact with their previous replays by passing objects between themselves and by recording conversations. With these *interactive* stacked recordings, single users can interact with themselves as they would with another person which creates a believable illusion of multiple people (see Section 3).
- Second, we create a dialogue dataset of 20 dialogues with two amateur actors who were recording dialogues together using the normal record and replay tool and on their own using stacked recordings which leads to a total of 80 recordings. We also modify the actors' voices using an online AI-based voice transformer. We want to know if single-user stacked recordings are qualitatively comparable to regular recordings with two users and if people can tell the difference between a single-user recording and a multi-user recording. We intentionally use simple cartoony avatars with head, hands and torso since we record our dataset in VR with only head and hand tracking and more complete avatars would not add any extra information for people. Besides, simple avatars like this are still widely used in SVR experiences (e.g., Rec Room).
- Third, we run two user studies, one online and one in VR with a total of 89 participants (see Section 4).

Results show that participants cannot tell the difference between single-user and multi-user recordings and even prefer single-user recordings (see Section 5). Finally, we elaborate on our study results in Section 6. The code and the materials for the studies, including the recordings, are available at https://doi.org/10.5522/04/23947278.v1.

2 RELATED WORK

2.1 Record and Replay

The concept of recording and replaying sessions in VR has been around for over 20 years. Greenhalgh et al. [14] separated their virtual environment into linked local areas called 'locales' and introduced temporal links that would connect current locales with recordings of other locales. They used temporal links for content creation in VR, to show flashbacks within a story between virtual characters, and for virtual messaging.

In more recent years, record and replay has also become a helpful tool for analysing VR experiences. Lopez et al. [33] developed a Unity plugin to record and play back user movements and the user's field of view to gain a better understanding of user experience in VR. Howie and Gilardi [19] also record user input information in addition to tracking data. Steed et al. [50] used a record and replay module in Unity that logs user and application behaviour for diagnostic purposes and content creation. Lilija et al. [31] introduced spatial recordings that enabled users to navigate through recordings by manipulating objects to see when the state of the object changed. Hubenschmid et al. [20] proposed the mixed-immersion tool *ReLive* with which users can analyse mixed reality studies by reliving them in-situ, while also offering a synchronised desktop view for ex-situ analysis of more complex data.

Recording and replaying VR sessions has also become increasingly important for virtual training applications as replaying offers new perspectives on recorded actions and can provide a better understanding of the tasks at hand. Kloiber et al. [29] developed an immersive motion analysis system for visualising human motion paths from multiple recorded VR sessions for assessing performance improvements in training applications. Similar to Kloiber et al. [29], Kamarianakis et al. [27] also use VR recording and replaying for training purposes, in particular, surgical training. Maria et al. [35] focus on post-operative debriefing for surgical training and use record and replay not only to let surgeons rewatch their own and other surgeons' actions but also to let them relive surgeries from the perspective of other surgeons to foster empathy and other non-technical skills. Xu et al. [56] replay football games in VR generated with data from real football matches, such as player position and speed, and a motion-capture dataset from actual football players. Virtual characters are then animated from the data using an AI-based motion controller, and provide football players and coaches with a better way of analysing their past matches. Mahadevan et al. [34] developed the world-in-miniature VR system *Tesseract*, that allows designers to search through miniature spatial design recordings and rewatch them life-sized. It supports designers in understanding new workflows and design processes.

Record and replay can be used for sharing information or experiences with others. With *ReliveReality*, Wang et al. [54] give users the opportunity to share real-world experiences from different perspectives in a multi-user environment in VR. They use a single RGB camera and deep-learning-based computer vision techniques to reconstruct 3D people and environments. Wang et al.[55] also created another tool, *ReliveInVR*, that would let users share their own VR experiences socially with others in the same virtual environment. Fender and Holz's [10] *AsyncReality* system captures physical events in real-time and conceals them from the user who is immersed in a virtual workspace. Later, the user can revisit these events causally correctly to update their reality with what has been happening around them while they were working.

Further, recording and replaying in VR can facilitate motion-capture and character animation tasks since no other motion-capture systems are necessary and the recorded animations can be played back immediately in the virtual environment. It also simplifies the capturing of multiple people which can be difficult for conventional

optical motion-capture systems due to marker occlusions. Gorisse et al. [13] developed a tool for Unity that enables virtual reality users to record their movements, export skeletal data, and replay the data on virtual characters. Steed et al. [49] integrated a record and replay tool into their Unity networking library Ubiq to easily populate virtual environments with characters, gather motion capture data of multiple people, and analyse user behaviour. Yin et al. [57] use record and replay in a similar way. They introduced the 'one-man-crowd paradigm' and enable a single user to create crowd motion step-by-step by recording themselves together with previous replays. Hofmann et al. [17] present a motion data pipeline to facilitate and automate VR development. Their pipeline includes a recording stage to capture motion data and a review stage to replay and verify the captured data.

Several companies have developed record and replay tools or integrated them into their products for analysis, content creation, or training purposes. For instance, FlipsideXR [11] enables users to record their animated shows and performances in VR with a wide range of virtual characters and lets users live-stream the content on social media. Mindshow [21] offered a very similar product but it is now discontinued. Subsequently, they focused on creating high-quality animation and content in VR for TV productions. RecordXR [36] is a commercial tool that allows users, especially educators, to import their data and record their interactions and explanations with it in VR. They can replay the recorded content and interact with the data and share it with other users and students. The tool is integrated into Medicalholodeck, a VR surgical training and medical education platform. With Wist [24], users can take a video on their phone which then gets turned into an immersive 3D memory that people can experience in VR. The web framework A-FRAME [40] supports recording and replaying of motion-capture data from VR devices to improve VR development, and NVIDIA [9] offers a Virtual Reality Capture and Replay tool for performance testing and debugging.

Record and replay has also made its way into the VR gaming industry. The art and game studio Tender Claws [5] used record and replay elements in their immersive theatre VR game 'The Under Presents' where players see past versions of themselves and have to work together to solve small puzzles. The VR game 'The Last Clockwinder' [42] by the studio Pontoco based its entire gameplay on record and replay. Players have to create their own looping automatons to build contraptions.

2.2 Virtual Self-Conversations

Our work, in particular the way in which we ask users to dialogue with themselves, is related to the technique of virtual self-conversations. Osimo et al. [41] and Slater et al. [48] used self-conversations for VR counselling. Participants alternately embodied themselves and a virtual therapist and would talk about their problems while also giving advice to their virtual selves as the therapist. A similar approach was used by Anastasiadou et al. [1] to help participants change to a healthier lifestyle. These studies explore counselling and body ownership. The user alternates between the virtual characters, immediately responding to the previously embodied character, but the applications do not support the recording and replaying of one continuous, seamless conversation.

2.3 Acting in VR

VR has long been used to support acting. Slater et al. [47] showed that rehearsing performances in VR with two remote actors and a producer is a good basis for successful live performances. Later, Steptoe et al. [51] developed a multimodal mixed reality environment for two remote actors and a remote producer where one actor was located in a meeting room, seeing the other actor on a projection screen, while the other actor was immersed in VR. The producer was observing the performance from a CAVE-like system. Batras et al. [2] presented a VR platform for improvising non-verbal dialogues between a real actor and a virtual agent. Kammerlander et al. [28] use VR to enable the acting of differently-scaled characters. Sakuma et al. [45] tried to inspire empathy in users by letting them role-play avatars with different personalities and backgrounds. In 2021 and 2022, the art and games studio Tender Claws [6] ran an interactive Shakespeare-inspired VR show with live actors who included the audience in the story.

While the concepts of acting as well as recording and replaying in VR are not new, we introduce interactivity in the form of single-user object interaction and dialogues. These open new possibilities for a wide range of content that can be created by just a single person. There has also been little work on comparing single-user versus multi-user recordings in content creation.

3 INTERACTIVE RECORD AND REPLAY TOOL

We extended our record and replay tool [49] for the Unity networking library Ubiq [12] that enabled users to record multi-user sessions in VR and perform networked replays for remote users. Stacked recordings could be created on top of existing replays, providing an easy way to populate virtual environments with multiple characters. We improved the existing record and replay tool to facilitate the creation of interactive scenarios for a single user.

A drawback of the virtual characters created by a single user with the record and replay tool was that early recorded characters could not react to later recorded characters since they did not exist yet. Recording simulated conversations with a previously recorded avatar was difficult since it was hard to estimate the length of the pauses between spoken sentences of the first recorded avatar and remember longer conversations. Naturally, passing objects between recorded avatars was not doable since the first recorded avatar would not have a counterpart to pass the object to.

3.1 Single-User Dialogues

To give the impression that two recorded avatars could have normal conversations with each other we implemented supporting tools that would make it easier for a single user to fabricate conversations. We implemented *audio indicators* that would hover above the head of recorded avatars and always face the user (see Figure 1 left). The audio indicator displayed the audio waveform of an avatar for the whole recording. A moving red cursor indicated the current progress of an ongoing recording. With the help of the audio indicator, the user could anticipate when a previous avatar would speak and prepare responses accordingly. Additionally, we added a portable semi-transparent *script canvas* that the user could place freely within the virtual environment. The script canvas contained

pre-scripted conversations the user could perform with their recordings. The user could place the semi-transparent canvas in front of them while recording and reading a conversation and would still be able to see the gestures of previously recorded avatars through the canvas. The canvas was scrollable using a button on the controller. This way, the user could scroll through the conversation while staying in character.

3.2 Single-User Object Interaction

To enable a user to pass objects between themselves and their recorded avatar, we implemented *recordable interactable objects* (RIOs). An RIO is an object that does not get created during a replay but remains in the environment for the recorded avatars to interact with. During a recording, only the initial position of the RIO and interactions with it were recorded. For example, in the first recording, the user could grab the RIO and pretend to hand it to another avatar. They could also pretend to take the RIO back from the other avatar. Once the RIO was released by the user, physics took over and it would fall to the ground, but this was not recorded. In the second recording, the user played the second avatar who took the RIO from the first avatar before it could fall to the ground. The user knew that the first avatar expected the RIO back and could hand it back when the first avatar reached for the RIO again. The RIO automatically attached to the first avatar's hand when the first avatar tried to grab it once it got released by the second avatar. We additionally used the audio indicators to highlight whenever an avatar was grabbing the RIO to make it easier for the user to time the grabbing and releasing of the RIO (see Figure 1). The workflow for creating single-user dialogues and object interaction is showcased in an accompanying video.

4 USER STUDIES

To determine whether single-user recordings could be indistinguishable from multi-user recordings, we conducted two user studies where participants had to watch and rate recordings made by one or two actors. In our studies, we only focused on comparing single-user with multi-user dialogues and did not include object interaction. The first study was conducted online and the second study was conducted in person in VR. We refer to them as *Web* study and *VR* study. In the following, we describe the process of the dialogue dataset creation for the studies and the design of the studies themselves.

4.1 Dialogue Dataset Creation

We recruited two female non-professional actors who recorded 20 short dialogues together and separately in VR on a Quest 2 and an Oculus Rift using the record and replay tool. The 20 dialogues were taken from Amazon's Commonsense-Dialogues dataset [58] which consists of roughly 11,000 dialogues. Each dialogue was written by Amazon Mechanical Turkers based on a given social context and had 4-6 turns between two people. We sometimes changed pronouns in the dialogues to match the female actors. A dialogue example is shown below:

> *Context: Sydney met Carson's mother for the first time last week. He (changed by the authors to 'She') liked her.*
>
> *"I met Carson's mother last week for the first time."*
> *"How was she?"*
> *"She turned out to be really nice. I like her."*
> *"That's good to hear."*
> *"It is, especially since Carson and I are getting serious."*
> *"Well, at least you'll like your in-law if you guys get married."*

The dialogues were recorded remotely by the actors. We sent them the record and replay application and gave them instructions about the recording procedure over voice chat. We also provided them with an instructional video that demonstrated the workflow of recording and replaying in VR. The actors did not know the purpose of the study. They were only told to record the given dialogues together and separately in VR and that they should try to be expressive because the avatars are very simple and would benefit from expressivity. They were also told to speak very clearly as we would need to transform their voices later on. The record and replay tool only allows one user at a time to record and replay the scene. Hence, the recording authority was given to the actor who would begin the dialogue, so they could start recording.

In order to create a well-balanced dialogue dataset and to avoid it being biased towards either single-user recordings, multi-user recordings, or one of the actors, the actors had to follow a specific protocol during the recording sessions. The recordings took place in a virtual kitchen environment (see Figure 1). As the dialogues could be from everyday life situations, we considered a kitchen a good setting. We placed a carpet in the middle of the scene and told the actors to always stand on the carpet when performing. Additionally, we rendered the whole kitchen scene including the actors from an in-game camera on a canvas that was positioned outside the kitchen where the camera was. The canvas showed the actors how participants would later see the scene in the videos that were created for the Web study and should help the actors to position themselves in the center of the scene. This way, we could make sure that the actors would always stand at similar distances away from the camera and from each other and that the videos we created from the stacked recordings would look alike.

For the single-user recordings, the actors always had to perform the lines of the first character in the dialogue before replaying and recording the lines of the second character on top of the first recording. Because the actors had to switch positions each time they recorded a new character in the single-user recording, we also wanted to make the actors switch positions after each dialogue they recorded together. This would prevent the actors from standing in the exact same position while performing the dialogues and would provide as much variability in positioning for the multi-user recordings as for the single-user recordings.

Because the actors were performing both parts of the dialogues during the single-user recordings, we had to make sure that they would also perform both parts during the multi-user recordings. This led to the following recording procedure: In the first multi-user round, Actor 1 was responsible for recording the first 10 dialogues, switching positions with the other actor after each dialogue. Actor 2 was responsible for recording the last 10 dialogues. This gave them some time to familiarise themselves with the record and replay tool

and with the dialogues. After that, each actor had to do the single-user recordings separately. In the second multi-user round, Actor 2 recorded the first 10 dialogues and Actor 1 the last 10 dialogues. The actor who spoke first in the dialogue was always responsible for the recording. The recording procedure is outlined in the supplemental material in Table A.1.

We made sure that each actor spoke each line (and each character) once in both single- and multi-user sessions and was positioned equally on the left and right side in the kitchen environment. This resulted in two sets of multi-user recordings, R1 and R2, and two sets of single-user recordings, A1 and A2.

The actors could record the dialogues at their own pace without the authors present, and they could rerecord dialogues as often as they wanted until they thought it was good enough. The first multi-user round took them roughly 2 hours. This was partly because they had to rehearse each dialogue a few times and also rerecord some of them because they made mistakes. They also occasionally experienced audio issues and could not talk to each other anymore; reconnecting to the application fixed the problem. The single-user recordings took them about 1 hour, and the second multi-user round took them 1.5 hours.

We also asked the actors what strategies they used for estimating the pauses between lines during the recording of the first character in the single-user recordings. Actor 1 said that they were reading the lines of the second character in their head and then added two more breaths to it. Actor 2 was reading the lines in their head slightly slower than they normally would. This usually gave them enough time to fill in the second character's lines in the subsequent recording.

4.2 Post-Processing of the Recordings

In order to anonymise the recordings and to hide the fact that the characters in the single-user recordings had the same voice, we used the online AI voice transformer *Koe: Recast* [44] to alter the voices in all recordings, including the multi-user ones. Raw audio data for each recording was saved in separate files from the motion data. We converted each actor's audio data to Waveform files which served as input for the AI voice transformer. The transformed audio file was then converted back into raw audio data and was used in the recording instead of the original audio data. We selected three different female AI voices and three different female avatar textures corresponding to the voices. We selected the AI voices based on how good and natural they sounded.

The audio that was recorded through the record and replay tool was often very noisy and the voice transformer was not able to transform the audio into something understandable. Normalising the audio and removing some of the noise helped to improve the quality of the transformed audio. Actor 1 had a German accent which did not work well for some voices. Actor 2 sometimes spoke too fast for the audio transformer. We ignored these dialogues and selected 6 dialogues that had the most understandable audio transformation for all sets R1, R2, A1, and A2 for the study.

We used the record and replay tool to create videos for the Web study from the recorded motion data and the transformed audio. For the VR study, we simply replayed the recordings with the record and replay tool. Because the dialogues were recorded remotely by the actors, we naturally had some latency in the recordings and the local actor, who was recording the dialogues had a higher framerate than the remote actor. Latency was an issue that had nothing to do with the actor's performance, but it impacted the recordings and made it obvious who was who since the remote actor's movements were less smooth. Therefore, we linearly interpolated between successive frames captured in the recording to smooth the motion output.

4.3 Study Design

For the Web study, we designed a Unity WebGL application (Unity version 2020.3) and hosted it on Netlify [39]. Participants were recruited on Prolific [43] and compensated with £6 (GBP). The VR study was run in person on a Pico 4 and we compensated participants with £8. In order to start with the study, participants had to give their consent for data collection. No personally identifying information was gathered. The studies were approved by the UCL Research Ethics Committee.

We used a within-subjects design. Every participant watched the same recordings. The study was divided into two rounds, a *Preference Round*, and a *Detection Round*. In each round, participants had to watch 24 short recordings (15-20 seconds each) of 6 dialogues. Each dialogue was recorded twice with two actors (R1, R2) and twice with one actor (A1, A2), hence 4 recordings per dialogue. In the Preference Round, we showed them 12 sets of 2 recordings. Each set consisted of one multi-user and one single-user recording of the same dialogue. We used a Balanced Latin square design to determine the order of single- and multi-user recordings within the sets. After each set, participants had to select which of the two recordings they preferred. Participants were not yet told about the single-user recordings. After the Preference Round, we asked participants: *1.) What were your deciding factors when you chose one recording over the other? 2.) Why did you prefer one recording over the other?* In the Web study, the answers had to be at least 60 characters long in order to proceed with the study. This concluded the first round and we revealed to participants that some of the recordings were only recorded by a single user playing both characters.

In the Detection Round, we showed them the same 24 recordings again, this time one by one and in a different random order. The recording order for the Preference and Detection Round is shown in the supplemental material in Table A.2. After each recording, participants had to select whether the dialogue was performed by one or two actors. At the end of the Detection Round, participants were asked again: *1.) What were your deciding factors when you selected the number of actors? 2.) What did you look for in the recordings to figure out if they were recorded by one person?* At the end of the study, we asked participants about their gender, age, VR experience, and how much they play video games. We also wanted participants to guess the purpose of the study and they could add optional feedback. Participant responses were uploaded to our own internal data collection server at the end of the study.

For the VR study, participants were immersed in the kitchen environment and watched the recordings of the Preference and Detection Rounds as a bystander in the same pseudo-random order. To make the whole study more interactive, participants could press physical buttons in front of them to answer the questions.

After each round, participants had to take off the headset to answer the text questions on a PC to spare them the trouble of

having to type in VR. The study procedure was the same as in the Web study with the exception that their written answers did not need to be at least 60 characters long. We implemented this in the Web study to prevent participants from clicking through the questions without answering them.

Additionally, we added an extra between-subjects *distance* condition to the VR study that was based on proxemics [15]. Participants would watch the recorded avatars either from 1 meter (personal distance), 2 meters (social distance), or 4 meters (public distance) away. We wanted to see if the distance to the avatars might affect participants' detection accuracy. On top of that, participants had to answer at the end of each round on a 7-point Likert scale (from 'strongly disagree' to 'strongly agree'): *1.) I felt that the avatars were ignoring me. 2.) The avatars were not polite.* We hypothesised that participants might feel ignored by the avatars and perceive the avatars as less polite when being at a personal distance. At this distance, participants might feel part of the conversation from which they would be obviously excluded by the avatars. Participants who would watch the recordings from further away (social or public distance), might feel more like observers and therefore feel less ignored by the avatars and perceive them as more polite.

5 RESULTS WEB AND VR STUDY

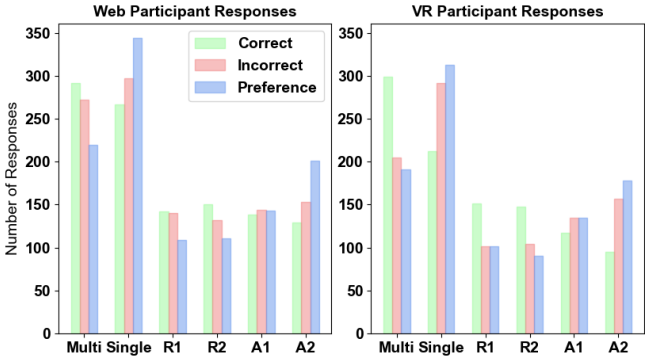

Figure 2: Overview of results from Preference and Detection Round for Web and VR study visualised for single- and multi-user recordings separately. Additionally, single- and multi-user recordings are broken down into R1 and R2, and A1 and A2, respectively.

For the **Web** study, We recruited 50 participants on Prolific (13 female) who were proficient in English. We excluded 3 participants from the evaluation because they gave the same answers for all recordings in the Detection Round. We analysed the results from the remaining $n = 47$ participants (12 female). Participants' age (mean + standard deviation) was 32 ± 11.5. The average time they needed to complete the study was 26.6 ± 5.1 mins.

For the **VR** study, we recruited 44 participants from which we had to exclude 2 participants because they gave the same answers for all recordings in the Detection Round. The majority of participants were undergraduate students from UCL. Of the $n = 42$ participants, 25 were male and 17 were female. Participants' average age was 21-25 years and the average completion time was 24.9 ± 2.6 mins.

Table 1: Normalized overall performance of correctly and incorrectly rated recordings for Web and VR study.

		Actual Number of Actors			
		Web		VR	
		1 actor	2 actors	1 actor	2 actors
Participant Response	1 actor	**0.47**	**0.48**	**0.42**	**0.41**
	2 actors	**0.53**	**0.52**	**0.58**	**0.59**

5.1 Overview

In the Web and VR study (47 and 42 participants), 24 recordings were rated which is 1,128 responses for the Web study and 1,008 responses for the VR study in total. In the Preference Round, they had to select from sets of single- and multi-user recordings the one they preferred. In the Detection Round, they had to select for each recording whether one or two actors recorded it. Figure 2 gives an overview of the preference results of the Preference Round and the detection results of the Detection Round. The overall preference (blue) for single-user recordings was higher than for multi-user recordings in both studies. There was an overall preference for the recordings of Actor 2. Overall, participants guessed more multi-user recordings correctly than single-user recordings. Table 1 shows the overall normalized performance of participants in the Detection Round. In total, 559 (Web) and 511 (VR) recordings were rated correctly, and 564 (Web) and 497 (VR) recordings were rated incorrectly. In the VR study, overall detection accuracy was slightly higher with 0.59 for the multi-user recordings than in the Web study (0.52) and slightly worse for single-user recordings with 0.42 (0.47 in the Web study).

5.2 Analysis Preference Round

We wanted to know if there was a significant difference in preference for single-user recordings compared to multi-user recordings. We compared participants' preference scores for the single- and multi-user recordings. We also compared preferences for Actor 1 and Actor 2. If not otherwise stated, we used Shapiro-Wilk's test and examinations of Q-Q plots to determine normality. For normally distributed data, we performed paired samples t-tests. For the sake of correctness, we performed non-parametric Wilcoxon signed-rank tests on data that failed Shapiro-Wilk's test despite being very close to normally distributed. We also kept any outliers in the data as they never changed the significance of the analysis. We provide additional results of t-tests that confirm the results of the non-parametric tests in the supplemental material as t-tests are fairly robust to violations of normality [52] (Section A.2).

5.2.1 Single-User vs. Multi-User Recordings. The differences between single-user preferences and multi-user preferences were normally distributed ($p = .108$ for Web and $p = .143$ for VR). In the Web study, participants significantly preferred the single-user recordings in 7.32 ± 1.656 sets (mean ± standard deviation) of the 12 recordings sets over the multi-user recordings (4.68 ± 1.656 sets), $t(46) = 5.46, p < .001, d = .796$. These results are comparable with the results from the VR study where the single-user recordings were preferred in 7.45 ± 1.797 sets of the 12 recording sets over multi-user

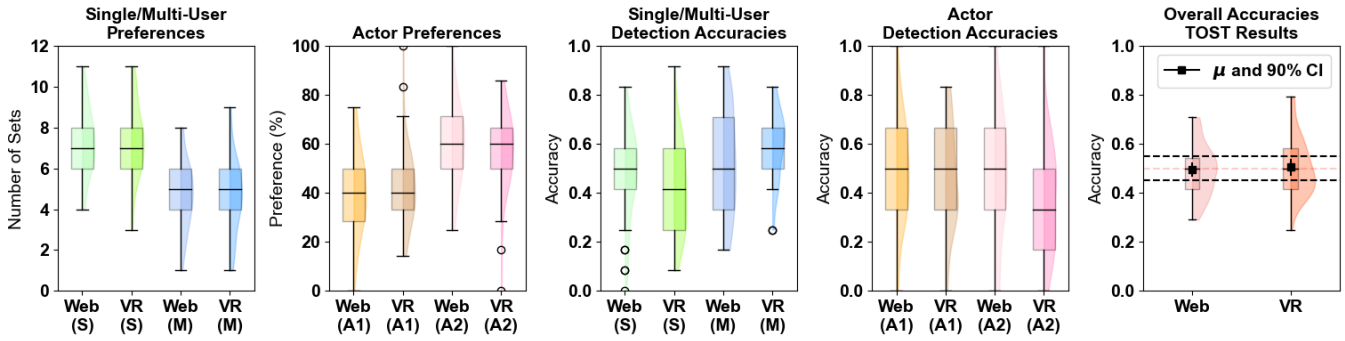

Figure 3: Preference results for Web and VR study, showing the number of sets in which participants preferred either the single- (S) or multi-user (M) recording and the normalized preferences for Actor 1 (A1) and Actor 2 (A2). Detection accuracies for Web and VR study, showing accuracies for single- and multi-user recordings and for recordings of Actor 1 and Actor 2. And finally, overall detection accuracies and equivalence margins of the TOST procedure. The means and the 90% CIs lie within the equivalence margin [.45, .55], therefore, we can assume that detection accuracies are equivalent to guessing average (dotted red line).

recordings (4.55 ± 1.797 sets), $t(41) = 5.238, p < .001, d = .808$. The distributions of single- and multi-user preferences are shown in Figure 3 on the left.

5.2.2 Actor 1 vs. Actor 2. Given the overall higher preference for the recordings of Actor 2, we also determined whether there was a significant preference for Actor 2. Because participants' preference scores for single-user recordings varied, we normalised the scores for how often they selected recordings of Actor 1 or Actor 2. In the Web study, the data were normally distributed ($p = .743$). Among the single-user preferences, participants significantly preferred Actor 2's recordings $59.3 \pm 16.6\%$ of the time over Actor 1's recordings ($40.7 \pm 16.6\%$), $t(46) = 3.849, p < .001, d = .561$. In the VR study, the difference scores were not normally distributed ($p = .021$). There was a statistically significant median increase of 2 recordings for Actor 2 (6 recordings) compared to Actor 1 (4 recordings), $z = 2.896, p = .004$ (Wilcoxon signed-rank test). The distributions of the actor preferences are shown in Figure 3.

5.3 Analysis Detection Round

In the Detection Round, we analysed whether participants considered single-user recordings equal to multi-user recordings. We defined participant accuracy as the percentage with which participants correctly identified the number of actors in the recordings. We specified that if participant accuracy for detecting the number of actors would be between 45% and 55% and thus close enough to the accuracy of random guessing (50%), we could infer that participants cannot distinguish between single-user recordings and multi-user recordings. We conducted two one-sided t-tests (TOST) and Bayes factor (BF) equivalence tests to determine whether our effect sizes δ lie within our predefined equivalence margin ($[0.45, 0.55]$ in raw units which corresponds to $[-0.42, 0.5]$ in standardised units using Cohen's d [7]) and can therefore be considered equivalent to random guessing [30, 32]. Detailed explanations about the TOST and BF procedures are given in the supplemental material (Section A.1).

5.3.1 TOST Equivalence Test. We ran a null hypothesis significance test (NHST) and a one-sample TOST equivalence test in SPSS (v29). We tested participant accuracies for equivalence within the raw equivalence bounds [0.45, 0.55] against the random guessing average of 0.5. The participant accuracies in the Web study were normally distributed ($p = .134$). The null hypothesis of the NHST was $H_0 : \mu = 0.5$. Participant accuracies were not significantly lower by .004 (95% CI, −.036 to .027) than the guessing average, $t(46) = -.280$, $p = .781$, Cohen's $d = .041$ (95% CI, −.245 to .327). The null hypotheses for the TOST procedure were $\mu \leq 0.45$ for the lower bound and $\mu \geq 0.55$ for the upper bound. Participant accuracies were significantly higher by .046 than the lower bound, $t(46) = 2.877$, $p = .003$, Cohen's $d = .420$ (95% CI, .119 to .716), and significantly lower by .054 than the upper bound, $t(46) = -3.436$, $p < .001$, Cohen's $d = .501$ (95% CI, .195 to .802). The participant accuracies in the VR study were not normally distributed ($p = .033$). We ran a Wilcoxon signed rank TOST equivalence test with continuity correction in R (v.4.2.2) using the TOSTER package [4]. The null hypothesis test ($H_0 : \mu = 0.5$ is the median in the non-parametric test) was non-significant, $V = 353, p = .759$. The equivalence test was significant with $V = 667, p = .004$ for the lower bound (0.45) and $V = 247, p = .005$ for the upper bound (0.55). The V-statistic is the sum of ranks assigned to positive differences. The bigger the V-value, the more different the medians and the smaller the p-value. The 90% CI was $[0.458, 0.542]$, which lies within our defined equivalence bounds. The result of the TOST procedure with equivalence margins and 90% CI is visualized in Figure 3 on the right. The CI for TOST is only 90% instead of the conventional 95%. This is because TOST computes the two one-sided tests for the upper and lower equivalence margin separately, and each of the two t-tests uses $\alpha = 0.05$ and a 95% CI. Therefore, the combined equivalence test has $\alpha = 0.1$ and a CI of 90%.

5.3.2 BF Equivalence Test. We further used the open-source software JASP to compute the BF for a one-sample equivalence test [25] using a standardized equivalence margin $[-0.42, 0.5]$ based on Cohen's d [7]. Our hypotheses are $H_0 : \delta \notin [-0.42, 0.5]$ and $H_1 : \delta \in [-0.42, 0.5]$. We used a Cauchy prior centered at 0 with different width parameters $\omega = \{0.5/\sqrt{2} \approx 0.354,\ 1/\sqrt{2} \approx$

$0.707,\ 2/\sqrt{2} \approx 1.414\}$. We computed the BF for these width parameters to provide a sensitivity analysis to verify the robustness of results given different priors [46, 53]. For the Web study, we get $BF_{10} = 356$ for $\omega = 0.354$, this means that our participant accuracy is 356 times more likely under H_1 than under H_0 which means that we have extreme evidence that our effect size δ is within the equivalence margin. For $\omega = 0.707$, we get $BF_{10} = 493$ and for $\omega = 1.414$, we get $BF_{10} = 908.5$. The BFs for the VR study were $BF_{10} = \{618, 813, 1440\}$ for $\omega = \{0.345, 0.707, 1.414\}$ respectively which all indicate extreme evidence for equivalence. Prior and posterior distributions, as well as the BF for $\omega = 0.707$, are shown in Figure A.1 in the supplemental material.

5.3.3 Detection Accuracy of Single- and Multi-User Recordings. We further analysed the differences between participant accuracy of single-user detection and their accuracy of multi-user detection. In the Web and VR study, the differences between the accuracies were normally distributed ($p = .571$ for Web, $p = .446$ for VR). In the Web study, Participants were not significantly worse at detecting single-user recordings ($.473 \pm .189$) than at detecting multi-user recordings ($.518 \pm .219$), $t(46) = -.872, p = .388$, $d = -.127$. In the VR study, the detection accuracy for multi-user recordings ($.593 \pm .022$) was significantly higher by .173 than for single-user recordings ($.421 \pm .028$), $t(41) = 4.784$, $p < .001$, $d = .738$. Detection accuracies are visualized in Figure 3.

5.3.4 Detection Accuracy of Recordings of Actor 1 and Actor 2. We also analysed whether there was a difference in participant accuracy between detecting single-user recordings from Actor 1 and Actor 2. The differences between the accuracies were normally distributed for the Web study ($p = .053$) but not for the VR study ($p = .005$). In the Web study, participants detected Actor 1's recordings with a $.032 \pm .043$ higher accuracy, but the difference was not significant, $t(46) = .739$, $p = .464$, $d = .108$. In the VR study, there was a statistically significant median increase of .083 for recordings of Actor 1 (.5) compared to recordings of Actor 2 (.333), $z = 2.267$, $p = .023$. Detection accuracies are visualised in Figure 3.

5.3.5 Effect of Distance on Detection Accuracies and Perception of Politeness. We gathered additional data during the VR study only. Participants were watching the recordings from 3 different distances (between-subjects condition) based on proxemics (personal, social, and public distance), and we wanted to know whether distance would affect participants' detection accuracy and their perception of politeness of the avatars. We conducted a one-way ANOVA for the detection accuracies and a Kruskal-Wallis H test for the perception of politeness. There was no statistically significant difference in accuracies, $F(2, 39) = .045$, $p = .956$, and no statistical significance with regards to the perception of politeness, $\chi^2(2) = .049$, $p = .976$. To save space, more details can be found in the supplemental material (Section A.2).

6 DISCUSSION

6.1 Discussion Preference Round

Participants in both the Web and the VR study preferred single-user recordings to multi-user recordings. This was an unexpected outcome but we suspect that it has to do with the fluency of the conversations. In both studies, participants frequently mentioned that body language (gestures) (mentioned by 21 Web and 22 VR participants), naturalness (20 Web and 17 VR participants), and clarity/quality of voice of the avatars (18 Web and 14 VR participants) influenced which recording they preferred. Also, speaking speed was important for some participants (13 Web and 12 VR participants). Table A.3 in the supplemental material lists all the deciding factors that were named by participants for the Preference Round during both studies.

The multi-user recordings were created remotely, with both actors in different locations. This naturally created a bit of latency between the actors' responses. We did not try to correct the latency, but we smoothed all recordings (including the single-user ones) to compensate for the lower framerate in the multi-user recordings. For some participants, these minimal delays in the conversation might have made it less natural, especially since most participants seemed to prefer recordings with a faster speaking speed. In the single-user recordings, actors' responses often came quickly since they had to fit responses into the allocated gaps of the first recording. Some participants did not like when one avatar was almost interrupting the other avatar, but they still mostly preferred the more fast-paced recordings.

Participants might have been only subconsciously sensitive to the latency and explained their preferences in terms of naturalness. Only very few participants actually mentioned smoothness, fluency, or latency in their responses (8 Web and 0 VR participants).

Participants in both studies also preferred recordings of Actor 2 over recordings of Actor 1. Actor 2 who was a native English speaker spoke, in general, faster than Actor 1 which made her dialogues sound a bit more engaging. Also, her movements and hand gestures were often a bit faster and more energetic. Actor 2 spoke and moved slower. Some participants found this unnatural and some said that the movements looked robotic. Participants also thought that we sped up and slowed down some of the recordings although we did no such thing.

6.2 Discussion Detection Round

Participants' detection accuracies were equivalent to the random guessing average in both, the Web and the VR study. Participants mentioned that it was quite difficult and that they often were not sure. Participants primarily listened to the tone, pitch, and volume of the voices in order to figure out the number of actors (mentioned by 38 Web and 35 VR participants). Some noticed the different accents in the voices and were trying to use this as an indication (13 Web and 15 VR participants). Surprisingly few were looking at hand gestures and body movements (6 Web and 11 VR participants). Especially in VR, we would have expected that participants might be more interested in the movements than the voice. When we asked participants why they did not focus more on the movements some said that movements might have been computer generated. The AI-transformed voices were mostly convincing, only a few words occasionally sounded computer-generated, so participants probably assumed that most of the voices were natural. Another reason could be that the voices sounded natural and human to them, whereas the simplified cartoony avatar bodies were too abstract. We suspect that some participants did not treat the avatars with

their voices and movements as one entity but as two separate parts, where the voice was just added on top of some previously animated avatars.

There were a few factors that distinguished Actor 1 and Actor 2. Apart from the difference in speaking and movement speed, it was possible to detect the different accents of the actors. One participant who was mainly focusing on the accent was able to reach a detection accuracy of 79.2% (19/24 correct). During breaks, Actor 2 often had both hands slightly lifted, whereas Actor 1 left her hands dangling on her sides. One participant noticed these differences and also reached a detection accuracy of 79.2%. This shows that it could have been possible to pick out the actors based on either accent or body language alone.

We also compared participants' detection accuracies for single- and multi-user recordings separately. In the Web study, there was no significant difference in accuracies, however, in the VR study, multi-user recordings were detected significantly more often. We assumed that when there was no obvious indication of a recording being single-user, participants might have been more inclined to select two actors as this is what one might normally expect.

In the Web and the VR study, detection accuracies for Actor 1 were significantly higher than detection accuracies for Actor 2. Since participants also preferred recordings from Actor 2, we think that Actor 2's recordings were perceived as more natural than Actor 1's recordings and therefore believed to be more likely multi-user.

In the VR study, the social distances we used had no significant effect on participants' detection accuracies. Participants were predominantly listening to the voice of the avatars, hence the distance to the avatars did not seem important. We also hypothesised that participants who were positioned 1 meter away from the avatars (personal zone) might perceive them as less polite since the participants would be close enough to be part of the conversations while at the same time being ignored by the avatars. We theorised that participants who were positioned 2 and 4 meters away from the avatars (social and public zone) might feel more like observers than part of the conversation and thus rate the avatars as more polite. However, this was not the case; the majority of participants felt that the avatars were ignoring them but considered them polite regardless of the distance. It is possible that participants did not think that the avatars were realistic enough to relate to them at all.

6.3 Further Observations

Because the actors were non-professionals, they did not *act* like different characters in the single-user dialogues but always like themselves. We believe this made it easier for participants to detect single-user recordings than using professional actors. We also think that the record and replay tool might be easier to use for professional actors as they are used to learning lines by heart and misspeak less.

With regard to realism, we are aware that the avatars we used are simplistic and do not have any facial expressions. We did not use full-body avatars because our recordings were only driven by head and hand tracking and a full body would not have added any extra movement information for participants. We did not use face tracking and facial expressions as we were more interested in the gestures and because face tracking is not that widespread yet in consumer headsets and on SVR platforms. As it gains in importance we would consider this for future work.

Participants also looked at recordings only for about 15-20 seconds each time. If participants had had the time to look at recordings longer, they might have eventually figured out who is who. Since we want to use stacked recordings for creating NPCs in the background, users in SVR scenarios might not focus on them for more than a few seconds anyways.

6.4 Future Work

In this work, we compared single-user dialogues with multi-user dialogues but we did not cover object interaction. Single-user object interaction is a more challenging task that might require training of the actors to time the precise takeover of objects. Therefore, we chose to start with the easier task of performing dialogues. Investigating single-user object interactions in more detail could be interesting future work.

The creation of single-user dialogues works well in practice, however, for very long dialogues the audio indicators can get hard to read since they always show the whole dialogue and pauses between sentences might get increasingly difficult to estimate. An improvement could be to use a moving window to only show parts of the current dialogue at a time. For long dialogues, users have to manually scroll down to locate the text which can be tedious. It would be possible to synchronize the recorded audio with the text in the dialogues to improve usability.

Future work also aims to enhance stacked recordings using machine-learning methods to make them more adaptable to changes in the environment.

7 CONCLUSION

We presented interactive stacked recordings in VR that enable a single user to create interactive content with multiple characters to populate virtual environments with NPCs that behave naturally. Having given a single user the ability to recreate multi-user interactions, we wanted to investigate whether stacked recordings were qualitatively as good as regular recordings with multiple users and whether people were able to distinguish them from one another. We recruited two amateur actors to record a dialogues dataset together and separately and conducted two user studies, one online and one in VR with 89 participants in total, where we showed participants several single- and multi-user dialogues. Our results revealed that participants could not tell the difference between stacked and regular recordings and even slightly preferred the stacked recordings. We believe that single-user stacked recordings can provide an easy solution for animating multiple NPCs in VR in a believable way. The code and the materials for the studies, including the recordings, are available at https://doi.org/10.5522/04/23947278.v1.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860768 (CLIPE project).

REFERENCES

[1] Dimitra Anastasiadou, Pol Herrero, Julia Vázquez-De Sebastián, Paula Garcia-Royo, Bernhard Spanlang, Elena Álvarez de la Campa, Mel Slater, Andreea Ciudin,

Marta Comas, J. Antoni Ramos-Quiroga, and Pilar Lusilla-Palacios. 2023. Virtual Self-Conversation Using Motivational Interviewing Techniques to Promote Healthy Eating and Physical Activity: A Usability Study. *Frontiers in Psychiatry* 14 (2023). https://www.frontiersin.org/articles/10.3389/fpsyt.2023.999656

[2] Dimitrios Batras, Judith Guez, Jean-François Jégo, and Marie-Hélène Tramus. 2016. A Virtual Reality Agent-Based Platform for Improvisation between Real and Virtual Actors Using Gestures. In *Proceedings of the 2016 Virtual Reality International Conference (VRIC '16)*. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/2927929.2927947

[3] Klara Brandstätter. 2023. Immersive Record and Replay for Lively Virtual Environments. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Shanghai, CN, 979–980. https://doi.org/10.1109/VRW58643.2023.00332

[4] Aaron R Caldwell. 2022. Exploring Equivalence Testing with the Updated TOSTER R Package. *PsyArXiv* (2022). https://doi.org/10.31234/osf.io/ty8de

[5] Tender Claws. 2023. The Under Presents. https://tenderclaws.com/theunderpresents. Accessed: 2023-05-18.

[6] Tender Claws. 2023. The Under Presents Tempest. https://tenderclaws.com/tempest. Accessed: 2023-05-18.

[7] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, UK.

[8] Roblox Corporation. 2023. Roblox. https://www.roblox.com/. Accessed: 2023-05-22.

[9] NVIDIA Developer. 2022. VR Capture and Replay (VCR). https://developer.nvidia.com/vcr-early-access. Accessed: 2022-12-12.

[10] Andreas Rene Fender and Christian Holz. 2022. Causality-Preserving Asynchronous Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501836

[11] FlipsideXR. 2023. FlipsideXR Website. https://www.flipsidexr.com/. Accessed: 2022-05-18.

[12] Sebastian J. Friston, Ben J. Congdon, David Swapp, Lisa Izzouzi, Klara Brandstätter, Daniel Archer, Otto Olkkonen, Felix Johannes Thiel, and Anthony Steed. 2021. Ubiq: A System to Build Flexible Social Virtual Reality Experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (Osaka, Japan) *(VRST '21)*. ACM, New York, NY, USA, Article 6, 11 pages. https://doi.org/10.1145/3489849.3489871

[13] Geoffrey Gorisse, Olivier Christmann, and Charlotte Dubosc. 2022. REC: A Unity Tool to Replay, Export and Capture Tracked Movements for 3D and Virtual Reality Applications. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI 2022)*. ACM, New York, NY, USA, 1–3. https://doi.org/10.1145/3531073.3534472

[14] C. Greenhalgh, M. Flintham, J. Purbrick, and S. Benford. 2002. Applications of Temporal Links: Recording and Replaying Virtual Environments. In *Proceedings IEEE Virtual Reality 2002*. IEEE, Orlando, FL, USA, 101–108. https://doi.org/10.1109/VR.2002.996512 ISSN: 1087-8270.

[15] Edward T. Hall, Ray L. Birdwhistell, Bernhard Bock, Paul Bohannan, A. Richard Diebold, Marshall Durbin, Munro S. Edmonson, J. L. Fischer, Dell Hymes, Solon T. Kimball, Weston La Barre, S. J. Frank Lynch, J. E. McClellan, Donald S. Marshall, G. B. Milner, Harvey B. Sarles, George L Trager, and Andrew P. Vayda. 1968. Proxemics [and Comments and Replies]. *Current Anthropology* 9, 2/3 (1968), 83–108. http://www.jstor.org/stable/2740724

[16] Lewis G. Halsey. 2019. The Reign of the P-Value is Over: What Alternative Analyses Could We Employ to Fill the Power Vacuum? *Biology Letters* 15, 5 (2019), 20190174. https://doi.org/10.1098/rsbl.2019.0174 arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsbl.2019.0174

[17] Sarah Hofmann, Cem Özdemir, and Sebastian von Mammen. 2023. Record, Review, Edit, Apply: A Motion Data Pipeline for Virtual Reality Development & Design. In *Proceedings of the 18th International Conference on the Foundations of Digital Games (FDG '23)*. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/3582437.3587191

[18] Daniel Holden. 2018. Robust Solving of Optical Motion Capture Data by Denoising. *ACM Trans. Graph.* 37, 4, Article 165 (jul 2018), 12 pages. https://doi.org/10.1145/3197517.3201302

[19] Scott Ronald Howie and Marco Gilardi. 2019. Virtual Observation of Virtual Reality Simulations. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312836

[20] Sebastian Hubenschmid, Jonathan Wieland, Daniel Immanuel Fink, Andrea Batch, Johannes Zagermann, Niklas Elmqvist, and Harald Reiterer. 2022. ReLive: Bridging In-Situ and Ex-Situ Visual Analytics for Analyzing Mixed Reality User Studies. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, 1–20. https://doi.org/10.1145/3491102.3517550

[21] Mindshow Inc. 2023. Mindshow Website. https://mindshow.com/. Accessed: 2023-05-18.

[22] Rec Room Inc. 2023. Rec Room. https://recroom.com/. Accessed: 2023-05-22.

[23] VRChat Inc. 2023. VRChat. https://hello.vrchat.com/. Accessed: 2023-05-22.

[24] Wist Labs Inc. 2023. Wist – Step Inside Your Memories. https://wistlabs.com/. Accessed: 2023-05-18.

[25] JASP Team. 2022. JASP (Version 0.17.1)[Computer software]. https://jasp-stats.org/

[26] Harold Jeffreys. 1961. *Theory of Probability*. Oxford University Press, Oxford, UK.

[27] Manos Kamarianakis, Ilias Chrysovergis, Nick Lydatakis, Mike Kentros, and George Papagiannakis. 2022. Less is More: Efficient Networked VR Transformation Handling Using Geometric Algebra. *Advances in Applied Clifford Algebras* 33 (Dec. 2022). https://doi.org/10.1007/s00006-022-01253-9

[28] Robin K. Kammerlander, André Pereira, and Simon Alexanderson. 2021. Using Virtual Reality to Support Acting in Motion Capture with Differently Scaled Characters. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, Lisboa, PT, 402–410. https://doi.org/10.1109/VR50410.2021.00063 ISSN: 2642-5254.

[29] Simon Kloiber, Volker Settgast, Christoph Schinko, Martin Weinzerl, Johannes Fritz, Tobias Schreck, and Reinhold Preiner. 2020. Immersive Analysis of User Motion in VR Applications. *The Visual Computer: International Journal of Computer Graphics* 36, 10-12 (Oct. 2020), 1937–1949. https://doi.org/10.1007/s00371-020-01942-1

[30] Daniël Lakens. 2017. Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science* 8, 4 (2017), 355–362. https://doi.org/10.1177/1948550617697177 arXiv:https://doi.org/10.1177/1948550617697177 PMID: 28736600.

[31] Klemen Lilija, Henning Pohl, and Kasper Hornbæk. 2020. Who Put That There? Temporal Navigation of Spatial Recordings by Direct Manipulation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376604

[32] Maximilian Linde, Jorge Tendeiro, Ravi Selker, Eric-Jan Wagenmakers, and Don van Ravenzwaaij. 2021. Decisions about Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. *Psychological Methods* (11 2021). https://doi.org/10.1037/met0000402

[33] Thomas Lopez, Olivier Dumas, Fabien Danieau, Bertrand Leroy, Nicolas Mollet, and Jean-François Vial. 2017. A Playback Tool for Reviewing VR Experiences. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology (VRST '17)*. ACM, New York, NY, USA, 1–2. https://doi.org/10.1145/3139131.3141776

[34] Karthik Mahadevan, Qian Zhou, George Fitzmaurice, Tovi Grossman, and Fraser Anderson. 2023. Tesseract: Querying Spatial Design Recordings by Manipulating Worlds in Miniature. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3580876

[35] Sophie Maria, Solène Lambert, and Ignacio Avellino. 2022. From Déjà vu to Déjà vécu: Reliving Surgery in Post-Operative Debriefing. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Christchurch, NZ, 462–465. https://doi.org/10.1109/VRW55335.2022.00102

[36] Medicalholodeck. 2023. RecordXR – Record, Replay, Share in Virtual Reality. https://www.medicalholodeck.com/en/record-replay-share-in-virtual-reality-record-XR/. Accessed: 2023-05-18.

[37] Meta. 2023. Meta Horizon Worlds. https://www.meta.com/gb/horizon-worlds/. Accessed: 2023-05-22.

[38] Richard Morey and Jeffrey Rouder. 2011. Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological methods* 16 (07 2011), 406–19. https://doi.org/10.1037/a0024377

[39] Netlify. 2023. Netlify Homepage. https://www.netlify.com/?attr=homepage-modal

[40] Kevin Ngo. 2023. Rapid Motion-Capture-Powered VR Development. https://aframe.io/blog/motion-capture/. Accessed: 2023-05-18.

[41] Sofia Adelaide Osimo, Rodrigo Pizarro, Bernhard Spanlang, and Mel Slater. 2015. Conversations Between Self and Self as Sigmund Freud—A Virtual Body Ownership Paradigm for Self Counselling. *Scientific Reports* 5, 1 (Sept. 2015), 13899. https://doi.org/10.1038/srep13899

[42] Pontoco. 2023. The Last Clockwinder. https://pontoco.com/the-last-clockwinder. Accessed: 2023-05-18.

[43] Prolific. 2023. Prolific Homepage. https://www.prolific.co/

[44] Koe: Recast. 2022. Koe: Recast Website. https://koe.ai/. Accessed: 2022-12-12.

[45] Hiroshi Sakuma, Hideyuki Takahashi, Kohei Ogawa, and Hiroshi Ishiguro. 2023. Immersive Role-Playing with Avatars Leads to Adoption of Others' Personalities. *Frontiers in Virtual Reality* 4 (2023). https://www.frontiersin.org/articles/10.3389/frvir.2023.1025526

[46] Xenia Schmalz, José Biurrun Manresa, and Lei Zhang. 2021. What is a Bayes Factor? *Psychological Methods* (11 2021). https://doi.org/10.1037/met0000421

[47] M. Slater, J. Howell, A. Steed, D-P. Pertaub, and M. Garau. 2000. Acting in Virtual Reality. In *Proceedings of the third international conference on Collaborative virtual environments*. ACM, San Francisco California USA, 103–110. https://doi.org/10.1145/351006.351020

[48] Mel Slater, Solène Neyret, Tania Johnston, Guillermo Iruretagoyena, Mercè Álvarez de la Campa Crespo, Miquel Alabèrnia-Segura, Bernhard Spanlang, and Guillem Feixas. 2019. An Experimental Study of a Virtual Reality Counselling

Paradigm Using Embodied Self-Dialogue. *Scientific Reports* 9, 1 (July 2019), 10903. https://doi.org/10.1038/s41598-019-46877-3

[49] Anthony Steed, Lisa Izzouzi, Klara Brandstätter, Sebastian Friston, Ben Congdon, Otto Olkkonen, Daniele Giunchi, Nels Numan, and David Swapp. 2022. Ubiq-Exp: A Toolkit to Build and Run Remote and Distributed Mixed Reality Experiments. *Frontiers in Virtual Reality* 3 (2022), 131. https://doi.org/10.3389/frvir.2022.912078

[50] Anthony Steed, Mingqian Wang, and Jason Drummond. 2019. *Recording and Replaying Virtual Environments for Development and Diagnosis.* CRC Press, Boca Raton, FL, USA, 349–361. https://doi.org/10.1201/b21598-18

[51] William Steptoe, Jean-Marie Normand, Oyewole Oyekoya, Fabrizio Pece, Elias Giannopoulos, Franco Tecchia, Anthony Steed, Tim Weyrich, Jan Kautz, and Mel Slater. 2012. Acting Rehearsal in Collaborative Multimodal Mixed Reality Environments. *Presence: Teleoperators and Virtual Environments* 21, 4 (Nov. 2012), 406–422. https://doi.org/10.1162/PRES_a_00109

[52] Katherine Sunderland, Harvey Keselman, James Algina, Lisa Lix, and Rand Wilcox. 2003. Conventional And Robust Paired And Independent-Samples t-Tests: Type I Error And Power Rates. *Journal of Modern Applied Statistical Methods* 2 (11 2003), 481–496. https://doi.org/10.22237/jmasm/1067646120

[53] Don van Ravenzwaaij, Rei Monden, Jorge Tendeiro, and John Ioannidis. 2019. Bayes factors for Superiority, Non-Inferiority, and Equivalence Designs. *BMC Medical Research Methodology* 19 (03 2019). https://doi.org/10.1186/s12874-019-0699-7

[54] Cheng Yao Wang, Shengguang Bai, and Andrea Stevenson Won. 2020. ReliveReality: Enabling Socially Reliving Experiences in Virtual Reality via a Single RGB camera. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).* IEEE, Atlanta, GA, USA, 710–711. https://doi.org/10.1109/VRW50115.2020.00206

[55] Cheng Yao Wang, Mose Sakashita, Upol Ehsan, Jingjin Li, and Andrea Stevenson Won. 2020. Again, Together: Socially Reliving Virtual Reality Experiences When Separated. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376642

[56] Weilai Xu, Ismail Alarab, Charlie Lloyd-Buckingham, Steve Bowden, Benjamin Noer, Fred Charles, Simant Prakoonwit, Andrew Callaway, Shelly Ellis, and Chris Jones. 2022. Re-enacting Football Matches in VR Using Virtual Agents' Realistic Behaviours. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR).* IEEE, CA, USA, 119–123. https://doi.org/10.1109/AIVR56993.2022.00024 ISSN: 2771-7453.

[57] Tairan Yin, Ludovic Hoyet, Marc Christie, Marie-Paule Cani, and Julien Pettré. 2022. The One-Man-Crowd: Single User Generation of Crowd Motions Using Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 5 (May 2022), 2245–2255. https://doi.org/10.1109/TVCG.2022.3150507 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[58] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue.* Association for Computational Linguistics, Singapore and Online, 121–132. https://arxiv.org/abs/2109.06427