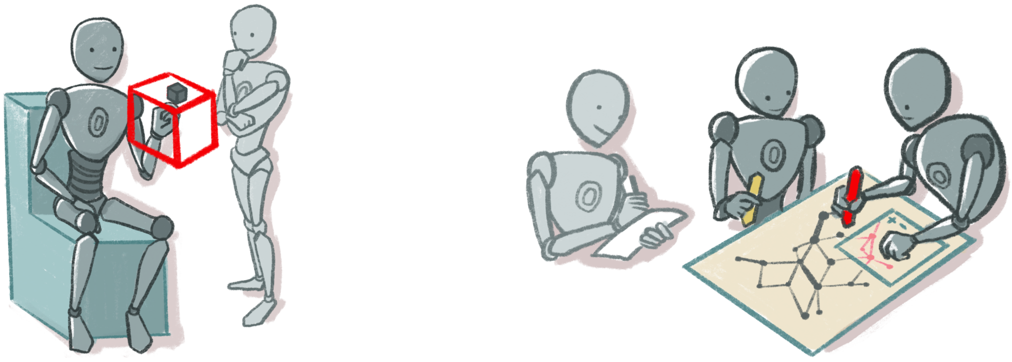
Remote and Collaborative Virtual Reality Experiments via Social VR Platforms

David Saffo*
Northeastern University
Khoury College of Computer Sciences
saffo.d@northeastern.edu

Sara Di Bartolomeo
Northeastern University
Khoury College of Computer Sciences
dibartolomeo.s@northeastern.edu

Caglar Yildirim
Northeastern University
Khoury College of Computer Sciences
c.yildirim@northeastern.edu

Cody Dunne
Northeastern University
Khoury College of Computer Sciences
c.dunne@northeastern.edu



(a) Quantitative Study: Fitts' Law in 3D VR

(b) Qualitative Study: Tabletop Collaboration

Figure 1: Depiction of the two user studies we replicated using VRChat, a social VR platform. The darker robots represent the avatars of the participants. The lighter robots represent the avatars of the researchers, observing the participants and collecting data. *Left:* In quantitative Study 1 — extending Fitts' Law in 3D VR [8] — a participant selects a target by moving the cursor to intersect it. *Right:* In qualitative Study 2 — a trial from the tabletop study from Tang et al. [34] — two participants collaborate to find the best path between two nodes in a network.

ABSTRACT

Virtual reality (VR) researchers struggle to conduct remote studies. Previous work has focused on working around limitations imposed by traditional crowdsourcing methods. However, the potential for leveraging social VR platforms for HCI evaluations is largely unexplored. These platforms have large VR-ready user populations, distributed synchronous virtual environments, and support for user-generated content. We demonstrate how social VR platforms can be used to practically and ethically produce valid research results by replicating two studies using one such platform (VRChat): a quantitative study on Fitts' Law and a qualitative study on tabletop collaboration. Our replication studies exhibited analogous results to the originals, indicating the research validity of this approach. Moreover, we easily recruited experienced VR users with their own hardware for synchronous, remote, and collaborative participation. We further provide lessons learned for future researchers experimenting using social VR platforms. This paper and all supplemental materials are available at `osf.io/c2amz`.

CCS CONCEPTS

• **Human-centered computing → Virtual reality**; *Usability testing*; • **Information systems → Crowdsourcing**.

KEYWORDS

Virtual Reality, Social VR, Crowdsourcing, Replication Study, Transferability Study, Quantitative Study, Qualitative study

ACM Reference Format:
David Saffo, Sara Di Bartolomeo, Caglar Yildirim, and Cody Dunne. 2021. Remote and Collaborative Virtual Reality Experiments via Social VR Platforms. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445426

*ORCIDS: Saffo, Di Bartolomeo, Yildirim, Dunne.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445426

1 INTRODUCTION

The advent of crowdsourced evaluation methods have reshaped the way many scientists construct and conduct scientific experiments. In human computer interaction (HCI) research, crowdsourcing is used extensively to recruit remote participants for a wide variety of user studies [2, 11, 15, 17]. Crowdsourcing can have advantages over traditional in-lab evaluation methods — faster turnaround times, asynchronous participation, larger and varied recruitment populations [26] — though may also introduce challenges such as catching speeders and cheaters [17]. But virtual reality (VR) evaluations are still conducted primarily in-lab. This is largely due to the specialized head-mounted displays (HMDs) often required for these studies [31], as well as the difficulty and overhead associated with distributing VR applications.

As a result, VR researchers have not been able to leverage the full potential of crowdsourced evaluation methods. VR user studies have been mainly limited to what could be practically studied in-lab — i.e., smaller sample sizes, student populations, limited access to experienced VR users, and little collaboration. The COVID-19 pandemic has only served to exacerbate these issues, as in-lab studies became ethically questionable and practically difficult in many locales. Existing research has examined various crowdsourcing methods for VR experiments [16, 18, 30]. But one potential method unique to VR has yet to be explored in detail — social VR platforms.

Social VR platforms connect large populations of VR users and allow people to meet, chat, play, and explore together in a distributed, synchronous virtual environment. Many platforms allow uploading user-generated content in the form of custom worlds, animations, objects, and avatars. This provides an opportunity for researchers to implement and conduct VR user studies, which may have otherwise been difficult or impossible. E.g., by using a social VR platform, researchers could implement and conduct a collaborative VR experiment where experimenters and participants are remotely present in the same synchronous virtual environment — without the overhead of application distribution and networking implementations. However, as with any new research method, there are a number of questions to explore in regards to the practicality, validity, and ethics of conducting evaluations in this manner.

In this paper, we explore the suitability of social VR platforms generally — and VRChat (`vrchat.com`) in particular — for conducting VR user studies. We contribute:

1. The results of two preregistered replication studies we conducted — one quantitative and one qualitative — which demonstrate the practicality, validity, and ethics of running user studies within a social VR platform;
2. Implementation details and open materials for these studies along with recommendations to guide future researchers in adopting our approach, particularly using VRChat; and
3. A discussion of the future and implications of our approach.

This paper and all supplemental materials — including preregistration, stimuli videos, experiment code, collected data, and analysis code — is freely available at `osf.io/c2amz`.

2 RELATED WORK

Crowdsourcing HCI user studies using platforms such as Mechanical Turk is now commonplace. Existing literature supports this method's validity and usefulness [1, 4, 6, 12, 24, 25, 29]. E.g., Heer & Bostock [13] successfully replicated Cleveland & McGill's in-lab graphical perception studies using Mechanical Turk [9].

Steed et al. [31] recently detailed the challenges of evaluating immersive experiences during the COVID-19 pandemic. E.g., such evaluations involve specialized hardware and often require proctoring, resulting in them being conducted in-lab. Steed et al. propose several potential solutions — using lab personnel, recruiting participants with the necessary hardware, and distributing hardware to participants — as well as guidance for remote experiment design.

These alternatives to in-lab experiments have been at least partly explored. Mottelson et al. [21] distributed Google Cardboard hardware to participants, while Steed et al. [30] recruited Google Cardboard and Samsung Gear users directly via web pages and emails. To recruit VR participants more broadly, Huber et al. [16] used a purpose-built scientific study platform (`labinthewild.org`). Ma et al. [18] instead used Mechanical Turk directly. 190 of their 242 surveyed Turkers reported having access to a smartphone-based VR head-mounted display (HMD) (Samsung Gear VR, Google Cardboard), while only 18 reported access to more advanced VR equipment (HTC Vive). The studies mentioned here all primarily used smartphone-based HMDs — e.g., Ma et al. had only their Samsung Gear VR users participate in their study.

Using a smartphone-based HMD for a VR user study is not always practical or desirable. Many researchers study more advanced VR technologies with high-resolution displays, positional tracking capabilities, and motion-control handhelds. Moreover, smartphone-based HMDs are disappearing — both Google Cardboard and Samsung Gear VR are discontinued. Participants with access to more advanced VR equipment do not represent enough of the population of common crowdsourcing websites for those sites to be practical for many VR researchers [18]. Distributing advanced VR hardware to pools of potential participants (proposed by Steed et al. [30]) could alleviate the problem — but would require extensive funding ($300–3000 USD per participant) and coordination by all involved.

We have observed several recent instances of researchers recruiting participants with more advanced hardware from VR-related web forums and specialized crowdsourcing websites. The XR Distributed Research Network (`xrdrn.org`) was also recently created to consolidate VR study participant recruitment, in response to the COVID-19 pandemic. Posts on these sites typically ask participants to download the study software, conduct the experiment themselves, then upload the results — requiring extensive implementation effort, especially to support collaborative studies.

Saffo & Di Bartolomeo et al. [27] first introduced the idea of using a social VR platform for crowdsourcing VR experiments. They explored the practicality of using custom virtual environments in VRChat to implement a usability study. They recruited participants within VRChat and asked them to search then exit a maze. While their results demonstrate the promise of using social VR platforms for running user studies, they did not evaluate the validity of the study results — arguing that full studies should be conducted before such conclusions can be reached. Additionally, they did not discuss the ethical concerns that may arise when using social VR platforms to recruit participants [31].

3 REPLICATION STUDY SELECTION

Careful study selection and design is critical for validating the efficacy of a novel evaluation method. Replicating a widely-known and well-regarded study is a common approach (e.g., Heer & Bostock [13]), in which the original results are compared against a novel methodology. It is also important to select studies with sufficient complexity to test the capabilities and practicality of an evaluation method, beyond simply the validity of the derived results. With these criteria in mind, we selected two classic studies to validate our proposed approach of conducting user studies within social VR platforms — (1) a new take on the classic topic of Fitts' Law and (2) a study on immersive collaboration, a topic of increasing academic and industry interest.

3.1 Defining Replication Success

Before detailing our replication selections, it is important to define what we consider to be a successful replication. Replication studies often employ some form of heterogeneity metrics and statistical analyses to determine if a replication was successful. These methods answer the question, "Do the replication results quantitatively match the original?" However, relying on heterogeneity metrics does not reflect the purpose of a replication study, as studies with different results can still come to the same conclusions.

Accordingly, we use the definition for replication success presented by Mathur and VanderWeele [19]: "The goal is not to determine whether the replication studies are similar to one another, but rather to determine whether they support the scientific effect under investigation." If we satisfy this criterion, then we can conclude our proposed evaluation methodology allowed for adequate control of experimental factors — and thus can produce efficacious results. If this criterion is not met, then we can conclude that our proposed method has additional factors that need to be accounted for — or, in the worst case, cannot produce efficacious results.

3.2 Study 1: Extending Fitts' Law in 3D VR

We first replicate Clark et al.'s [8] study. They extended Fitts' Law to model the time it takes to point to an object as a function of target size and distance — within the domain-specific constraints of VR applications. They compare the movement time predictive performance of Fitts' Law to popular 3D Fitts' Law extensions and their own empirically-driven model. Their study was conducted in-lab with 23 participants using a VR HMD and relied primarily on *quantitative* measures. Fitts' Law has been extensively studied, replicated, and extended to other contexts.

Our replication study (1) tests the generalizability of Clark et al.'s [8] empirical model and (2) demonstrates the validity of using social VR platforms for conducting remote yet proctored *quantitative* VR user studies.

The effects we investigated in the study were:

(1) Whether Fitts' index of difficulty (ID) would be a significant predictor of movement times;
(2) Whether the model would accurately predict movement times in VR; and
(3) Whether movement time would vary as a function of target size and depth.

3.3 Study 2: Tabletop Collaborative Coupling

Our second replication is Tang et al. [34]'s study examining the collaborative coupling style of participants interacting with a tabletop display. Tang et al. studied the coupling styles and table positions of participants over several trials featuring two tasks (independent or compromise) and two data layer interaction techniques (filter or lenses). Their study was conducted in-lab with 4 pairs of co-located participants and relied primarily on *qualitative* measures. The findings of Tang et al. have been extensively cited and have helped inform the design of related immersive applications.

Our replication (1) examines the transferability of results to a VR context, (2) demonstrates the use of a social VR platform to conduct remote yet proctored *collaborative* and *qualitative* VR research, and (3) identifies opportunities for future VR-specific extensions.

Collaboration over immersive technologies is a popular contemporary research topic that has seen renewed interest. Collaboration has become an increasingly common part of VR research [38], applications [5], and games [10]. That said, conducting this style of VR research is challenging — both in-lab and remotely. This research paradigm often requires a large amount of space, multiple sets of HMDs and equipment, VR avatars, inverse kinematics, and distributed networking. Furthermore, remote evaluation methods are not typically well-suited for collecting many qualitative measures, such as the talk-aloud protocol and video recordings employed in Tang et al. [34]. Using a social VR platform to implement this study synchronously takes care of most of this overhead automatically and highlights the utility of this approach for VR studies. To the best of our knowledge, this will be the first VR collaborative evaluation study conducted both remotely and synchronously — with participants and proctors all sharing the same virtual space.

The effects we investigated in the study were:

(1) The six coupling styles (fig. 3);
(2) How individuals would work independently with lenses; and
(3) The usage of perspective sharing when tightly coupled.

4 STUDY APPARATUS

The apparatus for this study was VRChat, a social VR platform where our proctors interacted with participants who were using their own head-mounted displays (HMDs). We only recruited participants with HMDs that support the PC version of VRChat, perform head tracking, and connect with two handheld motion controllers. More details on the HMDs participants had access to can be found in section 5 and table 4. Here we justify our decision for selecting VRChat and detail many of the intricacies of using the platform for VR user studies. VRChat was selected as our study apparatus as it met our postulated requirements for conducting evaluative studies: it was freely available, protective of user privacy, supports cross-platform compatibility; sustains an active and VR-equipped userbase; and extensively supports user-generated content.

4.1 Free, Private, and Cross-Platform

Our ideal social VR platform would be easily accessible to the most users (and researchers) as possible. I.e., freely available on variety of HMDs and platforms. Additionally, in order to ethically instruct participants to sign up and download a platform, it is important to consider how the platform will respect their privacy.

VRChat is a free-to-play game that is distributed through the Steam and Oculus digital distribution services. As of 2020-12 VRChat is not monetized by in-game purchases, advertisements, or by selling user data — instead, it is largely funded by external investors [37] and through an optional subscription that provides quality of life features. VRChat, Steam, and Oculus all require users to register for an account on the platform before it can be used. VRChat and Steam only require an email address, username, and password in order to sign up. Email addresses are not publicly displayed. As a result, VRChat & Steam are the best in terms of privacy — they require the least amount of personal information possible and offer the possibility of anonymity. Accessing VRChat through Oculus offers less privacy. Oculus accounts require a user's full name, and soon will only be accessible with a Facebook account [23].

VRChat features a Windows 10 PC version and an Android Oculus Quest version. The PC version officially supports many of the most popular HMDs[1], and it unofficially supports any HMD supported by Steam VR. The PC version can also be played without a HMD using the desktop version either with a keyboard and mouse or gamepad — which is useful for researchers to monitor participants and collect data. VRChat is widely accessible and we were ethically comfortable asking participants to use the platform.

4.2 Active VR-Equipped Userbase

Participant recruitment is a challenge in most user studies, and more so for remote studies. Recruiting users onto a social platform like VRChat to participate in a study would be a viable approach. However, recruiting participants who are already on the platform would be preferable in many scenarios. Unless the study in question requires a very specific population, recruiting active users with experience on the chosen platform would be ideal. These users will be familiar with the navigation, controls, and limitations of the platform, and thus will have an easier time joining and executing the study. Therefore, it is important for the selected social VR platform to have a large and active VR-equipped userbase.

VRChat is one of the most popular social VR platforms and one of the most popular VR applications in general in terms of active users. Accounting only for Steam users, the platform averaged more than 11,000 concurrent players in 2020-08 with a peak of over 17,000 concurrent players[2] — an increase of approximately 50% over 2019-08. According to VRChat, VR users accounted for 30% of the daily player population in 2018 [36]. Assuming this percentage has not changed significantly, we can extrapolate that there was on average more than 3,000 concurrent VR users through the month of 2020-08. This massive usrebase provides a large population of VR equipped and experienced participants from which to recruit.

4.3 Support for User-Generated Content

Social VR platforms generally allow for some level of user-generated content. However, the developers of these platforms need to weigh the benefits of allowing more creative freedom with the cost of losing some control over how the platform is used. Regardless, for a platform to be useful for a wide range of research purposes it must provide enough creative freedom to implement novel, programmatic, and complex stimuli, interactions, and applications. VRChat focuses heavily on enabling user-generated content in the form of custom worlds, avatars, animations, objects, and shaders. The following sections detail the necessary background for developing user-generated content on VRChat for implementing user studies.

4.3.1 Custom Worlds. VRChat allows users to create and upload custom worlds created with Unity game engine[3] and the VRChat SDK[4]. These worlds can feature any 2D or 3D objects, textures, or shaders that can be created inside Unity or can be imported as assets into Unity. The VRChat SDK3 also features Udon, a proprietary graphical programming language for creating custom scripts in a world. Udon supports standard programming language features such as data types, conditional statements, and loops. It also features higher-level functions for handling game events, variable and event synchronization between clients, and player tracking. The SDK comes with several pre-built components for handling common VR and VRChat interactions such as picking up objects, virtual chairs, and portals between worlds. Moreover, many native Unity components are also integrated into and exposed by the SDK.

That said, the current VRChat SDK3 does have limitations that will affect VR researchers and their ability to implement and conduct studies. Perhaps the biggest limitation of the VRChat SDK3 is that it currently does not allow any data to be sent or received outside of the VRChat client. This means that any data collected on the platform cannot be exported or saved to a server. Therefore, data must be recorded by other means. In addition, interactions are limited to the grab and trigger buttons of HMD motion controllers. All other buttons are reserved for VRChat-specific functions such as movement and menu controls. The last major limitation we encountered was the current client synchronization system. Synced variables are limited to around 150 bytes, and synced event messages can sometimes be dropped by the network.

Despite these limitations, and custom logic being constrained to the features of Udon, the range of what can be implemented using the SDK is vast. This is best seen in breadth of VRChat community creations — e.g., a convolutional neural network object detector [28], interactive cubic Bézier curves [35], and a portal gun from the popular game Portal [32]. The VRChat SDK3 provides researchers with the necessary tools to implement complex and novel VR studies — far beyond what was originally observed by Saffo and Di Bartolomeo et al. using VRChat SDK2 [27].

4.3.2 VR Avatars. Social VR platforms use avatars to represent their users in the virtual environment. Most platforms use an avatar with standardized dimensions that users can further customize to better represent themselves. VRChat does not use a standardized avatar and instead allows users to create and upload their own avatars using the VRChat SDK. Avatars can be copied off other players or selected from avatar pedestals placed in custom worlds. The avatar SDK allows for models with support for partial or full body tracking, simulated eye movements, simulated expressions, mouth movement, and walking animations.

[1] https://docs.vrchat.com/docs/controls
[2] https://steamcharts.com/app/438100
[3] https://unity.com/
[4] https://docs.vrchat.com/docs/choosing-your-sdk

A user's in-game dimensions, e.g., height and arm length, are defined by their selected avatar. As an avatar can be just about anything, it is important for researchers to select one standardized avatar for all participants to use — unless the experiment requires otherwise. Additionally, avatars do not scale proportional to the size of the person controlling them. This may affect experimental results. Researchers could consider having multiple avatars of different sizes to accommodate people of different heights and arm lengths, or ensuring that the experiment works well for a single avatar that reasonably fits their average participants.

4.3.3 Community Support. The amount of community support for developing with Udon warrants discussion as well. Despite Udon being less than a year old, there already exist many user guides and tutorials for beginners. Furthermore, the official VRChat forum and Discord both have channels where users can post questions and get help or feedback. There also exists many user-created Udon resources, components, and toolkits that make developing for the platform easier. One notable example of this is UdonSharp [20], a user-created C# variation to Udon compiler. This compiler allowed us to write all of our experiment code in a familiar language instead of relying on the Udon graphical programming interface.

5 RECRUITMENT AND PARTICIPANT DEMOGRAPHICS

In order to ensure our research was conducted as ethically as possible, we submitted our study for approval by our institutional review board (IRB). As a part of this process, we reached out to the VRChat administrators to verify our proposed study was an appropriate use of their platform. We received permission to continue with our study as described, with the stipulation that we only recruit participants from outside the VRChat client. This means we would not approach players in game and ask them to participate, as was done by Saffo and Di Bartolomeo et al. [27]. Rather, we could recruit participants through online forums, social circles, or other places VR users could be reached — given that solicitation was allowed.

We recruited participants primarily through two channels: the VRChat Reddit page and the VRChat official forum. Before posting our IRB-approved recruitment advertisement, we made sure our post would not break any of the forum rules. Our post called for VR users who were interested in participating in one or both of our studies to sign up using an online survey. The first page of the survey presented the consent form and related details for the selected study. The subsequent pages asked participants several questions about their access to HMDs, relevant physical conditions, optional demographic information, VR experience, and contact information. The consent form and survey were approved by our IRB and can be viewed in the supplemental material and at `osf.io/c2amz`.

In order to qualify for our study, participants had to be 18 years of age or older, fluent with the English language, and have access to a VRChat-compatible HMD and the PC version of VRChat. After completing the survey, participants who qualified were redirected to a web page to sign up for a time to meet in-game for the study.

In less than 24 hours we reached our recruitment goal of 23 participants. Select participant data can be seen in table 4 and table 5. Additional data can be found in the supplemental material and at `osf.io/c2amz`. Of these participants, 15 signed up for the Fitts' Law study and 8 signed up for both. The average participant age was 23 (min=18, max=34, std=3.9), and the majority of our participants self-identified as men, were out of school with some level of high school or college degree, and resided within North America. Participants did not report any physical conditions that would significantly hinder their ability to complete either study. All participants had access to a compatible HMD, the most common of which was HTC Vive, followed by Valve Index and Oculus Rift(S).

We also asked participants several questions about their VR experience and usage. The majority of participants reported they were experienced with both VR in general and VRChat specifically. Additionally, the majority of participants reported that they do not easily experience motion sickness while using VR. Participants also commonly reported using VR and VRChat at least once a month or more, with only one participant reporting that they had never used VRChat. Finally, participants were provided with a free response field where they could provide additional details about themselves and why they wanted to participate. Many of the responses centered around interest in VR Research and curiosity about how VR studies can be conducted on social VR platforms.

6 EXTENDING FITTS' LAW IN 3D VR

Clark et al. [8] describe three phases of analysis for their VR Fitts' Law extension study. In phase 1, they use a mixed liner model to examine the effects of 3D target position on movement time. Phase 2 uses the results of the previous phase to construct an empirical model for movement time. Finally, phase 3 compares the empirically derived model to traditional Fitts' Law [14], as well as the other 3D Fitts' Law extensions presented by Murata and Iwase [22], Cha and Myung [7], and Machuca and Stuerzlinger [3]. Our replication study is focused on Clark et al.'s phase 3 analysis.

We did not have access to any of the original data, experiment resources, or analysis code. All of our experiment and analysis code was developed from the details provided in the original paper [8]. We also preregistered our replication plan and analysis code at `osf.io/awtvq`, which was developed with the help of a pilot study using 4 participants recruited from our personal connections. These materials, along with our experiment code and results data, can be found in supplemental materials and at `osf.io/c2amz`.

6.1 Methods

The experimental task consisted of 96 unique selection trials inside a cubical interaction space with the origin at the center. Trials could appear at 24 unique (X,Y,Z) positions from the center and at 4 sizes. Each trial began by selecting a reset target at the origin and ended when the trial target was selected. Targets were selected with a cube-shaped cursor held by participants. When the cursor collided with the target, the target changed colors and could be selected with the trigger button. Trial times were recorded from when the reset target was selected to when the trial target was selected.

Once the data was recorded, it was then converted to spherical coordinates which were used by the predictive models in our analysis. Index of difficulty (ID) scores were standardized across all models and calculated with the equation $ID = log_2(\frac{2R}{S+P})$ defined in the original study [8], where R is the radial distance, S is the target size, and P is the pointer size multiplied by a constant. Outliers

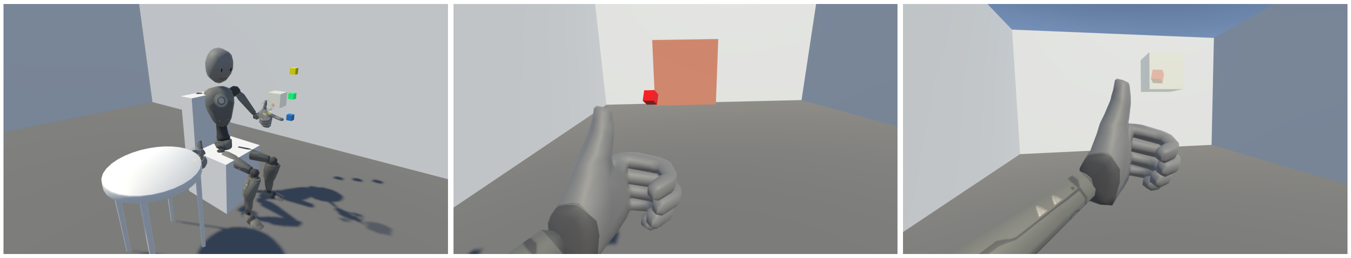


Figure 2: Screenshots from our Fitts' Law study. *Left:* How a participant looked while completing tasks from the researcher's perspective. The others show the participant's perspective. *Middle:* The participant moving the cursor towards a target. *Right:* The cursor colliding with the target, with the participant soon to select it and complete a trial.

Table 1: Candidate models for the Fitts' Law in 3D VR study.

Hoffmann [14]	$MT = a + b\ (ID)$
Murata & Iwase [22]	$MT = a + b(sin\varphi) + c(ID)$
Cha & Myung [7]	$MT = a + b(sin\varphi) + c(\theta) + d(ID)$
Machuca & Stuerzlinger [3]	$MT = a + b(ID) + c(CTD)$
Clark et al. [8]	$MT = a + b(ID) + c(\theta) + d(\theta xS)$

1.5 × IQR above the 75th percentile for each trial were removed. Additionally, after completing the study we found and removed 4 trials that often spawned on top of the selection cursor — causing issues due to the way Unity's "OnEnter" event works. Once outliers were removed, participants' data was averaged across each trial.

Each model, defined in table 1, was fit to its description in their respective publications. Model performance was then compared using R^2, standard error (ms), and an F-test. Each model parameter was also analyzed for its individual contribution using semi-partial correlation, standard error, and a t-test. Additional details on the experiment and analysis procedure can be found in the original study [8], the supplemental materials, and at `osf.io/awtvq`.

6.2 Implementation

The experiment space contained a chair — a simple prefab distributed in the VRChat SDK which allows characters to sit and be anchored to a fixed point in space. In front of the chair was a target cube and beside it the cursor. Also in the room were three buttons: start the tutorial, start the experiment, and respawn the cursor if it got lost. Three text canvases on a wall contained the instructions, the log of the results, and a detailed log of every event triggered in the room. The result log was one of the most important components, as it allowed us to retrieve the experiment results.

The cursor was a red cube that floats slightly distanced from the participant's hand. This allowed the participants to not have their field of view occluded by their avatar's VR hands and arms, while still performing the same movements as the original experiment. Rotation and any kind of physics on the cursor are disabled.

The original experiment stated that they considered 300 VR units as 1 meter. We scaled everything to Unity's measurement system, in which 1 unit is 1 meter, so that the target, the cursor, and the array of possible positions would maintain the same proportions.

It is important to keep in mind that, while multiple participants and researchers can be in the same instance of a room, their clients are separate and any changes on one client needs to be manually synchronized to the others. Only a small number of elements in VRChat are automatically synchronized. E.g., movements of the avatars. Everything else needs to be synchronized manually.

Remote function calls can be done by using the Udon SDK's `sendCustomNetworkEvent` function. But no parameters can be passed, so sharing variables has to be done another way. Udon offers a way to declare a variable as synced between two clients using the `[UdonSynced]` attribute in C# before the declaration of a variable[5]. The variable is then synced between the two clients every time they exchange data over the network. However, only limited data can be exchanged each time (126 characters), and the developer needs to take into account packet loss.

Thus we prioritized syncing the results at the expense of other variables such as the position of the target. This resulted in the researcher being able to observe the participant moving and the log printed after the trial, but not the target itself moving. This, though, did not present a problem for running the experiment or for recording the experiment data. Given the maximum synced string size of 126 characters, we decided to only synchronize the suffix of the entire log. All the previous results were stored locally every time. In retrospect, using some of those 126 characters to implement redundancy or error checking would have avoided some data lost due to packet loss. This cost us 1% of the data we collected.

VRChat also has a concept of ownership over in-game objects. Only the owner of an object can set synced variables associated with it. Ownership can change at runtime and can be linked to an in-game event, e.g., pressing a button. In our implementation, whoever presses the "init tutorial" or "init experiment" button becomes the experiment owner. Thus, variables are set by the participant's client, and the researcher's client only acts as a data receiver.

6.3 Participants

We recruited 23 participants for this experiment. Each participant was given a 30-minute timeslot — even though the experiment lasts around 10 minutes, we wanted to take into account the possible added overhead for setting up their VR equipment. Each participant was paid $15 USD for their time.

[5] https://ask.vrchat.com/t/how-to-send-data-over-the-network-using-synced-variables/2264

Table 2: Comparison of the models evaluated by Clark et al. in person [8] (their Table 3) and our replication in VRChat. Parameters: index of difficulty (ID), inclination angle (θ), azimuth angle (φ), change in target depth (CTD), target size (S). Measures: standard error (SE), semi-partial correlation (sr), p-value from the t-test, coefficient of determination (R^2), standard error (ms) (SE (ms)), and adjusted coefficient of determination (Adj. R^2).

Model	Parameters	Estimate		SE		sr		p		R^2		SE (ms)		Adj. R^2
		original	our result	original	our result	original	our result	original	our result	original	our result	original	our result	our result
Hoffmann (Fitts') [14]	Intercept	804.287	461.674	123.000	21.963	-	-	-	-	0.501	0.780	427.450	83.113	0.778
	ID	411.940	123.752	42.440	6.920	0.708	0.883	<0.001	<0.001	-	-	-	-	-
Murata & Iwase [22]	Intercept	805.404	426.211	123.907	21.691	-	-	-	-	0.501	0.788	429.697	82.077	0.783
	ID	411.760	123.173	42.678	6.842	0.707	0.878	<0.001	<0.001	-	-	-	-	-
	Sin(φ)	8.506	-23.678	61.130	13.065	0.010	-0.088	0.890	0.073	-	-	-	-	-
Cha & Myung [7]	Intercept	819.971	458.177	152.789	20.898	-	-	-	-	0.501	0.806	431.963	78.897	0.800
	ID	411.846	124.293	42.060	6.588	0.707	0.885	<0.001	<0.001	-	-	-	-	-
	Sin(φ)	6.429	-21.950	62.735	12.573	0.008	-0.082	0.919	0.084	-	-	-	-	-
	θ	-0.290	0.557	1.759	0.193	-0.012	0.089	0.135	0.005	-	-	-	-	-
Machuca & Stuerzlinger [3]	Intercept	795.413	458.381	118.073	21.190	-	-	-	-	0.545	0.798	410.202	80.070	0.794
	ID	415.547	124.640	40.742	6.674	0.713	0.889	<0.001	<0.001	-	-	-	-	-
	CTD	3.505	0.651	1.164	0.230	0.211	0.134	0.003	0.006	-	-	-	-	-
Clark et al. [8]	Intercept	732.369	454.952	105.456	21.286	-	-	-	-	0.645	0.802	364.162	79.790	0.795
	ID	433.127	125.566	36.319	6.690	0.741	0.890	<0.001	<0.001	-	-	-	-	-
	θ	-2.681	0.373	1.170	0.277	-0.142	0.064	0.024	0.181	-	-	-	-	-
	S x θ	0.336	0.012	0.060	0.012	0.346	0.048	<0.001	0.313	-	-	-	-	-

6.4 Procedure

Participants were asked to point to and select a series of targets, which then changed size and position. The original experiment had participants move a virtual cursor to the target cube's position and then select it to complete a trial. In our version, participants held a virtual cursor in their avatar's hand, so that the movement required to bring the cursor to the target would be the same. A researcher was present in the same virtual room at all times.

Participants were initially asked to change their avatar to a standard one (as seen in fig. 2). They all used the same avatar so no difference in avatars would influence the results. This avatar was tested previously to ensure its height and arm length adequately fit the experiment, and participants raised no concerns regarding the avatar. Participants were then asked to play seated and with their avatar siting in the virtual chair. This was to help them keep the same position and point of view over the trials and put the cursor and trial targets in reach. The researcher then explained to them how the experiment would work. The instructions were printed on a wall for participants to read as well, in case they could not hear the researcher's microphone for any reason or needed a reminder. As in the original experiment, participants first completed a 10-step tutorial, followed by the actual experiment which comprised 96 trials. Each trial had participants first select an orange cube positioned in the center of the matrix of possible target positions.

After every trial, the trial index, timing, target position and target size would be printed on the result log. At the end of the 96 trials, participants were asked if they had any trouble completing any of the trials or additional feedback they would like to share. Finally, the researcher would dismiss the participant and then take pictures of the result log on the wall. The pictures of the result log would then be uploaded to an OCR converter to extract the data used in the analysis. For this purpose we used a free online service that would convert the results embedded in these images into text.[6]

6.5 Results

The results from our study and the original study can be seen in table 2. We compared the model performance using adjusted R^2 to fairly account for models with additional parameters. All the tested models significantly predict movement time (alpha=0.05, $p < 0.001$), and could explain 77%-80% of the variance according to adjusted R^2. ID was consistently the parameter that contributed most to movement time prediction (alpha=0.05, $p < 0.001$). Other parameters Sin(φ), θ, and CTD were also found to be significant predictors (alpha=0.05). These parameters helped 3D specific models to outperform traditional Fitts' Law up to 3% ($Adj.R^2$). The model presented by Cha and Myung [7] was overall the best predictor (Adj. $R^2 = 0.800$), followed closely by Clark et al. [8] (Adj. $R^2 = 0.795$) and Machuca and Stuerzlinger [3] (Adj. $R^2 = 0.794$).

Our results clearly support effect (1) — ID was consistently the best predictor in all models in terms of partial correlation (sr). Our results also support effect (2), showing that all models are effective predictors of movement time. However, we do observe differences in the performance and effect size between models. The original study observed a 10% difference in performance between the empirical model and the next best one (Machuca & Stuerzlinger [3]), while we found the largest effect size between models (after accounting for multiple predictors) to only be 3%. Additionally, our results support the empirical model's generalizability to other populations as the difference between it and our observed best model (Cha & Myung [7]) is only 1%. We believe that the differences between our studies can at least in part be attributed to the VR expertise of

[6] https://ocr.space/

our participants, who completed the tasks roughly 2–4× faster and more consistently (lower standard error (SE)).

Our correspondence with the original authors illuminated another potential difference in our studies. They believe our study had more depth cues provided by better contrast in our virtual environment and our avatar modeled with visible limbs — which could have made the targets easier to select leading to lower variance. The original authors stated that these differences support effect (3), as they demonstrate how this effect can generalize to tasks that provide a more robust set of depth cues.

7 TABLETOP COLLABORATIVE COUPLING

Tang et al. [34] present two observational studies aimed at investigating the coupling styles of users collaborating over a tabletop display. Their results are cited extensively and inform the design of tabletop interfaces. Our transferability study is focused on their Study 2, where the stimuli was more abstract and experimental design further refined. Study 2 featured two task types, individual and compromise, and two interaction methods, filter and lens. The experimental task was to draw either one or two paths between two nodes on a graph. The graph had two data layers (i.e., cost and time) which were encoded on the graph edges using color.

The experiment and analysis procedure was developed from the published details [34], and through email communication with the original authors. Our replication plan preregistration is available at `osf.io/3sqmj`, and our analysis code, experiment code, and results data can be found in supplemental materials and at `osf.io/c2amz`.

7.1 Methods

The stimuli for this study were network graph visualizations with two data layers (price and time). Each data layer featured edges that represented a value: one, two, or three. The data layers were encoded on the visualization using sequential color scales with green for price and blue for time.

For the individual task, participants were assigned one data layer each and were asked to draw the lowest cost path between two nodes for their assigned data layer — one path for price and one path for time. For the compromise task, participants were asked to work together to draw one path that represented the lowest cost for both price and time. Participants then had access to either a filter interaction that would switch the data layer encoded on the network graph globally, or two lenses that locally displayed the selected data layer as illustrated in fig. 1b. The lenses and filters are designed to exacerbate the modes of interaction that participants can adopt: individual lenses (which allow local data layer changes) might work best when two people are working on different problems, but a global data layer switch (the filter) can instead encourage the participants to collaborate more, as they need to manage access to the data layers cooperatively.

Following the original experimental design, we conducted a within-subjects study with a 2 (filter vs. lenses) X 2 (individual vs compromise) design. The presentation order of the experimental conditions for groups were counterbalanced using a Latin square. Additionally, the edge weights and destination nodes were randomized for each experimental condition. Participants were instructed to talk aloud as they worked, and each session was recorded from two angles (overhead and experimenter point of view). The recordings were reviewed and analyzed for the codes and table arrangements observed in the original study, which can be seen in fig. 3. For additional details, please refer to the original publication [34], our supplemental material, and our online supplement at `osf.io/c2amz`.

7.2 Implementation

The experiment was implemented as a private VRChat world. It consisted of a small office-like environment with a table at the center, which was approximately proportional to the tabletop display in the original study. We manually recreated the exact graph used in the original study and displayed it over the virtual table. The filter interaction was implemented with two buttons, one on either side of the table. The button would switch between a green or blue graph to represent the separate data layers. The lens interaction was implemented using world space textures. As the lens was moved around the world space, it displayed different sections of the graph. Each lens also featured buttons to change the displayed data layer and to increase or reduce its size. The experimental condition could be changed with buttons placed at the back of the room.

We also imported an existing Udon 3D pen implementation[7] for participants to draw with. In order to ensure we stayed as faithful as possible to the original experiment, we shared a demo video with the authors to review our implementation. The authors agreed that we had created a faithful implementation with only a few notable differences [33]. E.g., our lenses were more "fluid" than the original, as they could be angled, moved off the table, and overlap with each other. Additionally, participants could partially walk through our virtual table to make it easier to reach the center. These differences were implemented in an effort to balance faithfulness to the original experiment and the affordances of VR interactions.

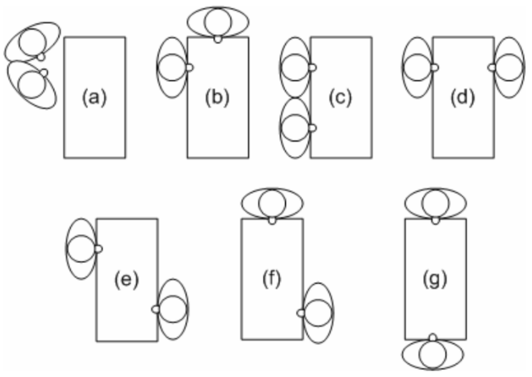
7.3 Participants

In this experiment, 8 participants were recruited to work in 4 pairs. After completing the consent and survey form, participants could choose one of four scheduled experiment times. Each participant pair was given a 1 hour time slot for this experiment and the average experiment time was approximately 40 minutes. Each participant was compensated for their time with the choice of an Amazon or Target gift card worth $25 USD.

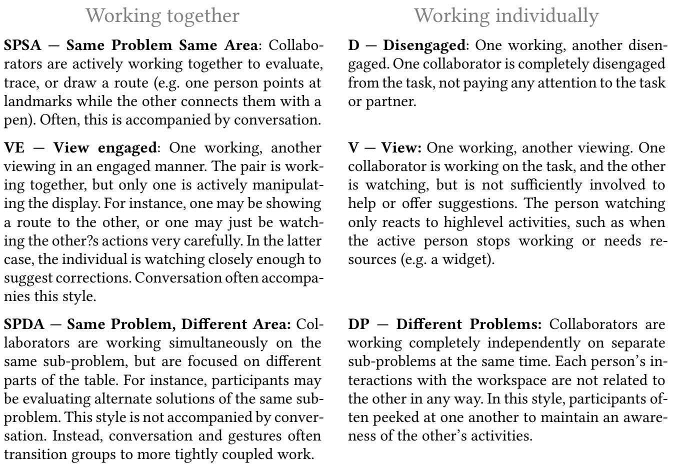
7.4 Procedure

Each participant was sent a reminder email 30 minutes before their scheduled experiment time. This email requested that they add our VRChat account to their friends list and to be ready in-game at their scheduled time. Once online, participants received an invite from our account to join our private world. Two researchers were present at all times in the virtual room with the participants — one explaining the steps of the experiment and the other managing the overhead video recording. When both participants were in our world, they were asked to change their avatars to the default VRChat robot and given a brief overview of the experiment. Instructions for each experimental condition were given before the condition began. Participants were asked to talk aloud during the experiment and to use the provided tools in anyway they deemed fit.

[7] https://booth.pm/ja/items/1555789



Position arrangements reported in Tang et al.'s Fig.5: (a) together, (b) kitty corner, (c) side by side, (d) straight across, (e) angle across, (f) end side, and (g) opposite ends.



Working together

SPSA — Same Problem Same Area: Collaborators are actively working together to evaluate, trace, or draw a route (e.g. one person points at landmarks while the other connects them with a pen). Often, this is accompanied by conversation.

VE — View engaged: One working, another viewing in an engaged manner. The pair is working together, but only one is actively manipulating the display. For instance, one may be showing a route to the other, or one may just be watching the other?s actions very carefully. In the latter case, the individual is watching closely enough to suggest corrections. Conversation often accompanies this style.

SPDA — Same Problem, Different Area: Collaborators are working simultaneously on the same sub-problem, but are focused on different parts of the table. For instance, participants may be evaluating alternate solutions of the same sub-problem. This style is not accompanied by conversation. Instead, conversation and gestures often transition groups to more tightly coupled work.

Working individually

D — Disengaged: One working, another disengaged. One collaborator is completely disengaged from the task, not paying any attention to the task or partner.

V — View: One working, another viewing. One collaborator is working on the task, and the other is watching, but is not sufficiently involved to help or offer suggestions. The person watching only reacts to highlevel activities, such as when the active person stops working or needs resources (e.g. a widget).

DP — Different Problems: Collaborators are working completely independently on separate sub-problems at the same time. Each person's interactions with the workspace are not related to the other in any way. In this style, participants often peeked at one another to maintain an awareness of the other's activities.

Coupling styles reported by Tang et al. (pp.1187).

Figure 3: Position arrangements and coupling styles in Tang et al. [34].

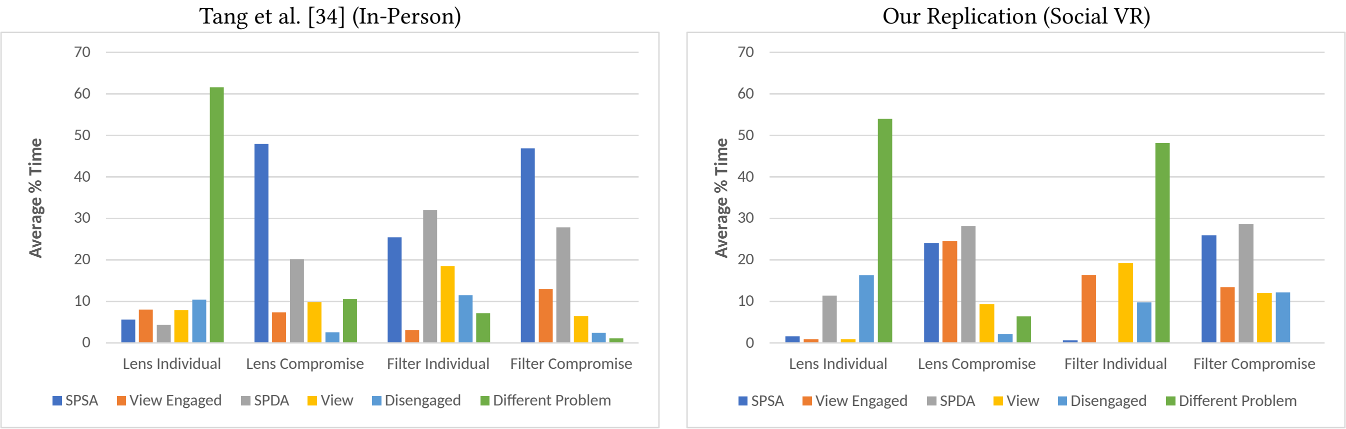


Figure 4: Coupling styles in each condition observed by Tang et al. in person [34] (their Fig. 4) and in our VRChat replication. The two experiments share some interesting commonalities, considering the fact that the first was run in person, and the second was run in VR. This suggests that people do have similar behaviors in person and in VR. Notably, the Lens Individual setup is very similar, with participants mostly working on Different Problems in both experiments. The Lens Compromise and Filter Compromise setups had participants spending more time in the Same Problem Same Area (SPSA) and Same Problem Different Area (SPDA) coupling styles in both cases — as expected in situations which require more collaboration.

7.5 Results

After reviewing the recorded sessions, we observed participant coupling styles consistent with those reported by Tang et al. [34]. Our replicated results on average time for each coupling style and experimental condition are in fig. 4. As in the original study, participants were more tightly coupled when working on compromise tasks and global filters, compared to individual tasks and local lenses.

In the Lens Individual condition participants worked separately on their respective sub-problem 54% of the time. Participants also spent most of their time (48%) on different problems during the Filter Individual task. However, participants during this trial spent more time viewing than actively working on the display compared to the Lens Individual condition. Participants also spent less time working together overall during this task compared to the original study. Instead of helping with each other's sub-problems, participants could often been seen looking for the shortest path or doing math while waiting to view their data layer (fig. 5 ❶). During the compromise task, participants spent most of their time tightly coupled. For both lens and filter conditions participants spent 28% of their time working on the same problem and same space. Only one of our groups used the lenses to work on the compromise task in parallel — starting on opposite ends of the board. The other three groups worked together on this task by enlarging one lens to cover the whole table and using it as a global filter instead (fig. 5 ❺). Through our communication with the authors, we know that this behavior was also observed in the original study [33].

We also analyzed the session recordings for participant table arrangements, broken down by coupling style (table 3). Unlike the original study [34], we did not observe participants consistently standing closer together when working collaboratively. Rather, in the virtual environment the table arrangement of participants related more to access to the table instead of proximity to collaborators. When working closely with each other, participants were most often arranged kitty corner (fig. 3 b), side by side (fig. 3 c), or end side (fig. 3 f). These positions afford participants equal access to the same parts of the table (fig. 5 ③ ④). When working loosely together or not at all, participants worked in arrangements that were most convenient for the parts of the table they needed to access.

Participants rarely interacted with areas of the table immediately near their collaborators. When participants wanted to show something to their collaborators, they often indicated the location with a gesture and then moved out of the way. These behaviors closely resemble territorial behavior observed in the original study.

Interference between participants was rarely observed. In VR, the proximity of another person feels much less intrusive, and the only real interference you can cause is occluding another participant's field of view. That can be easily avoided, as participants can use teleport locomotion to move to other areas of the room. As a result of these factors, it is trivial for experienced VRChat users to avoid interfering with one another. The only instances of interference we observed came when an inexperienced user was struggling with the controls, or when participants briefly overlapped their lenses.

We observed user behavior in line with all six of the coupling styles (fig. 3) described in the original study [34], supporting effect (1). We also did not observe any behavior outside of these styles. As in the original study, we participants were more tightly coupled during the compromise task and while using perspective sharing global filters, supporting effects (2, 3). The differences observed in the proportion of coupling styles and table arrangements can largely be explained by the change in context (such as avatars not being as expressive as humans and users needing to rely on VR locomotion instead of in-person collaboration), individual differences in participants (working more independently during individual conditions), and by different researchers performing the coding. Our results show that several of the findings of the original study, such as perspective sharing being important for tabletop collaboration, are transferable to the new Social VR context.

Table 3: Percentage of time spent working in each coupling style and physical arrangement observed by Tang et al. in person [34] (their Table 1) and in our VRChat replication.

White: < 1%
░ Light grey: 1-4%
▒ Dark grey: > 4%

	SPSA	VE	SPDA	V	D	DP
	Tang et al. [34] (In-Person)					
(a) Together	7.8	1.6	3.4	0.5	0.2	0.5
(b) Kitty corner	9.4	1.9	5.2	2.4	0.9	1.9
(c) Side by side	2.5	1.0	2.3	0.9	0.9	3.1
(d) Straight across	9.2	2.3	8.7	3.3	2.3	1.0
(e) Angle across	3.8	1.4	2.4	2.3	1.4	6.2
(f) End side	0.5	0.1	0.1	0.3	0.3	4.9
(g) Opposite sides	0.0	0.0	0.0	0.1	0.0	3.1
	Our Replication (Social VR)					
(a) Together	0.5	0.6	0	0	0	0
(b) Kitty corner	8.3	12	1.4	4.1	1	0.9
(c) Side by side	2.8	4.3	1.2	1.9	1	0.8
(d) Straight across	3.9	3.6	4.9	1.6	0.8	3.7
(e) Angle across	1.4	0.8	6.6	1.4	1.9	2.2
(f) End side	5.5	3.9	7.9	8.2	2.7	8.4
(g) Opposite sides	0.8	1.7	6.7	2.4	0	2.5

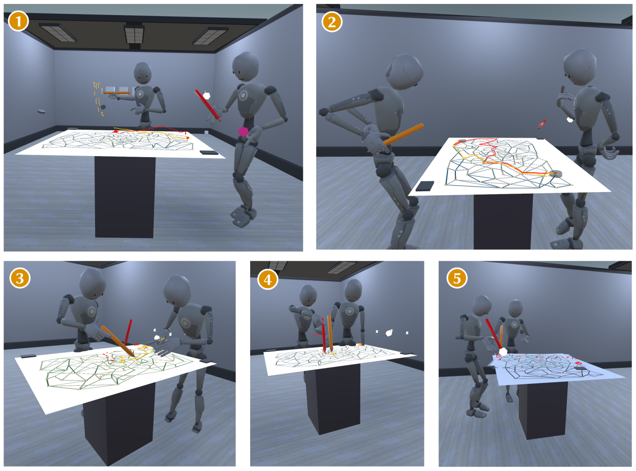


Figure 5: Select screenshots taken in VRChat during the experiment. ① One participant annotating and counting, while the other works independently on their sub-problem. ② Participants exhibiting typical thinking mannerisms (hip holding, foot tapping, hand on face) while checking their paths. ③ Participants closely working together, one drawing while the other guides. ④ Participants simultaneously investigating the same area of the graph. ⑤ Participants creating one large lens and using it as a global filter.

8 DISCUSSION

In this section we will reflect on our experience conducting our studies on a social VR platform. We will begin by discussing our experience broadly, detailing the aspects we found most interesting or helpful. Then, we will discuss aspects of the studies that we would approach differently in hindsight. Next, we summarize these findings into a general guide for conducting research on social VR platforms. Finally, we present a discussion on the future of this method and remote VR studies.

8.1 Conducting Remote Studies via Social VR

After completing these studies we can confidently say that social VR platforms can be used to conduct remote and collaborative VR studies and collect both quantitative and qualitative observations. These studies were safely conducted synchronously, during a global pandemic, by two experimenters working remotely in different cities. The recruitment process was surprisingly easy, taking only a day for us to exceed our recruitment goal of at least 23 participants. In total we had 26 participants sign up and 23 participants who made their scheduled meeting time(s). All of our participants were genuinely nice, respectful, and often interested in our research. Many participants inquired further about the study they had just participated in, and some even went further providing us with development tips and design feedback. It is important to note that, because of how the recruitment was performed, we had access to a pool of mostly expert VRChat users, and people who were genuinely interested in the nature of our research. Developing for VRChat with Udon was made easier by the amount of community resources that exist for both Udon and Unity. Researchers who are already experienced with Unity development will have an especially easy time developing for VRChat.

8.1.1 Potential advantages of conducting remote VR studies. We believe this method can open up new opportunities for VR research that was not possible before — or at least very hard to achieve. For example, this setup makes it possible to run experiments including a large number of VR users at the same time without the need for a lab to own a large number of VR headsets and their respective setups. At the time of writing this paper, VRChat worlds have a limit of 30 users per instance. If a researcher wanted to run a VR experiment with multiple participants at the same time, without the option of running a remote study, they would have to have as many HMDs — a high barrier of entry for collaborative VR research. Additionally, using an established platform largely takes care of the development overhead required to implement the networking logic and systems required for such an experiment.

8.1.2 Extending Fitts' Law in 3D VR. We were able to accurately implement the experimental task, replicating nearly identical trial data (ID, radial distance, inclination angle, etc.) to the original. In line with related studies, we demonstrated the utility of Fitts' Law and similar models for predicting movement time in three dimensions — indicating the validity of our results. We also demonstrated how experienced VR users can influence the results of a VR study. We believe our participants were much faster and consistent at the experimental task, likely as a result of their experience with VR and familiarity with the controls of their HMD.

Conducting the study synchronously allowed us to verbally explain the experiment task to participants and ensure that they understood them. It also allowed us to observe participants throughout the study and converse with them afterwards, to ensure the study was conducted correctly. Given the novelty of the experiment and the reliance required on the participants' equipment, some technical issues were to be expected. Interestingly, the expressiveness offered by the avatars in VRChat allowed us to overcome some of the technical difficulties. When a participant showed up with a malfunctioning microphone, we were still able to communicate with them by having them say "yes" or "no" by just moving their head, and they were still able to successfully complete the experiment. Additionally, many of the participants informed us after the study of a bug that would occur when the trial target spawned on top of the cursor. This bug prevented participants from selecting the target until they moved the cursor out of the target and back in. As a result of this communication, we were able to identify and remove the 4 trials affected by this bug.

8.1.3 Tabletop Collaborative Coupling. Through communication with the authors of the original study, we faithfully recreated the experiment and demonstrated the transferability of the original results to a VR context. We observed strikingly similar behavior from our participants to what was described by Tang et al. [34].

Initially, we were concerned that the virtual avatars would not be expressive enough for us to accurately analyze the coupling behavior. Our standard avatar did not have complex emotive features, only a basic mouth movement animation automatically played while users were speaking. Despite this limitation, we were surprised to observe just how expressive these avatars allowed participants to be. Still, VRChat does offer the option to implement more complex avatars with a more precise simulated mouth movement, eye tracking, and expressions. Standard avatars in VRChat have finger joints, in addition to all the major joints in the body. Non-verbal communication such as pointing or waving clearly indicated when participants were switching between loose and tight coupling styles. It was often clear if participants were engaged or simply viewing by just observing the body language of their avatar. Participants could be seen exhibiting typical thinking mannerisms such as holding their chin or hips; nodding their heads or waging their finger when counting; and tapping their feet or shuffling back and forth if they had full body tracking (fig. 5 ❷).

We rarely observed participants acting in a disengaged manner, only doing so briefly to adjust their headset or wait for their collaborator to wrap up their task. We suspect this is due to the immersion provided by a virtual environment with potentially less distractions from the real world.

Demonstrating that these results transfer to a collaborative virtual environment helps establish the foundation of future research in this area. The experimental conditions were designed to be as close as possible to the original study, which were in turn designed with the constraint of a physical tabletop display. There are many aspects of the interactions that could be improved in the virtual context. E.g., the filter interaction could occur locally or globally depending on collaborator needs. Exploring VR-specific collaborative tabletop interactions will be easier knowing that Tang et al.'s tabletop design guidelines [34] still largely hold in VR.

8.2 What We Would Do Differently

Overall, we did not encounter any critical problems conducting our studies. However, there are several things that we would consider doing differently if given the chance.

Out of an abundance of caution, we spaced out all our participant meeting times for the Fitts' Law study. These were over the course of two weeks and at least one hour apart. The study itself only lasted around 10 to 15 minutes and participants could have been scheduled closer together. We also encountered issues with client synchronization during this experiment, resulting in 1% (20 trials) of data being lost. This could have been avoided with more robust synchronization code, or implementing data redundancy. Additionally, extracting the data with OCR caused several errors that had to be resolved manually. This process could have been improved with more advanced methods of data extraction such as generating

QR codes or checksums. The Fitts' Law experiment needed so little interaction from the researcher that the full experiment could have been run asynchronously. In order to run it asynchronously, a researcher would need to access the same instance of the room as the participants to have access to the same generated data, but there would be no action required while the experiments were being run. By including more than one experiment stall in a VRChat world, multiple participants could also be run at the same time. This could make the entire process of running participants much faster and less cumbersome for the researcher, and more akin to a classic Mechanical Turk setup.

Conversely, the tabletop collaboration study was more complex and required more participation from the researcher — something that could not be easily avoided with clever tricks in the implementation. A number of minor details could have been improved to make the experience feel smoother for the participants — such as pens that only write when touching the surface of the table, instead of being able to write in the air, which allows for more space but can cause issues when looking at written text from a different perspective. Overall, though, we do not believe these small details could have influenced the results significantly, and solving them would not have made the experiment less onerous on the researcher.

8.3 Guidelines For Social VR Studies

We believe this research method will prove helpful for many VR researchers; however, we acknowledge that it may not meet the needs of every VR study. This method is best suited for conducting foundational studies (such as the Fitts' Law experiment used in our study), evaluating prototype interactions, and collaborative studies. Implementations for these social VR platforms will often be proprietary to the platform it was developed for. As a result evaluating standalone VR applications with this method is not practical or recommended.

To conduct user studies with social VR platforms as ethically as possible, we recommend adhering to the following guidelines.

Ask for approval first: Receiving IRB approval will help ensure the study procedures are ethical. When conducting users studies on a particular platform that has not already approved this practice, it is of great importance to seek explicate permission from the relevant authorities prior to the onset of the study. Social VR platforms often have community managers who will be able to assist with this step.

Recruitment: Next, it is generally a good idea to avoid recruiting participants directly in game, because direct solicitation is often against the rules and commonly seen as a nuisance. Before recruiting participants from external forums, ensure solicitation is allowed and will not be breaking any rules. When recruiting experienced VR users make sure to adequately compensate them for their time, equipment, and expertise. In this study we remunerated participants at a higher rate than we typically would to account for their expertise and the overhead of setting up their own VR equipment.

Participant privacy: Recruiting participants provides the opportunity for anonymity by allowing them to provide in-game usernames instead of real names, and only refer to participants by these usernames. If participants are added to an in-game friend list, remove them after the study is complete. Friends lists show when users are online and where they are in-game. It would be unethical to track participants on social VR platforms, just as it is unethical to track them in real life. Communicating with participants to handle consent forms, scheduling, and compensation is best done through email. Communication on the platform should be limited as much as possible to matters related to the experimental task.

Challenges given by the platform: Before beginning to implement a VR study, familiarize yourself with the capabilities and limitations of the targeted platform. With these factors in mind, consider the variables that will need to be controlled. For example, should players be seated or standing? What locomotion method should be used? How big of a play space do participants need? Should participants all use the same avatar? What would be the most suitable and inclusive avatar? When these factors are hard to control for, the study should be designed to account for potential differences. In general we would recommend designing the studies in a way that accommodates the most users possible unless specific user population requirements are absolutely necessary.

Experiment design: The experimental task should be easy for participants to understand, and should be designed around the affordances of the platform they are conducted on. Participants should be given brief tutorials before the experiment begins to ensure they understand the required task. The experiment should not last more than roughly one hour to avoid fatiguing participants. Additionally, the experimental task and stimuli should be designed to avoid common causes of VR motion sickness (e.g., rapid movement and virtual locomotion). Instruct participants to inform experimenters if they begin to experience motion sickness or any discomfort — so that the study may be stopped.

Data retrieval: If data cannot be automatically saved outside of the platform, it should be kept as simple as possible. E.g., our data consisted of either recorded videos or simple comma separated lists. Once this data was extracted from the platform, it was further analyzed to generate the final study results. If the data is being manually extracted, implement methods for validating the data, such as a checksum. Once the study is complete, ensure the data is fully anonymized — removing usernames, identifiable audio, and other personal information.

8.4 VRChat-specific recommendations

Observing the participants: In VRChat, in order to synchronously witness an experiment, the researcher's avatar needs to be in the same virtual room as the participants. For some experiments, if participants are aware they are being observed it could influence their actions. This can be mitigated to some degree with clever virtual world design. The VRChat SDK allows for virtual cameras to display views onto world objects — a technique typically used to create mirrors. A researcher could use these cameras to observe the participants from another location out of participant's field of view. However, participants would still be aware that the researcher is present somewhere in the virtual environment.

Extracting data from VRChat: As of 2020-12, there is no officially supported way to programmatically serialize data in or out of the VRChat client. To work around this limitation we used video recordings and in game screen shots to extract our data. These methods worked well for our relatively simple data, however they would not scale well with more complex data that can be captured within VRChat — such as tracking and movement data. Data that can be obtained through the VRChat API can be seen in the UdonSharp documentation [8]. We believe it would be possible to extract more complex data with methods such as a QR code or Morse code generator, still this creates more overhead for the researcher. It is possible that VRChat will eventually implement a solution for this — HTTP request functionality is a highly requested feature among users.

This comes back to the debate between more creative freedom versus more administrative control. There are several user privacy and safety concerns around allowing data to be sent and received outside the VRChat client and servers. Current and future social VR platforms may be able to address this by adopting a robust user consent model. Such models can be seen on mobile devices where applications must receive consent from users before accessing their data. This concept can be translated to a virtual environment where users will have to consent before data can be sent or received.

8.5 The Future of Remote VR Studies and Social VR

The need for remote research methods for VR studies will continue to increase along side the democratization of VR devices and platforms. We have demonstrated that social VR platforms can be used to at least partially meet this need. However, social VR platforms like VRChat can change overtime or even become discontinued. Aspects that make VR studies possible, such as user privacy or custom content, are all subject to change as well. This calls into question the long term feasibility of this research method. We believe there are two possible paths that can be taken to address this.

The first path is to partner directly with a social VR platform to create official channels for conducting VR studies. For example, VRChat has a menu for browsing different user worlds organized by category. One such category could be dedicated to ongoing VR studies — where interested users could sign up and participate directly through the platform. Additionally, existing social VR platforms could allow researchers to obtain a licenses for conducting research on the platform. This licenses could allow for additional SDK features, such as HTTP requests, to be made available to licensees. This would create a system more akin to traditional crowdsourcing platforms (i.e., MTurk) and could be set up with more guarantees for user privacy and platform longevity.

The second path would be for the VR research community to create a social VR platform explicitly made for conducting user studies. This platform would follow the same basic architecture of other social VR platforms. Mainly, it would provide researchers with a platform on which to build their VR studies — taking care of the overhead of application distribution and distributed networking. By limiting the custom content on the platform to only approved researchers, researchers could be granted more creative freedom with access to all features of a common programming language or game engine. Creating a platform like this will take a significant amount of effort, time, and funding. Other proposed long term solutions, such as hardware seeding [31], also face the same issue. However, we believe that a custom social VR platform presents the greatest long term benefit for researchers seeking to conduct remote, synchronous, and collaborative studies — making them easier to implement and conduct, just as we experienced in this study using VRChat.

9 CONCLUSION

We have demonstrated how social VR platforms can be used to implement remote, synchronous, and collaborative VR experiments. We successfully replicated two studies, an individual study based on extending Fitts' Law in VR, and a collaborative one that required participants to find a path in a network. All the studies were carried out remotely on VRChat, a popular social VR platform that allows developers to upload custom worlds containing custom code. We show how we leveraged the advantages offered by this method, such as being able to conduct the experiment remotely, recruiting participants with access to HMDs, and conducting the study in synchronous, distributed virtual environments. The results we obtained are comparable to those in the original studies, and demonstrate the practicality and validity of this research method.

We believe this method can open up new possibilities in how VR experiments are performed and for what experiments are viable, such as large scale collaborative VR studies. Social VR platforms present researchers with the opportunity to more easily conduct remote, synchronous, collaborative user studies than otherwise possible. By demonstrating their efficacy, we hope that future researchers will be able to utilize these social VR platforms to further the cutting edge in VR research.

ACKNOWLEDGMENTS

We graciously thank members of the Data Visualization @ Khoury lab for their feedback, Giorgio Severi and Mary Makarious for pilot testing, and the VRChat staff and community for their assistance.

REFERENCES

[1] Fouad Alallah, Ali Neshati, Nima Sheibani, Yumiko Sakamoto, Andrea Bunt, Pourang Irani, and Khalad Hasan. 2018. Crowdsourcing vs Laboratory-Style Social Acceptability Studies? Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173884

[2] Sara Albakry, Kami Vaniea, and Maria K. Wolters. 2020. What is This URL's Destination? Empirical Evaluation of Users' URL Reading. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376168

[3] Mayra Donaji Barrera Machuca and Wolfgang Stuerzlinger. 2019. The Effect of Stereo Display Deficiencies on Virtual Hand Pointing. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300437

[4] Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368. https://doi.org/10.1093/pan/mpr057

[5] Conrad Boton. 2018. Supporting constructability analysis meetings with Immersive Virtual Reality-based collaborative BIM 4D simulation. *Automation in Construction* 96 (2018), 1 – 15. https://doi.org/10.1016/j.autcon.2018.08.020

[6] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1 (23 8 2011), 3–5. https://doi.org/10.1177/1745691610393980

[7] Yeonjoo Cha and Rohae Myung. 2013. Extended Fitts' law for 3D pointing tasks using 3D target arrangements. *International Journal of Industrial Ergonomics* 43, 4 (2013), 350 – 355. https://doi.org/10.1016/j.ergon.2013.05.005

[8] Logan D. Clark, Aakash B. Bhagat, and Sara L. Riggs. 2020. Extending Fitts' law in three-dimensional virtual environments with current low-cost virtual reality

[8] https://github.com/MerlinVR/UdonSharp/wiki/vrchat-api

technology. *International Journal of Human-Computer Studies* 139 (2020), 102413. https://doi.org/10.1016/j.ijhcs.2020.102413

[9] William S. Cleveland and Robert McGill. 1984. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554. https://doi.org/10.1080/01621459.1984.10478080

[10] Arindam Dey, Thammathip Piumsomboon, Youngho Lee, and Mark Billinghurst. 2017. Effects of Sharing Physiological States of Players in a Collaborative Virtual Reality Gameplay. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4045–4056. https://doi.org/10.1145/3025453.3026028

[11] Sara Di Bartolomeo, Aditeya Pandey, Aristotelis Leventidis, David Saffo, Uzma Haque Syeda, Elin Carstensdottir, Magy Seif El-Nasr, Michelle A. Borkin, and Cody Dunne. 2020. Evaluating the Effect of Timeline Shape on Visualization Task Performance. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376237

[12] Leah Findlater, Joan Zhang, Jon E. Froehlich, and Karyn Moffatt. 2017. Differences in Crowdsourced vs. Lab-Based Mobile and Desktop Input Performance Data. In *Proc. 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6813–6824. https://doi.org/10.1145/3025453.3025820

[13] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

[14] Errol R. Hoffman. 1995. Effective target tolerance in an inverted Fitts task. *Ergonomics* 38, 4 (1995), 828–836. https://doi.org/10.1080/00140139508925153 arXiv:https://doi.org/10.1080/00140139508925153

[15] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the Perception of Machine Teaching. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376428

[16] Bernd Huber and Krzysztof Z. Gajos. 2020. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *PLOS ONE* 15, 1 (01 2020), 1–17. https://doi.org/10.1371/journal.pone.0227629

[17] Aristotelis Leventidis, Jiahui Zhang, Cody Dunne, Wolfgang Gatterbauer, H.V. Jagadish, and Mirek Riedewald. 2020. QueryVis: Logic-Based Diagrams Help Users Understand Complicated SQL Queries Faster. In *Proc. 2020 ACM SIGMOD International Conference on Management of Data* (Portland, OR, USA) *(SIGMOD '20)*. 2303–2318. https://doi.org/10.1145/3318464.3389767

[18] Xiao Ma, Megan Cackett, Leslie Park, Eric Chien, and Mor Naaman. 2018. Web-Based VR Experiments Powered by the Crowd. In *Proc. 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 33–43. https://doi.org/10.1145/3178876.3186034

[19] Maya Mathur and Tyler VanderWeele. 2019. Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges and Schauer (2019). *Psychological Methods* 24 (10 2019), 571–575. https://doi.org/10.1037/met0000223

[20] MerlinVR. 2020. UdonSharp. https://github.com/MerlinVR/UdonSharp.

[21] Aske Mottelson and Kasper Hornbundefinedk. 2017. Virtual Reality Studies Outside the Laboratory. In *Proc. 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. Association for Computing Machinery, New York, NY, USA, Article 9, 10 pages. https://doi.org/10.1145/3139131.3139141

[22] Atsuo Murata and Hirokazu Iwase. 2001. Extending Fitts' law to a three-dimensional pointing task. *Human Movement Science* 20, 6 (2001), 791 – 805. https://doi.org/10.1016/S0167-9457(01)00058-6

[23] Oculus. 2020. A Single Way to Log Into Oculus and Unlock Social Features. https://www.oculus.com/blog/a-single-way-to-log-into-oculus-and-unlock-social-features/

[24] G. Paolacci, J. Chandler, and P.G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://ssrn.com/abstract=1626226

[25] David G. Rand. 2012. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology* 299 (2012), 172 – 179. https://doi.org/10.1016/j.jtbi.2011.03.004 Evolution of Cooperation.

[26] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

[27] David Saffo, Caglar Yildirim, Sara Di Bartolomeo, and Cody Dunne. 2020. Crowdsourcing Virtual Reality Experiments Using VRChat. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382829

[28] SCRN-VRC. 2020. SimpNet-Deep-Learning-in-a-Shader. https://github.com/SCRN-VRC/SimpNet-Deep-Learning-in-a-Shader.

[29] Danielle N. Shapiro, Jesse Chandler, and Pam A. Mueller. 2013. Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science* 1, 2 (2013), 213–220. https://doi.org/10.1177/2167702612469015 arXiv:https://doi.org/10.1177/2167702612469015

[30] A. Steed, Sebastian Friston, María Murcia-López, Jason Drummond, Ye Pan, and D. Swapp. 2016. An 'In the Wild' Experiment on Presence and Embodiment using Consumer Virtual Reality Equipment. *IEEE Transactions on Visualization and Computer Graphics* 22 (2016), 1406–1414.

[31] Anthony Steed, Francisco R. Ortega, Adam S. Williams, Ernst Kruijff, Wolfgang Stuerzlinger, Anil Ufuk Batmaz, Andrea Stevenson Won, Evan Suma Rosenberg, Adalberto L. Simeone, and Aleshia Hayes. 2020. Evaluating Immersive Experiences during Covid-19 and Beyond. *Interactions* 27, 4 (2020), 62–67. https://doi.org/10.1145/3406098

[32] suzuki_i. 2020. Portal Gun (UDON). https://vrchat.com/home/launch?worldId=wrld_b0396e28-6d1c-4b8a-86b6-31b8389154e5.

[33] Anthony Tang. 2020.

[34] Anthony Tang, Melanie Tory, Barry Po, Petra Neumann, and Sheelagh Carpendale. 2006. Collaborative Coupling over Tabletop Displays. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1181–1190. https://doi.org/10.1145/1124772.1124950

[35] Varneon. 2020. Extended clip from my upcoming monthly development vlog showing the #UdonSharp Cubic Bezier Curve I did few weeks ago. #VRChat #MadeWithUdon. https://twitter.com/VarneonOfficial/status/1300438000745078786.

[36] @VRChat. 2019. Thanks to our community for making 2018 #VRChat's best year yet! Tweet. Retrieved 02 Nov 2019 from https://twitter.com/VRChat/status/1086389685268635648.

[37] VRChat. 2019. VRChat Partners with HTC and Makers Fund to Close $10m Series C. https://medium.com/vrchat/vrchat-partners-with-htc-and-makers-fund-to-close-10m-series-c-50ea04d5ca23

[38] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376642

A ADDITIONAL FIGURES AND TABLES

Table 4: Participants self-reported device and demographic information.

Participant Information	# count	% total
HMDs		
HTC Vive	**7**	**30.4%**
Valve Index	5	21.7%
Oculus Rift S	3	13.0%
HTC Vive Pro	3	13.0%
Oculus Rift	2	8.7%
Samsung Odyssey +	2	8.7%
Oculus Quest	1	4.3%
Demographics (optional)		
Gender		
Man	**18**	**81.8%**
Woman	2	9.1%
Non-binary/Third Gender	2	9.1%
Education		
Some college but no degree	**7**	**31.8%**
Bachelor's degree	6	27.3%
High school graduate or equivalent	5	22.7%
Less than high school degree	2	9.1%
Associate degree	1	4.5%
Master's degree	1	4.5%
Employment		
Working (paid employee)	**11**	**50.0%**
Student	5	22.7%
Not working (looking for work)	4	18.2%
Not working (temporary layoff)	2	9.1%
Continent		
North America	**16**	**68.8%**
Europe	4	25.0%
Asia	1	6.3%

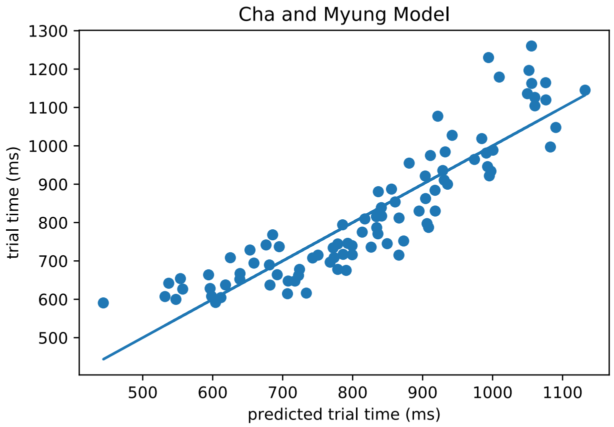


Figure 6: Model fit for Cha & Myung [7].

Table 5: Participants self-reported VR experience and usage.

White: < 25%
░ Light grey: 25-50%
▓ Dark grey: > 50%

	Strongly disagree	**Somewhat disagree**	**Neutral**	**Somewhat agree**	**Strongly agree**
I am experienced with using VR	1	0	0	6	16
I am an experienced VRChat user	2	1	2	4	14
I easily get motion sickness in VR	16	4	1	1	1

	Once a day	**Once a week**	**Once a month**	**Once a year**	**Never**
How often do you use VR?	7	12	3	1	0
How often do you play VRChat?	8	11	2	1	1
How often do you play videogames, both in VR and otherwise?	19	3	1	0	0