VIRTUAL REALITY TRACKING DATA: INSIGHTS, RISKS, OPPORTUNITIES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mark Roman Miller
June 2023

This dissertation is online at: https://purl.stanford.edu/ft715kr0160

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jeremy Bailenson, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

James Landay, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Maneesh Agrawala

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Nilam Ram

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Abstract

Much has been studied concerning the use of virtual reality (VR) as a display. However, relatively less has been performed with VR as a sensing tool, or as I lay out in this dissertation, an *observatory*. Using three studies, I demonstrate that the collection of head and hands pose data in current consumer VR, an instantiation of the paradigm of VR as an observatory, enables collection of body motion data in high temporal and spatial fidelity in large groups which is valuable for quantitative analysis regarding the study of nonverbal communication and poses risks to immersant privacy through biometrics. The three studies are (1) a study of person-to-person interactions through gaze and proxemics, (2) a study of synchrony measurement by exploring content validity, consistency, and predictive validity among 9300 measures of synchrony, (3) a demonstration of the identifiability of this data, and the effects of time upon identifiability.

Acknowledgments

A dissertation is the culmination of much work, and work is never done alone.

Beginning with my research community, I'd like to thank Jeremy Bailenson first as my advisor and mentor through this PhD. I hope to pass along what I've received as a researcher and a person to the students I will advise in the coming decades. I would like to thank James Landay for structure when I need it and flexibility when I don't. I would also like to thank the rest of my committee, Maneesh Agrawala for also having the opportunity to teach in the HCI Frontiers and Foundations course to solidify my understanding of HCI, and Nilam Ram for being an ever-helpful methods master. Finally, I would like to thank Gabrielle Harari for chairing this committee and introducing me to the formal study of media as an young PhD student.

I would like to thank Ade Mabogunje for his mentorship, providence, encouragement, and ever-intriguing conversations, and Byron Reeves, whose challenge that "new methods require new theories" helped me synthesize my focus on numbers with a need for theoretical impact.

I would like to thank my labmates at the Virtual Human Interaction Lab, beginning with Ketaki Shriram, who first met with me to describe the lab, Catherine Oh, who taught me about synchrony, Fernanda Herrera, who guided me through the halls of the lab's institutional knowledge, and Hanseul Jun, my (mutual) sounding board on anything related to computer science or mathematics in the lab. I would also like to thank Anna Queiroz and Geraldine Fauville, who constantly inspired me to press through difficulty and uncertainty, Eugy Han, co-manager and co-coordinator of this amazing dataset, Cyan DeVeaux, the design expert, and Portia Wang, who has continued on my tradition of using too much math. A big thank-you to the lab managers and coordinators who make all of this run, including Tobin Asher, who gave me the lab tour when I was some random PhD student, Brian Beams, who I constantly interrupt, and Talia, Elise, and Crystal.

I would like to thank Daniel Akselrad, Ryan Moore, and Mark York for their support with the course that turned into this dataset, and thank Benjamin Liao, Casey Manning, Benjamin Martinez, Umar Patel, and Neha Vinjapuri for their help with developing the virtual environments for these studies. Finally, I would like to thank all the people I have had the chance to collaborate with over the years, including Inrak Choi, Sean Follmer, Greg Welch, Larry Leifer, Kristine Nowak, Ryan Chen, Eliot Jones, and Michael Arruza-Cruz, and all the administrative staff in the Mechanical Engineering, Computer Science, and Communication departments who have to deal with the logistics

of my choice to do interdisciplinary work.

Turning to my friends and family, I would like to thank Aldo and Daniel for many provocative conversations on the nature of virtual reality, computing, and technology, the “P90Xodus" crew of Tony, Robert, and Diogo, additional support (musical and memetic) from Eric Lebel and Jon Timcheck, and the Catholic Community at Stanford, whom I have had the opportunity to serve for several years here.

I would like to thank my father, Scott, and my mother, Annie, for their love and support from my very beginning. I owe the kindling of my love of computer science to my mother, who shares my love of the clear and logical elegance of a good program, and I owe the spark to my father, who decided to write–and explain–a program to answer a silly math question I had in primary school. I also owe the confidence to pursue a PhD to him, who was the first in his family to attend college and still continued all the way to a doctorate, and I owe my turn towards design, psychology, and communication to my mother, who instilled in me the empathy that makes this work worthwhile.

My wife, Jordan, has been my rock and guiding star, my true partner, traveling from Illinois to California, my cellmate in our tiny studio during the pandemic (oh, with Cookie, our dog, too), my hiking buddy, breadwinner, and head businesswoman, my support and reason for supporting. Thank you, Jordan.

Finally, praise and thanks to God the Father, Son, and Holy Spirit: the giver of good gifts, the one who loved me first, the unmoved mover, the hope of ultimate rest.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Virtual reality (VR), as it has been envisioned, is a medium in which a person called the *immersant* is insulated from real-world stimuli and instead experiences a coherent, interactive, computer-generated world. VR has been a research topic for more than a half-century and has recently become popular in the consumer technology space. Much of the focus on virtual reality has been on its affordances as a type of output, i.e., its ability to produce experiences that would otherwise be expensive (remote collaboration), dangerous (flight training), counterproductive (viewing environmentally protected places), or downright impossible (navigating outer space or the insides of cells). However, just as in any other interactive system, output requires input. *If a computer is to appropriately produce a coherent, interactive world as output (which becomes the immersant's sensory input), it will need to observe all activity of the immersant - the immersant's output, so to speak - as device input.* To speak in Sutherland's terminology [113], the ultimate display requires the ultimate observatory.

In this dissertation, I demonstrate that the collection of head and hands pose data in current consumer VR, an instantiation of the paradigm of VR as an observatory, enables collection of body motion data in high temporal and spatial fidelity in large groups which is valuable for quantitative analysis regarding the study of nonverbal communication and poses risks to immersant privacy through biometrics.

1.1 Head and Hands Pose Data

First, I must explain the collection that currently occurs in consumer VR headsets. To begin, consider what is necessary to produce the perception of a virtual world. One of the properties of this world that we are familiar with is that where objects appear in our vision is relative to our head position and orientation and relative to our eye orientation. If one is to simulate a virtual world like this one through a head-mounted screen, the system needs to account for the immersant's head motion so that the system can place the image of the virtual object in what is perceived as a stationary location in the world, but in fact has a changing location on the screen. The necessity of

this for realism is detailed further in section 2.3 and its technical components are discussed further in section 2.2. This is why user tracking is necessary for VR.

The data that I have been studying is 3D position and 3D orientation data of the headset and hand controllers from a current VR device, the Meta Quest 2. The device itself is described more in section 3.1. This position and orientation data is often referred to as eighteen-degrees-of-freedom or 18DOF for short, as there are eighteen values – three positional plus three rotational, each represented in three tracked points – to describe the state of the system at any given time. The bulk of this dissertation is concerned with analysis and inference upon this data.

1.2 VR as Observatory

The use of VR as an method for observation and data collection sits within a larger conception of VR. As described above, much of the focus in the development and use of VR is through its affordances as a display: the immersant experiences a completely computer-generated world that is perceived and responded to, in large part, as if it is real. Much less has been considered about VR's tracking capabilities, representing VR as an input device. I argue that a vision of totalized VR as an ultimate display [113] implies, as a corollary, a vision of VR similarly totalized as an *ultimate observatory*.

To describe this, consider Sutherland's original description of this hypothetical display:

> The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. Handcuffs displayed in such a room would be confining, and a bullet displayed in such a room would be fatal. With appropriate programming such a display could literally be the Wonderland into which Alice walked [cite Sutherland].

However, that appropriate programming - if it is interactive - requires information, and a *lot* of information at that. In order to sit in a chair, the computer needs to know the distribution of weight in the chair so that the chair can be permitted to break (or hold, if appropriate) at the designed point. The computer needs to also know the position and pressure on any lock-picks nearby so that the handcuffs can appropriately respond. And for the computer to locate the bullet, it needs to know the location of the victim's head or other vital organs.[1]

[1] I cannot use Sutherland's example in good conscience without calling out how it is such a fundamentally disoriented vision of a device. Technology does not exist for its own purposes: it exists - or rather, it is designed by people - to promote human well-being. A technology that has the capacity to kill if *it* is not served correctly is the clear and direct inversion of that relationship. Furthermore, the giving of [moral] agency to the computer is precisely the misdirection that precludes algorithm designers and those approving an algorithm's use from taking responsibility in settings such as biased parole decisions. It is not perfectly accurate, but it is certainly more accurate to say that the *designers* of the VR experience and device - not the computer - have in fact killed the immersant.

This image of a powerful display causing death is not a passing comment limited to one man in the 1960s. Palmer Luckey, co-founder of Oculus, the company that led to the current consumer VR revolution, has even gone so far as to have developed a prototype headset-that-can-kill-you in his office, complete with live explosives: https://palmerluckey.com/if-you-die-in-the-game-you-die-in-real-life/. I have wondered if there is something about VR that tempts one to think in this way.

Critically, if a computer is to appropriately *operate* all matter in a place as output, it will need to *observe* all matter in a place as input. This implies the *ultimate display* is also the *ultimate observatory*.

I believe this claim extends beyond simply the motion capture data that is the subject of this dissertation. Motion capture as a technology is limited to - well, capturing motion. Virtual reality extends beyond simply motion, as evidenced by how new headsets include face tracking, eye tracking, heart rate, pupillometry, skin conductance, and beyond. Why these additional channels fit so well with the current vision of VR is not immediately obvious unless VR is both an observatory and a display. Several of these developments are motivated by higher-fidelity avatars: to faithfully represent the immersant in all their actions to others. This is precisely how the observatory is the dual to the display. While my work here is framed within motion capture data, I leave it to the reader to judge whether the principle here convincingly extends to the entirety of action and interaction within the VR device.

Naturally, current VR devices do not control the existence of all matter in a room, or even all stimuli for a person. Headsets do not block out all light, sounds from barking dogs and friendly reminders still come to the immersant's ears, bodysuits are not common, and work in virtual smell and taste is nascent. Furthermore, complete sensory fidelity is not universally desired (even if possible). Nevertheless, this link between display and observatory is not limited to the endpoint. The degree to which VR is an interactive display is the degree to which it functions also as an observatory.

It is also worth noting that augmented reality (AR) systems still operate in this principle, even if the real world is visible through the headset - the headset is still operating as the interface to the rest of the world, and retains a "right of first refusal" for any sensory information passing through. A perfect AR headset would, by most accounts, be perfectly aware of the world it is placed in.

The use of VR as an observatory underlies the work that has been done in this dissertation. I bring it into the foreground here so that it can be noticed in other research work as well and become a focus of insight and criticism.

1.3 High Temporal and Spatial Resolution

Virtual reality is not the only way of collecting head and hands pose data. However, it is particularly valuable because of its high temporal and spatial fidelity. Yaremych and Persky discuss this work well in their review of the use of VR for what they call *behavioral tracing*, the "covert and continuous collection of behavioral data at high spatial and temporal resolutions" [132]. One example of this is the use of *path tortuousity*, a metric from fractal analysis indicating the indirectness of a path, in the context of navigating a virtual buffet to predict a drop in parents' guilt about their children's eating habits [131]. The mental effort the parents undertook (which caused the drop in guilt) was available through the micro-movements when browsing the buffet. This is only possible as a behaviorally traced measure, as weaker spatial or temporal resolution would not have caught these movements.

This high resolution is in contrast to many studies from previous work in studying behavior. For example, one of the most common measures of interpersonal space is the stop-distance task [52]. A person approaches the study participant until the study participant is uncomfortable, at which point the participant indicates as such. The spatial resolution is fair, as it is limited by the participant's and confederate's reaction time, but the temporal resolution is on the order of tens of seconds at the very least.

Another example, this time in the study of interpersonal synchrony, is the use of film and micro-coding to assess events that occur at frame-level temporal resolution [26]. However, these events are coded with coarse, body-part-level spatial resolution, and simple direction identifiers (up, down, left, right).

Because of VR's built-in tracking system already designed for high spatial and temporal fidelity to avoid simulator sickness and to produce presence (see section 2.3), this data is provided out-of-the-box.

1.4 Group Interaction

Virtual reality is not the only system that can provide both high spatial and temporal resolution of head and hands pose data. For example, motion capture systems do the same. However, VR is well suited to the study of group interactions.

The study of group interactions is highly valuable. Collaboration is the means by which almost all human work is performed; yet for its importance and ubiquity, its processes are not fully understood. Two of the three projects I describe are in fact studies of a type of group behavior. Without the use of VR, the study described in chapter 3 would involve significant time outfitting all participants with motion capture gear, and would be visible to all participants throughout the experiment. Furthermore, this data was collected remotely, with participants all over the United States and one or two in other countries. The ability for groups to come together with less work on the part of the researcher is a great benefit.

1.5 Gaze and Proxemics

The first research work I discuss, chapter 4, is a study of the proxemics and gaze of participants.

Proxemics is the study of person-to-person proximity and its relations with affect, behavior, and cognition. We experience the dynamics of personal space throughout person-to-person interaction. Understanding proxemics is important because it relates to several constructs of interest including liking, communication, and warmth [19]. Gaze is another important social signal, as it can signal attention, intention, and intimacy [89, 20].

Clearly, if positions of participants are tracked in virtual reality, one can collect information about their positions, orientations, and relative distances. These values capture proxemics. Less obviously tracked by VR is the direction of an immersant's gaze. This is done by tracking the direction of the

head and noting that the angle between the head direction and eye direction (defined as eccentricity [106]) is less than 20° for 95% of the time.

In this study, we found that participants spread out farther over time because they could hear each other, which I hypothesize is caused by a lack of drop-off in sound volume in the virtual environment. Then, in order to maintain an appropriate level of intimacy at this increase in distance, participants also increase the amount of gaze they share over time.

I also find that the environment affected interpersonal distance such that larger spaces encouraged larger distances, and pairs of participants showed some degree of consistency in their interpersonal distances, and I hypothesize these findings generalize to nonverbal behavior beyond virtual reality.

1.6 Synchrony

In the second work, chapter 5, I study synchrony. In short, synchrony is a ubiquitous property of person-to-person interaction that the interactants synchronize their actions with each other given little to no intentionality. The high spatial and temporal resolution of this data, as well as its use in group settings, led to a uniquely large data set on motion in conversation and task-based interactions.

Previous work differs widely on what actions and what time patterns count as synchrony, so I have leveraged this rich motion data to compare nearly ten thousand methods of measuring synchrony. The 9300 measures of synchrony are constructed based upon defensible choices provided in previous work, and are evaluated on content validity, consistency, and predictive validity using a multiverse analysis. Overall, I find high content validity, varying consistency, and low predictive validity. With the exception of one effect due to technical issues with recording, I expect these findings to carry over to a similar setting outside of VR. These results provoke questions as to the nature of synchrony.

1.7 Identification

In the third, chapter 6, I show that this motion data collected by virtual reality is powerful enough to be a risk to privacy through behavioral biometrics and re-identification attacks. The motion data collected over time also provides evidence that this re-identification is most likely when the duration between collection and identification is small. This large dataset and the natural setting in which it was collected also prompted the use of a different accuracy metric to compare effectiveness of models independent of participant sizes and the development of a novel parameterization of space relative to the user's body to permit the use of horizontal movement independent of a global coordinate system.

1.8 Roadmap

This dissertation consists of six chapters that follow, articulating three research studies upon the same dataset. Chapter 2 gives background on VR from the interdisciplinary perspectives I take in this work. Chapter 3 describes the dataset, how it was collected, its extent, and its properties. Chapter 4 describes an investigation of gaze and proxemics in this dataset. Chapter 5 leverages the size and duration of this dataset to conduct a comprehensive study of the conditions for nonverbal synchrony. Chapter 6 describes the ability of the motion tracked in this dataset to be identifiable data, and the relationship of duration on identifiability unique to this dataset. Then, Chapters 7 and 8 wrap up the findings of these studies and synthesize the results.

Chapter 2

Background

Because this work is interdisciplinary, it is important to frame virtual reality within each of the different perspectives that inform this work. In particular, I use the perspectives of design, computer science, psychology, and communication. The design (or rather, the designs) of VR articulated through a history of visions of VR each describe a relationship between a person and a device to achieve some good. Computer science brings a technical perspective to VR as an apparatus with an information flow consisting of tracking, rendering, and display technologies. From the perspective of psychology, VR induces *presence*, the perception of non-mediation [65], and poses questions about human perception more broadly. From the perspective of communication, VR is a medium which supports common and novel interactions and provide a methodological tool to investigate person-to-person interaction. Each of these perspectives are provided to establish the background in which I study virtual reality.

2.1 Visions

This section details the visions of virtual reality. Each of these visions is in fact a design in how it supports a class of experiences and sets the defining factors of those experiences in relationship to a context and a use. This collection of visions is best structured historically, as these visions are developed, experienced, critiqued, and replied to over the course of time.

The starting point of VR as is commonly attributed is the the work of Ivan Sutherland and his collaborators in the 1960s at MIT and the University of Utah. In the development of a device nicknamed the "Sword of Damocles" [111], he and his collaborators established the fundamental aspects of an virtual reality system, the tracking, rendering, and display, and demonstrated the coherencies of this process ultimately produced the illusion that the virtual object responds as if it were real. In this particular device, the display showed several wire frame shapes, such as a cube, a house, and a molecule, produced using common graphical pipeline transformations (projection, clipping, etc.) and displayed onto a transparent display. The display being transparent often makes

this device also the commonly described origin of augmented reality as well.

This device was an instantiation toward the vision of what Sutherland called "the ultimate display" [113]. In his time as a PhD student at MIT, he saw the development of pixels and screens as display technologies. In his own work with with Sketchpad [112], he demonstrated the value of computer-aided design and how computation could be an aid to - put another way, he began to articulate some of the affordances of virtual things. These threads prompted a reflection on what a display is for, which was extrapolated into the "ultimate display", which he described as "a room within which the computer can control the existence of matter."

The next prominent vision of virtual reality was expressed well by Jaron Lanier, who in fact coined the term “virtual reality". In an article describing an interview with him that was published by the Whole Earth Review in 1989 [59], he says:

> Virtual Reality is not a computer. We are speaking about a technology that uses computerized clothing to synthesize shared reality. It recreates our relationship with the physical world in a new plane, no more, not less. It doesn't affect the subjective world; it doesn't have anything to do directly with what's going on inside your brain. It only has to do with what your sense organs perceive.

The system Lanier demonstrated was complete with full-body suits and color graphics, and used a head-mounted display. The interviewer recounted the development of a new virtual world that night, live, for the interview. For Lanier and his collaborators, virtual reality was "just as real as the physical world, no more, no less."

This design of virtual reality stepping into the popular discussion provided a point of contrast for Marc Weiser in his 1991 article “A Computer for the 21st Century" [126] in which he directly critiques Lanier's vision of VR in relation to the vision of “ubiquitious computing."

> Perhaps most diametrically opposed to our vision is the notion of virtual reality, which attempts to make a world inside the computer. Users don special goggles that project an artificial scene onto their eyes; they wear gloves or even bodysuits that sense their motions and gestures so that they can move about and manipulate virtual objects. Although it may have its purpose in allowing people to explore realms otherwise inaccessible – the insides of cells, the surfaces of distant planets, the information web of data bases – virtual reality is only a map, not a territory. It excludes desks, offices, other people not wearing goggles and bodysuits, weather, trees, walks, chance encounters and, in general, the infinite richness of the universe. Virtual reality focuses an enormous apparatus on simulating the world rather than on invisibly enhancing the world that already exists.

Both the vision Lanier expresses and the vision Weiser expresses align with Sutherland's ultimate display: both aim to place more and more of human interaction within a computationally-accessible context. However, the distinction is what to do until we reach that ultimatum. Lanier's method is to immediately shrink the circle of experience to what can be generated and maintained by a

computer, but opening up possibilities of interaction that are far wilder than permitted by physical reality. Weiser looks to retain our current approaches of interaction and slowly expand the circle of what can be computationally-accessible.

This tension between virtual reality and ubiquitous computing coincides with several developments in virtual reality technology that offer resolutions to this tension. First, Carolina Cruz-Neira and her collaborators at University of Chicago developed the "CAVE" [27], which uses projectors instead of a head-mounted display so multiple people can be present and real objects can be added into the scene with minimal overhead. Cruz-Neira and collaborators also explicitly name five design goals in the development of the CAVE, and two address Weiser's comment of the exclusionary design of VR. Numbers 3 and 4 are "The ability to mix VR imagery with real devices (like one's hand, for instance)" and "The need to guide and teach others in a reasonable way in artificial worlds."

The second development at a similar time was the coining of the term "augmented reality" by Caudell and Mizell [23] in which spatially-located information - for example, airplane part drilling instructions - were displayed to a user through a transparent heads-up display. Again, this begins to address the fact that only the digital is permitted in VR.

As this VR boom cooled, it remained accessible to high-end research labs in academia and industry. One place in which VR found a niche was in methodologies innovating on the study of social behavior. While further discussed in section 2.4, Blascovich, a professor of social psychology at University of California at Santa Barbara, began to argue for the value of VR for science, in particular benefits to realism, replication, and representation.

The current VR boom was sparked by Oculus, a startup company that designed consumer-grade virtual reality devices. With the proliferation of cheaper screens and sensors spurred on by smartphones, it became possible to develop passable VR displays at a price point similar to a smartphone or a game console. The possibility of a large market of new VR users sparked many variations in the design of virtual reality, including the proliferation of social VR applications and platforms like VR Chat, Altspace, and Mozilla Hubs in which participants embodied avatars in social settings, enabling meetups, open mic nights and creative expression, alongside enactments of racist memes and embodied digital harassment. This growth of social VR coinciding with other nascent technologies such as blockchain systems and virtual property represented by non-fungible tokens (NFTs) caused speculators to appropriate the term "metaverse" to a hypothetical, single, shared virtual world. It remains to be seen how this implicit vision of VR will be described and distinguished from those that precede it.

This account is not meant to be complete, and these visions are not represented by a single person (or group of people). Nevertheless, this provides the setting for explicit and implicit designs of virtual reality. Among these designs, there is a presumption is that the world we live in ought to be operated through computation just as much as through direct action, but there are a wide range of possibilities regarding the role of the physical world and its relationship to the virtual. Furthermore, the focus in this designs is the content displayed to the immersant, and the information tracked about the user is secondary or implicit.

2.2 Device

From a technical perspective, VR can be effectively described by its input, processing, and output. Its input is *tracking*, in which the position of the user is tracked. This position is necessary for the next step, the *rendering*, which generates the view of the virtual scene in which the immersant is represented by the camera position. Then, once the virtual scene is rendered, the scene reaches the user through a *display*. In today's systems, this process happens between 90 and 120 times a second in order to minimize simulator sickness and increase *presence* as discussed in section 2.3 .

Several tracking methods have been explored over the years, including *mechanical tracking* using angle and distance measurement, *magnetic tracking*, which uses the intensity and direction of a magnetic field to locate tracked points, and *optical tracking* with external cameras like OptiTrack or internal sensors like the Vive Lighthouse system. The most common tracking method in today's headsets is the use of computer vision to operate *simultaneous localization and mapping* so that no special trackers are necessary. The system bootstraps a model of the nearby space at the same time as tracking its own position and orientation. In all of these methods, a low-latency (<10ms), accurate position is the goal.

Rendering virtual content to the VR display is largely similar to real-time computer graphics more broadly, but there are a handful of differences. First, the lenses used in most VR displays are strong enough to cause chromatic aberrations, requiring correction to account for the different refractive indexes for red, green, and blue light. Rendering in stereo vision is, to a first approximation, twice as expensive as rendering in mono, and so it usually takes twice as long to redber a VR scene as a standard scene. Finally, the angular field of view of most VR displays is far larger than television, computer, phone, or movie screens, and so there is more compute necessary to reach the same angular resolution.

Once the virtual scene is produced, it is displayed to the user. In most cases, this is done through a head-mounted display, a device the user wears on their head that blocks out external stimuli. As alluded to in section 2.1, there is also the CAVE (a self-referential acronym for CAVE Automatic Virtual Environment) in which screens or projectors surround the user. This approach can be extended into projecting onto nearby objects [94].

The technical aspect of virtual reality consists of tracking, rendering, and display. This is done to produce spatially coherent given motion of the viewpoint, the most important way one interacts with a virtual world.

2.3 Experience

For the psychologist, virtual reality is interesting for the state it produces in the *immersant* (the person using the headset). This is commonly called *presence*, defined as the "perception of non-mediation" [65]. In this state, the virtual reality apparatus is not salient to the user, but instead, the content the device is displaying is.

Taking a step back, this condition is particularly interesting for the psychologist. It is strange that when interacting with this virtual world, the immersant believes what is “fake" and ignores what is “real". In fact, virtual reality has been a useful tool and test case in how this process of perception happens.

One theory of presence claims “successfully supported actions" [134] are the root of presence. When actions are made towards an object, the object "reacts", in some fashion, to the action made. When the object’s response is congruent with the person’s expectations of the response, the action is said to be successfully supported.

A very simple example of a successfully supported action is the counter-rotation of a virtual object when the user rotates their head. If a user rotates their head left to right, an object in front of them moves, relative to their field of view, from right to left. A second example would be a glass tipping over when bumped by a user’s hand. The action is the hand contacting the glass, and the support is the production of a realistic tipping over effect.

There is good evidence that motion, and support of responses to that motion, are the strongest factors in presence. This is why tracking, rendering, and display are so important and are in fact mentioned as important in the original HMD paper [111]. In a meta-analysis by Cummings and Bailenson [28], the two most dominant determinants of presence are update rate and number of tracked points.

2.4 Medium

VR is also a medium for person-to-person communication. This often ties into social psychology. The value of VR for social psychology was articulated well by Blascovich and collaborators [17] in the 2002 paper "Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology." In it, they argue for three affordances VR provides over multimedia or real-world studies: realism, replication, and representation.

It is strange to argue that virtual reality can be more real, but it is ’real’ in the sense of mundane realism. In a study, there is usually a trade-off between mundane realism (how representative the scene is of real life, as well as participant’s understanding of its realism) and control (how much is kept constant across conditions or manipulated exactly by condition). Virtual reality, if high presence is achieved, allows cheaper development of high-realism, high-control settings. Second, they propose it is easier to replicate studies in VR. In principle, this may be as simple as transferring files to another researcher, but in practice, there is more complexity. Nevertheless, less is left to chance with VR. Finally, they propose that studies can be more representative by removing constraints of participant sampling. Largely, this has not come to pass.

However, these all assume that VR experiences are only replicated from real life experiences. This doesn’t allow changes and variations not available to real-world experiences, as described in a paradigm called Transform Social Interaction [9]. For example, immersants may view a talk spoken by one person with gestures added by another, or receive a good degree of eye contact provided

automatically.

With this potential, how has VR been used? Surprisingly, in 2022, Han and collaborators [51] reviewed the literature and found 37 experimental studies that had multiple participants in immersive VR interacting with each other at the same time. The lack of studies in a large portion of this time can be attributed to the availability of the headset and logistics of running multi-user studies.

In popular use, social VR has become possible supported by the availability of consumer VR devices. Software that enables these experiences vary from entertainment platforms like VRChat, Rec Room, and AltspaceVR to professional platforms like Mozilla Hubs, Meta Horizon Workrooms, and ENGAGE. There has been good ethnographic work understanding the conceptualizations and usage of these platforms. One characteristic of current social VR is the interplay between grounding experiences in social context and adapting to the new affordances of the medium.

Chapter 3

Data Collection

The three works described in the following chapters all use the same dataset. For consistency and space, this dataset is described once, producing this chapter.

This work reports on data collected as part of the Stanford Longitudinal VR Classroom Dataset over the course of two periods of data collection of classroom immersive VR [50]. Students met in small groups ranging from two to 12, and consented to have their verbal, nonverbal, and performance continually tracked during each course, typically eight weekly sessions which lasted about 30 minutes per session. In addition, each student provided self-report data about their experience after each session (see [50] for a detailed description).

Each period was run using a social VR platform called ENGAGE. Each data collection period consisted of its own participant pool and conditions. Participant consent went through a rigorous process, approved by two separate organizations within the university. Moreover, there was a 3rd party arbiter who oversaw data collection during the course, and students had an interactive, hour-long discussion of the study procedures and data collection before deciding to consent. Data recorded included position and rotation of each participant's headset and hand controllers as well as some questionnaire data.

3.1 Apparatus

In both periods, participants used the Meta Oculus Quest 2 headsets (503g) and two hand controllers (126g) in their own personal environments. The combined field-of-view of the headset is 104.00° horizontal FOV, 98.00° FOV. In the avatar study period, two participants opted to participate with owned personal headsets (both PC-based Valve Index). The headsets did not perform eye tracking, and in order to complete this experiment at the scale necessary, I opted not to include add-on eye tracking devices. Note that here, head orientation is used as a proxy for gaze. In the first round of data collection, two participants opted to participate with owned personal headsets (both PC-based Valve Index). All participants in the second round used the Meta Quest 2.

Figure 3.1: Participants performing discussion activities in the VR environment. In the top left panel, participants illustrate the environmental impacts of an oil spill with a duck model covered in black smudges representing oil. In the top right panel, several students discuss the experience of the headset using hand gestures in front of their faces. In the bottom left panel, the participants used 3D drawing to illustrate the water cycle as well as 3D models to show livestock in fenced-in environments. In the bottom right panel, participants are spreading out into different corners of the space and brainstorming on how to visualize their ideas to the assigned prompt.

The software in use was the ENGAGE virtual communications platform, versions 1.7 through 2.0.1, produced by ENGAGE PLC. The virtual environments in which the participants met varied by period. In period 1, all participants met in the same "Engineering Workshop" room. In period 2, participants met in one of 192 uniquely-built environments each week. These environments differed in size of moving area, height, and whether it was indoors or outdoors. Figure 3.1 shows screen captures of anonymized virtual students in discussion.

The physical locations in which participants were based were not organized by the researchers. However, anecdotal data indicated that most participants in period 1 were not on campus but rather scattered across the US. In period 2, most participants were on campus and participated in the virtual world from their near-campus housing.

3.2 Participants

There were a total of 232 participants in the study across the two subject populations ($n_1 = 86$), ($n_2 = 146$). Participants were university students enrolled in one of two 10-week courses about VR. While all students who were part of the course took part in all the VR activities, only those who consented to participate in the study had their data included in the study. Of the 101 students in Study 1 and 171 in Study 2, 93 and 158 consented to participate in the study, respectively.

In Study 1 (Female = 30, Male = 47, Other = 2, declined or did not answer = 7), participants were between 18 and 58 years old (M = 22.3, SD = 5.2; $n_{18-23} = 68$, $n_{24-29} = 7$, $n_{30-34} = 3$,

$n_{35-39} = 1$, $n_{55-59} = 1$, $n_{declined} = 6$ and identified as African American or Black ($n = 11$), Asian or Asian American ($n = 30$), Hispanic or Latinx ($n = 9$), Middle Eastern ($n = 1$), White ($n = 21$), more than one race ($n = 5$), or declined to or did not respond ($n = 9$). Participants had varying levels of experience with VR, with 41 (51.2%) having never used VR before. Prior to the course, 38 participants were not familiar with anyone in their discussion group, and others reported knowing one ($n_1 = 13$) or more members ($n_2 = 12$, $n_3 = 1$, $n_4 = 2$, $n_5 = 2$).

In Study 2 (Female = 59, Male = 79, declined or did not respond = 4), participants were between 18 and 49 years old (M = 20.9, SD = 2.8; $n_{18-23} = 133$, $n_{24-29} = 4$, $n_{45-49} = 1$, $n_{declined} = 4$ and identified as African American or Black ($n = 12$), Asian or Asian American ($n = 47$), Hispanic or Latinx ($n = 8$), Indigenous/Native American, Alaska Native, First Nations ($n = 2$), Middle Eastern ($n = 1$), Native Hawaiian or other Pacific Islander ($n = 5$), White ($n = 41$), more than one race ($n = 19$), a racial group not listed ($n = 1$), or declined to or did not respond ($n = 2$). Participants had varying levels of experience with VR, with 50 (36.2%) having never used VR before. Prior to the course, 67 participants were not familiar with anyone in their discussion group, and others reported knowing one ($n_1 = 36$) or more members ($n_2 = 10$, $n_3 = 4$, $n_4 = 5$, $n_5 = 1$, $n_7 = 1$).

3.3 Procedure

Students opted in to the experiment at the beginning of the course with a consent form approved by the Stanford University institutional review board (IRB) under protocol IRB-61257, and the Stanford University Student Oversight Committee. This IRB process required that researchers and course staff did not know which students opted in as participants in the experiment until after the course finished, and an external third-party arbitrator controlled the consent process, so that there would be no plausible appearance of coercion to participate in the study. This also implied that all students were recorded in this study, and data was filtered out from non-consenting participants after the course finished. Furthermore, participants were not compensated because they performed the same activities regardless of whether their data was used. Upon the start of each session's recording, the system gave a visual notification that recording was taking place. Before consent was given, one of the authors gave a 30-minute lecture on data tracking privacy, and the benefits and risks of consenting in the study, and gave the students the opportunity to ask questions about the study. All experimental protocols were approved by the Stanford University IRB (Institutional Review Board), all participants' informed consent was obtained, and all methods were carried out in accordance with relevant guidelines and regulations.

Participants provided consent for the use of this data for education and basic research. The authors took consideration to separately query the IRB if any further approval was necessary for attempting to perform re-identification and inference of personal characteristics, but the review board determined that the work was out of scope for the IRB because the data had already been collected and was - from the perspective of the IRB - already de-identified, so the IRB concluded there was no risk to participants.

Weekly activities varied, but included large-group discussion on current readings, discussion in pairs or triads on course material, and VR building activities. Sessions were led by a researcher, who was part of the teaching team of the course. In all virtual environments, participants were able to walk/teleport freely, create 3D drawings, write on personal whiteboards/stickies, add immersive effects/3D objects, and display media content. There was a library of about one thousand virtual objects available for participants to create, move, organize, and delete in the virtual spaces. The platform accommodated use of 3D audio, which allowed for splitting off into smaller groups without audio overlap. Sessions took place eight times over the course of eight to nine weeks, and the duration was about thirty minutes per session.

Tasks have been reported in [50] and are re-printed here. The tasks during the first data collection period by week (1-8) were:

1. Acclimate participants to the headset and platform, leaving margin for technical or content issues, and Discussion on being inside VR and how the experience compared to that in Zoom
2. Full-group discussion reflecting on participants' experience visiting various sites in AltspaceVR (e.g., an art exhibition, solar system), and sketching of ideas of how one might teach and present content inside VR
3. Full-group discussion reflecting on recording and performing as avatars inside of VR
4. Small-group discussions reflecting on various VR empathy experiences
5. Small-group discussion on how VR is used for medical applications and well-being
6. Small-group activity in which participants chose a unique feature of VR and brainstormed how to communicate climate change based on this feature
7. Activity, done either individually or in small groups, on creating and playtesting a VR-based game
8. Small-group discussion reflecting on VR and its use cases, dangers, and potential direction

The tasks in the section period were more structured and consisted of a discussion followed by an activity. These tasks by week (1-8) were:

1. Discussion: Acclimate participants to the headset and platform, leaving margin for technical or content issues. Activity: Consider the affordances of VR and create a prototype of something that leverages the uniqueness of VR
2. Discussion: Full-group discussion on what activities heightened sense of presence in VR. Activity: Create something frightening that induces a feeling of high presence.
3. Discussion: Full-group discussion reflecting on participants' experience visiting various sites in AltspaceVR (e.g., an art exhibition, solar system). Activity: Consider the affordances of VR to make a difficult concept easier to understand

4. Discussion: Full-group discussion on how to improve ENGAGE's avatar if the participant were in charge of ENGAGE. Activity: Create something that reimagines avatars and representations of the self.

5. Discussion: Small-group discussions reflecting on various VR empathy experiences

6. Discussion: Full-group discussion how VR is used for medical applications and well-being. Activity: Create a meditation room or "safe-space"

7. Discussion: Full-group discussion on VR's role in people's attitudes and actions toward climate change. Activity: Brainstorm an idea of how to communicate a message about climate change

8. Discussion: Full-group discussion on VR's role in the future of sports and fitness. Activity: Create and playtest a VR-based game.

3.4 Conditions

3.4.1 Avatar Study

The avatar study consisted of two conditions with two levels each, counterbalanced across sessions in a Latin square. As there were eight weeks, this assignment was repeated for the second four weeks, and as there were eight groups, the assignment was identical for the second four groups.

Avatar: self vs. uniform. One condition that varied by group and week was the embodiment of either a *self-avatar*, in which a participant was told to create an avatar that 'looks and feels like you', or a predefined *uniform avatar*, determined through pre-testing to be the most gender- and racially-ambiguous among the options available.

Synchrony manipulation: present vs. absent. At the beginning of each session, participants performed either a synchronized motion activity, raising and lowering arms in unison with the rest of the group, or an individual drawing activity that matches the amount of motion in the synchrony activity condition, but not the synchronous nature of the activity.

3.4.2 Context Study

In the study on virtual context, there were three conditions with two levels each, counterbalanced across sessions with a Latin square. As there were twenty-four groups, three groups were assigned to each of the eight sequences in the Latin square. A total of 192 environments were created in a stimulus sampling paradigm [95], 48 per combination of the two environmental variables (view and setting).

View: panoramic vs. constrained. The environments varied in terms of the amount of space visible in horizontal and vertical directions. In panoramic environments, much more space was visible and accessible than in constrained environments.

Setting: indoors vs. outdoors. In order to tease apart the natural correlation between outdoor, panoramic spaces, and indoor, constrained spaces, I varied these two dimensions separately.

Motion: active vs. passive. Finally, some groups in some weeks were asked to minimize their UI-based motion, like teleporting and smooth movement. A manipulation check revealed this was only partially effective, i.e., there were differences in movement between the two conditions, but there was still substantial teleporting and smooth movement in the passive condition.

3.5 Data

To illustrate the scale and the structure of the data, Figure 3.2 shows the weekly session, session duration, group sizes, and participants this work. The highest level of data organization was the *study*, which was either the *avatar study*, collected in summer 2021, or *context study*, collected in fall 2021. These are represented in the figure as two separate plots. The next levels of organization are the week and the section. The *week* indicates which week of eight the data were obtained, and is laid out in columns. The *section* was the group and time participants met for discussion, and is laid out in rows. In the avatar study, there were eight sections, and in the context study, there were 24 sections. Each participant took part in only one section per week. Usually, a participant attended the same section week to week, but there were some exceptions. A *session* is one participant's data for one week. Each session lies entirely within one and only one section. In total, I obtained data on 1745 sessions that, on average, lasted 31.19 minutes (SD = 7.86 min).

In total, there were 1,683 sessions with an average length of 27.2 minutes (SD = 11.2 min). During a session, the data was collected at 30Hz and consisted of four tracked points. Three were the traditional head, left hand, and right hand, and the fourth is the 'root', the relationship between the participant's physical space and the virtual space. This was used in cases when a participant translated or rotated their position with a UI control (e.g., teleporting by pointing and clicking, tapping the controller joystick left to rotate left by 15 degrees). As I was interested in an attack made by a malicious user and not a compromised system, I computed the avatar's head and hands positions as visible to another user, which is relative to the virtual spaces' coordinate system.

Visual features in Figure 3.2 highlight several aspects of the data. Empty rectangles (e.g., Avatar, Week 2, Group 4) indicate both primary and backup recordings failed and data was lost, which happened occasionally but not often. Single-line recordings (e.g., Context, Week 3, Group 10) occurred when the recording failed but one group member attended a different section that week. Roughness on the left side of the plot (e.g., Group 12 Week 4 Inset) indicated variation in when participants arrived in the virtual world. A premature and sharp cutoff on the right side (e.g., Avatar Study, Week 2, Group 3) indicated a system crash, but a fuzzier break (e.g., Group 12 Week 4 Inset) indicated a normal group dismissal.

Several analyses use a *pair* as a unit of analysis. This is a complete pairing of participants within the same session, meaning in group of e.g., 6 participants, there will be $\binom{6}{2} = 15$ pairs for symmetric relations such as the distance between participant A and participant B and $6 \times 5 = 30$ pairs for

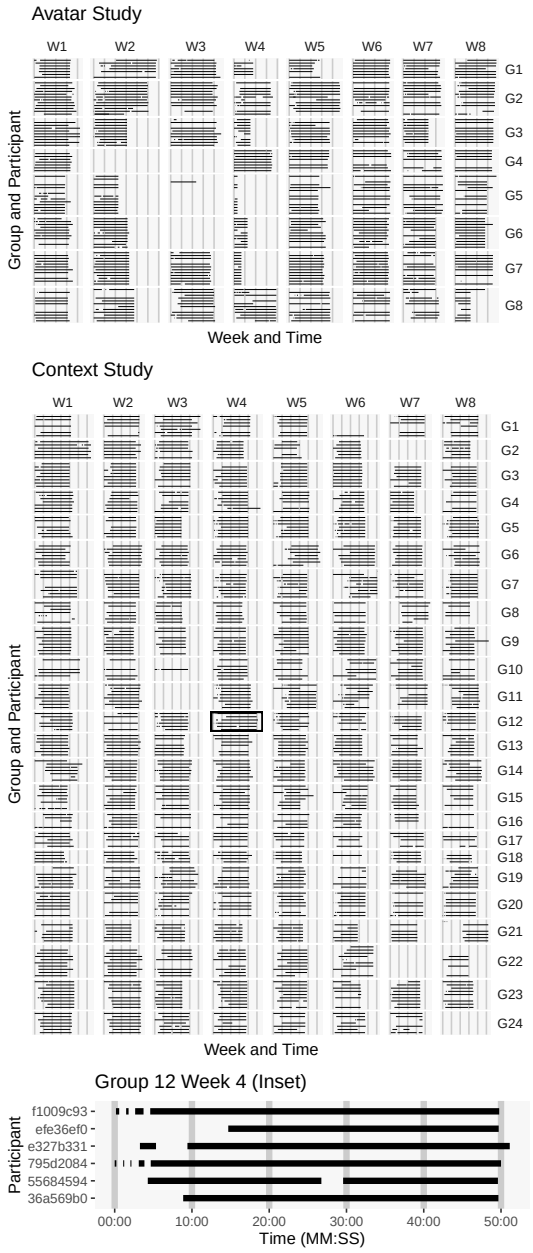

Figure 3.2: Figures showing the weekly sessions, duration, participants, and group size of the two studies. Panels 1 and 2 (Avatar Study and Context Study) consists of many facets. Each facet represents a week, given by its horizontal ordering, and a group, given by its vertical ordering. Within each facet, and in the final panel that shows a close-up of group 12 in week 4 from the context study, each participant receives a horizontal row on which a line is drawn if data is collected at that time. Vertical lines within each facet demarcate 10-minute intervals of time.

non-symmetric relations such as the percentage of time participant A is in participant B's field of view. On average, there were 21.9 pairs per session, with more per session in the avatar study (34.4 pairs per session) than the context study (17.9 pairs per session) due to the difference in group sizes (avatar, 9.25 people per session; context, 6.43 people per session).

During a session, the data was collected at 30Hz and consisted of four tracked points with six degrees of freedom each. Three of the points were the headset, left hand controller, and right hand controller, and the fourth is the 'root', the transformation between the participant's physical space and the virtual space. The root changed when a participant translated or rotated their position with a UI control (e.g., teleporting by pointing and clicking, rotating 15 degrees left by tapping the controller joystick).

The coordinate system of the data follows the conventions used by the Unity game engine; namely, a left-handed coordinate system with Y upwards, Z forwards, and X rightwards, and intrinsic rotations in the order of yaw (Y), pitch (X), and roll (Z), where positive values indicate a left-handed rotation relative to the positive direction along the axis.

Each of the studies uses this motion data. The first, chapter 4, uses this spatial data to infer interpersonal distance and gaze and attention. The second, chapter 5, uses the motion data along with ratings of entitativity, familiarity, and attraction. The final, chapter 6, uses motion data only.

Chapter 4

Gaze and Proxemics

4.1 Introduction

Virtual Reality (VR) captures an unprecedented richness of data from its users. In the common case, a device tracks head and hands position and rotation dozens of times per second. There are many questions regarding social behavior that this data can answer. Two of the most prominent are where people stand and were people look. These two constructs - proxemics and gaze - are easily available through this data and can be analyzed in high spatial and temporal fidelity in a paradigm known as behavioral tracing [132]. Understanding proxemics is important because it relates to several constructs of interest including liking, communication, and warmth [19]. Violations of personal space can be confusing and stressful, both when one is too far from another, or when one is too close. Gaze is similarly important, as it can signal attention and intimacy [89]. The study of proxemics and gaze is important to social virtual reality.

However, virtual reality is still a new medium, and effects of novelty and learning can influence findings. To this end, I studied the behavior of 232 participants who participated in social virtual reality eight times over eight weeks for about thirty minutes per session. I found several effects, including influences of facets of place, time, and dyad on these behaviors. For example, dyads increased their personal space over time, but also looked at each other more often over time. There was also a relationship between personal space and directness of gaze that corroborates what is known as equilibrium theory [5]. These results continue to show that social virtual reality carries over many of the patterns that I know from social interaction in real life. They also encourage future work studying the effects of virtual reality on behavior over time and within varying contexts.

4.2 Related Work

This work builds on several threads of previous research: longitudinal studies of social virtual reality, proxemics, and gaze. I review each of these threads and then bring them as context to my research

questions.

4.2.1 Social VR Over Time

Social VR has increased in popularity in recent years, spurred on by the availability of consumer VR devices. Software that enables these experiences vary from entertainment platforms like VRChat, Rec Room, and AltspaceVR to professional platforms like Mozilla Hubs, Meta Horizon Workrooms, and ENGAGE.

The earliest studies of social VR took place around 2000. For example, works led by Slater [108] and Garau [42] explore communication differences between virtual reality and face-to-face communication. Despite this early start, Han and colleagues [51] report only 37 social VR studies in their 2022 paper. There have been even fewer studies that have 3 or more participants sharing a virtual space in immersive virtual reality. Mütterlein and collaborators [78] studied groups of two to four, varying several facets of VR and observing their influence on intention to collaborate. Moustafa and Steed [77] performed an exploratory in-the-wild study of collaboration over headset-only VR with several groups of two to four participants. Roth and collaborators [97] studied groups of five participants in an augmented a virtual museum experience with visualized social signals like joint attention and eye contact. Finally, while not a study of immersive VR but rather a desktop-based virtual environment, Williamson and colleagues [128] studied the proxemics of 26 participants in a virtual workshop. Considering how many studies of virtual reality have been done, work on groups in VR has been difficult to come by.

It has been even rarer to find studies of social VR over time. Longitudinal studies are often difficult to coordinate, but are able to show adaptation to a given medium and study behavior of users who are well-acclimated to a system. Bailenson and Yee [7] studied three groups of three participants in 15 sessions over seven weeks and found substantial changes and adaptations over time in several variables. Roth and collaborators [98], Moustafa and Steed [77], and Khojasteh and Stevenson Won [60] all found several adaptations to the medium of virtual reality when studying participants behaviors over time. While individual adaptations depended on the affordances and frustrations of the VR hardware and software used in the study, it is clear that much of the adaptation was participants communicating more through the available social signals. This adaptation is a well-known result in computer-mediated communication [122]. To give an example, participants in the study by Moustafa and Steed [77] used a VR system with a headset and no hand controllers. Instead of a wave for a greeting or farewell, which would normally involve hand tracking, participants 'waved' using their heads, tilting their head left and right. Han and colleagues [51] found several effects of time, including greater presence, enjoyment, entitativity, and realism over time.

4.2.2 Proxemics and Gaze

Proxemics is the study of person-to-person proximity and its relations with affect, behavior, and cognition. Hall's work [49] on proxemics on middle-class American adults defined four levels of

proximity: intimate (< 0.45m), personal (0.45m-1.2m), social (1.2m-3.6m) and public (> 3.6m). These thresholds are not universal but rather vary depending on large-scale, cultural variables like the prevalence of contact [15] and the relative importance of individualism versus collectivism [68].

Virtual reality is amenable to studying proxemics due to the built-in capacity to track a user's position. This position data can then be leveraged as a continuous measure, oftentimes with high spatial and temporal fidelity [132]. Proxemics has been used to inform both independent variables and dependent variables. Bailenson and colleagues [11] leveraged proxemics to demonstrate virtual characters receive more personal space and seem more real when more behaviorally realistic. Bönsch and collaborators [19] show that angry characters receive more personal space than happy characters. Choudhary and collaborators [25] varied two affordances of social VR, volume and head size, to investigate their effect on distance estimation. Head size affected distance estimation, but volume did not. Takahashi and collaborators [115] varied the speaking volume of a character and noted that in a walking task, participant gave more distance to the character when the character spoke louder.

Gaze is another nonverbal form of communication that can intention, attention, and intimacy [20]. The value of this communication has even motivated technical developments in various stereo-like displays [87, 86]. Overall, the use of gaze in mixed reality has been substantial and has recently been reviewed [89]. Vertegaal and collaborators [118] showed that for a majority of time, people look at the speaker or the target in multiparty conversation. Gaze also signals turn-taking. When a virtual conversation was instrumented with automated random gaze, participants spoke in a greater number of turns and shorter duration turns compared to no changes in gaze, but teams using this random gaze model did not complete a task as quickly as teams using realistic gaze [119]. Particular kinds of gaze, like eye contact, do not necessarily signal effective or ineffective communication on their own, but can do so if other criteria are met [90]. One of the changes over time that Bailenson and Yee [7] found was that participants looked at others less over time. They explain this effect due to the weight and discomfort of the headset and the lack of any facial cues on the virtual avatars.

Finally, proximity and gaze can relate to each other through equilibrium theory [5]. Because both proximity and gaze are signals of intimacy, extrinsic changes to one variable (e.g., stepping into a small space like an elevator) lead to a change in the other variable (less mutual gaze) so as to maintain an appropriate level of intimacy. This has been demonstrated in several contexts, including virtual reality [10, 133].

4.2.3 Research Questions

RQ1: How does interpersonal distance adapt over time in virtual reality? As far as I are aware, there are no studies that investigate personal space over time in virtual reality. Considering there have been other adaptations over time, whether and how people's personal space develops is an important question.

RQ2: How can the virtual reality environment affect spacing? There has been some evidence of the influence of environment on proxemics in general and interpersonal distance in

particular [85], and sparse work using VR [52]. However, there is still much to explore, as there are quite many variables that define a space.

RQ3: What is the relative size of inter-dyad difference factors? It is known there are cultural differences in interpersonal distance, and some inter-dyad differences have been studied, like the gender composition of a pair [133]. However, the intersection of social VR studies that also collect data over time is small.

RQ4: How does gaze change over time? Given that the reasons gaze behavior changed in the work by Bailenson and collaborators [7] were discomfort due to headset weight and lack of nonverbal cues, what happens now that the headset is lighter and nonverbal cues like hand tracking are commonplace?

4.3 Results

The results section is organized with respect to the variables of interest. First, I discuss the effect of experimental manipulations and time upon interpersonal distance. Then, I discuss the distributions of and effects on gaze, proxied by the forward direction of the headset. Finally, I note the relationship between the two described in equilibrium theory [5].

The statistical analyses in this work used mixed-effect models using the 'lmer' and 'lmerTest' packages in the R programming language. In addition to linear models that have an output variable and an input variable, mixed effect models allow the specification of grouping factors for correlated random effects. For example, one random effect is the individual differences due to participant, which avoids both collapsing across observations as with an average and inflating significance with correlated errors.

4.3.1 Proxemics

I define my variable of interest, interpersonal distance, to be a function of the distance between participants' heads. For any given session, this is in fact a distribution of values, so there must be a summary function to collapse this distribution into one value [70]. In some previous work [11, 115], the minimum has been used. However, several concerns led me to select a different summary function. In contrast to face-to-face interaction, virtual reality allows spatial overlap between people. A participant can accidentally teleport into a position that is arbitrarily close to another participant. In addition to this, the sheer length of observation time (31 minutes average) increased the risk for this or other outliers in distance. Therefore, instead of computing the minimum value (first smallest) of the set, I compute the n-th smallest value. I selected $n = 150$ so that the values of five seconds worth of samples are ignored. I judged five seconds to be appropriate for this buffer because it is long enough for participants to react and move away if one participant accidentally moved too close to another. The results I report were robust to variations in this parameter.

Interpersonal distance values were highly right-skewed, and were thus log-transformed. All values

are reported given original units (meters) rather than the model term, log-meters. Consequently, differences between values, such as standard deviations and unstandardized effect sizes, become multipliers, which are written in this work as percentages. I also included the gender composition of the pairs as a covariate, as previous work [133, 54] found effects of pair gender composition on interpersonal distances.

Combined Datasets

The prototypical pair began week 1 at a distance of 1.44m and increased in distance by 7.0% per week over the eight weeks to 2.31m ($t(205.74) = 6.197$, $p < 0.001$). Although pairs in the Avatar study were 8.4% farther apart than pairs in the Context study (1.50m vs. 1.38m), this difference was not significant ($t(29.35) = 0.893$, $p = 0.379$). However, distance did differ across gender pairings (M-M = 1.33m, F-F = 1.42m, other pairs = 1.53m, $\chi^2(2, N = 5341) = 27.65$, $p < 0.001$). This result indicates interpersonal distance increases over time (**RQ1**).

The variance in distances was attributable to section, session, and pair differences. Variation uniquely due to section (the group) had a standard deviation of 18.4% ($\chi^2(1, N = 5341) = 11.92$, $p < 0.001$, M+SD = 1.70m, M-SD = 1.22m). Variation uniquely due to session had a standard deviation of 41.6% ($\chi^2(1, N = 5341) = 565.05$, $p < 0.001$, M+SD = 2.04m, M-SD = 1.02m). Finally, variation uniquely due to pair had a standard deviation of 24.8% ($\chi^2(1, N = 5341) = 83.33$, $p < 0.001$, M+SD = 1.80m, M-SD = 1.15m).

Avatar Study

In the Avatar study, the prototypical pair began week 1 at a distance of 1.61m and increased in distance by 5.4% per week over the eight weeks to 2.33m ($t(51.7) = 3.288$, $p = 0.002$). Neither condition of customized avatar nor shared task beforehand showed significant effects. The prototypical pair in the customized avatar condition was 8.0% closer than the prototypical pair in the uniform avatar condition pair (1.55m vs. 1.67m), which is not larger than would be expected by chance ($t(51.92) = -1.108$, $p = 0.273$). The prototypical pair in the synchrony activity condition was 2.7% closer than the prototypical pair in the condition with no synchrony activity (1.59m, 1.61m), which is not larger than would be expected by chance ($t(50.96) = -0.364$, $p = 0.718$). Distance did differ across gender pairings (M-M = 1.43m, F-F = 1.51m, other pairs = 1.77m, $\chi^2(2, N = 2028) = 18.38$, $p < 0.001$).

The variance among session and among pair was significant, but variation among section was not significant. Variation uniquely due to section had a standard deviation of 11.8% ($\chi^2(1, N = 2028) = 2.15$, $p = 0.142$, M+SD = 1.80m, M-SD = 1.44m). Variation uniquely due to session had a standard deviation of 27.9% ($\chi^2(1, N = 2028) = 88.02$, $p < 0.001$, M+SD = 2.06m, M-SD = 1.26m). Finally, variation uniquely due to pair had a standard deviation of 34.5% ($\chi^2(1, N = 2028) = 54.9$, $p < 0.001$, M+SD = 2.17m, M-SD = 1.20m). In order to compare the fixed effects to individual differences for **RQ3**, I compare the difference in means of each to the standard deviation uniquely

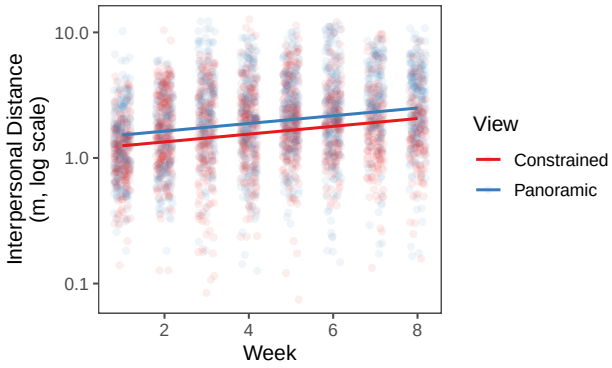

Figure 4.1: Plot of interpersonal distance as a function of week and view within the context study. Each pair is represented by a dot. Lines represent the prototypical pair in the constrained or panoramic condition, as indicated by the line's color.

due to the pair. The effect due to avatar was 0.35 times and synchrony activity was 0.09 times the standard deviation of distance due to pair. Both of these indicate that distance due to pair is much more than distance due to either independent variable.

Context Study

In the Context study, the prototypical pair began week 1 at a distance of 1.38m and increased in distance by 7.4% per week over the eight weeks to 2.27m ($t(162.51) = 5.744$, $p < 0.001$). The prototypical pair in the outdoor environment condition was 5.4% closer than the prototypical pair in the indoor environment condition (1.34m vs. 1.42m), which is not larger than would be expected by chance ($t(158.55) = -0.978$, $p = 0.329$). The prototypical pair in the active motion condition was 29.8% closer than the prototypical pair in the passive motion condition (1.21m vs. 1.57m), which is larger than would be expected by chance ($t(158.95) = 4.611$, $p < 0.001$). The prototypical pair in the panoramic view condition was 23.4% farther than the prototypical pair in the constrained view condition (1.53m vs. 1.24m), which is larger than would be expected by chance ($t(158.74) = 3.719$, $p < 0.001$). Distance did differ across gender pairings (M-M = 1.29m, F-F = 1.39m, other pairs = 1.42m, $\chi^2(2, N = 3313) = 9.06$, $p = 0.011$). Figure 4.1 shows the distribution of distances as well as the effect of week and view on distance. In regards to **RQ2**, I find evidence that panoramic views lead to larger interpersonal distance than constrained views, but no evidence that indoor or outdoor setting influences interpersonal distance.

The variance among each of section, session, and pair was significant. Variation uniquely due to section had a standard deviation of 22.4% ($\chi^2(1, N = 3313) = 16.54$, $p < 0.001$, M+SD = 1.69, M-SD = 1.13). Variation uniquely due to session had a standard deviation of 40.8% ($\chi^2(1, N = 3313) = 440.58$, $p < 0.001$, M+SD = 1.94m, M-SD = 0.98m). Finally, variation uniquely due to pair had a standard deviation of 16.1% ($\chi^2(1, N = 3313) = 19.94$, $p < 0.001$, M+SD = 1.60m, M-SD

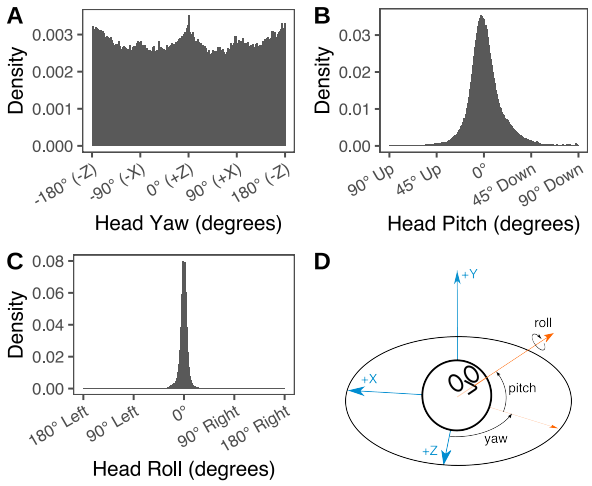

Figure 4.2: Panels showing histograms of Tait-Bryan angles of participants' headsets. Panel A displays yaw, Panel B displays pitch, and Panel C displays roll. Panel D displays a schematic of these angles relative to a participant's headset. Note that yaw and pitch are both in the negative direction in this schematic.

= 1.19m). In order to compare the fixed effects to individual differences for **RQ3**, I compare the difference in means of each to the standard deviation uniquely due to the pair. The effect due to environment condition was 0.35 times, motion was 1.75 times, and view was 1.41 times the standard deviation of distance due to pair.

4.3.2 Gaze through Head Orientation

Gaze, a useful measure of attention, can be inferred from headset direction. This is parameterized here in terms of yaw, pitch, and roll as shown in Figure 4.2 panel D.

Yaw, Pitch, and Roll

The distribution of yaw, shown in panel A of Figure 4.2 was relatively uniform. This reflects the fact that how one parameterizes the horizontal plane does not dramatically affect human behavior: I can rotate 30 degrees around the vertical axis and continue on a conversation. Within this uniformity, there were apparent peaks at 90-degree intervals. These effects are discussed further in the following section.

The distribution of pitch is shown in panel B of Figure 4.2. Two characteristics of the distribution are noteworthy. First, the bulk of time was spent looking nearly horizontally. Second, participants exhibited a trend that looking downward was more common than looking upward for a given angular displacement from horizontal.

The distribution of roll is shown in panel C of Figure 4.2. This distribution was highly concentrated, with 95% of samples falling between -17 and 17 degrees.

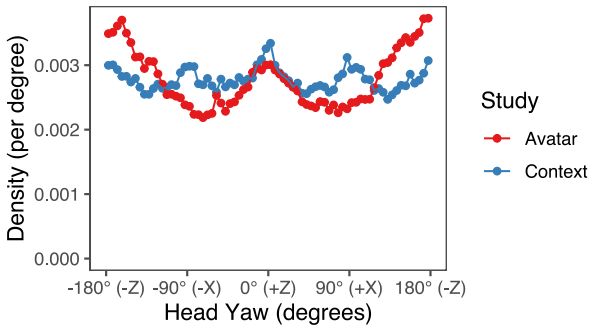

Figure 4.3: Distribution of yaw by study (avatar or context). Context study shows four peaks, avatar study shows two.

Head Orientation Rectilinearity

The distribution of yaw over time shows peaks when yaw is at 0, -90, 90, or 180 degrees. These are directions aligned with the X and Z axis of the underlying coordinate system. There are no obvious indicators of these axes in the virtual space, and in principle every virtual object in the scene can be rotated all together by an arbitrary amount without any perceptible change on the part of the user.

If this is the case, how did the direction of these axes influence participant's behavior? I believe these dimensions are visible through other rectangularly aligned objects, like the walls of a room or the orientation of a bench. I suggest that an environment designer finds it easier to align rectangular world elements with an underlying rectangular grid, and the global XZ grid provides that grid for the designer.

The relation of yaw to the environment is made visually apparent by considering the data from the two studies separately, as in Figure 4.3. In the avatar study there was one environment, a rectangular-shaped high-ceiling room. Instructions for the week's task were posted at the far ends of the room, at +Z and -Z directions. Of these, participants looked more often in the -Z direction, as it was closer. In the Context study, there were 192 separate environments participants saw, and so there was much larger potential variation, as well as three variables that may influence head orientation.

To statistically investigate the alignment of head orientation with the different axes, I define *head orientation rectilinearity*. Given a density function $f(\theta)$, $\theta \in (-\pi, \pi]$ representing the distribution of head yaw, head orientation rectilinearity is:

$$\int_{-\pi}^{\pi} f(\theta) cos(4\theta) d\theta$$

For intuition, consider that the $cos(4\theta)$ term weights angles around 90-degree intervals positively, and angles away from those interval points negatively.

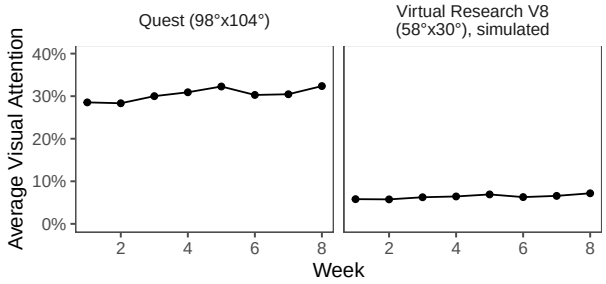

Figure 4.4: Proportion of time one participant was within field-of-view of the other, defined for both the headset in use (Oculus Quest 2) and the range for comparative previous work (Virtual Research V8).

A mixed-effect model was fit to the rectilinearity data with fixed effects of week, motion ability, environment, and visible space and random effects of individual, section, and session within section. All fixed effects were tested for significance. The only statistically significant result was the intercept was different from zero, i.e., that rectilinearity was on average positive ($t(20.59) = 4.801, p < 0.001$), all other terms $p > 0.102$). Participants were more likely to be looking in directions that aligned with the horizontal axes of the global environment's coordinate system than at angles diagonal to the coordinate system.

Head Orientation towards Others

In previous work that investigated gaze over time, it was found that attention to others decreased over the course of the study. Considering the rarity of datasets that collect gaze over a long span of time, I investigated this same effect in the present dataset. I define visual attention for two people to be the percentage of time one person's head is within another person's field of view. Note that this is subtly different from Bailenson and Yee [7] as they use the percentage of time a person has at least one person in their field of view. Averaging this visual attention across all participants across all sessions within a week, I find that the amount of visual attention increases over time, beginning at 28.28% in week one and increasing 0.47% per week ($t(6) = 3.003$, $p = 0.024$). These values are in Figure 4.4 in the left panel.

I also perform an auxiliary analysis on the same data and process, save that the size of the field-of-view is not the Quest's but rather a smaller region, namely the headset in the original work (Virtual Research V8). I do this in the case that the effect in previous work [7] is not due to the field-of-view per se, but rather to the angular dimensions that merely happened to be the field of view in that previous work. The effect was also significant and in the same direction. The average visual attention in the Virtual Research V8's field-of-view began at week one with 5.63% and increased 0.17% per week ($t(6) = 3.852$, $p = 0.008$). These values are in Figure 4.4 in the right panel. In regards to **RQ4**, both analyses indicate that participants looked at each other more over time.

4.3.3 Distance-Gaze Equilibrium

Equilibrium theory [5] posits that two people maintain a constant level of intimacy by balancing two cues, interpersonal distance and gaze directness. In previous work in naturalistic settings [133], balancing these two cues manifested as a negative correlation between indirectness of gaze and distance between people. In contrast to this previous work, I have data very rich in time. However, the non-independence of sample-level data adds significant complexity to a statistical model, and the equilibrium effect is in my case so strong that I do not attempt to summarize the distribution. Instead, I first consider only the moments in time for which the distance between participants is 3.66m or less. This threshold is the same as in [133] and stems from work by Hall [49]. From there, I randomly select a single moment in time from this thresholded set and use this moment in time as a pair-level data point.

To measure the indirectness of gaze, I follow Yee and collaborators [133] by calculating the *gaze sum*. The gaze sum is the the sum of two angles based upon two participant's head positions and orientations at a given moment. The first addend is the angle between participant A's forward vector and participant B's head with the vertex of the angle at participant A's head, and the second addend is analogous for participant B: the angle around participant B's head from participant B's forward direction to participant A's head. To measure interpersonal distance, I followed the same procedure as in subsection 4.3.1 but did not log-transform the data, as the thresholding step disrupted its log-normal distribution.

The analysis was performed with a mixed-effect model with fixed effects of week, gaze sum, and dataset, and random effects of section and session within section. The prototypical participant pair at a gaze-sum of 0 degrees, that is, directly looking at each other, had an distance of 2.66m in week 1 which decreased by a non-significant amount of 0.007m per week ($t(160.1) = -1.114$, $p = 0.267$). The effect of gaze sum on the prototypical pair was -0.00113m per degree ($t(4877) = -5.57.$ $p < 0.001$), meaning the prototypical pair facing the same direction (180 degree gaze sum) was 0.20m closer than the prototypical pair facing each other directly. There was also a significant effect of study on interpersonal distance, such that the prototypical pair in the Context study were 0.14m farther apart than the prototypical pair in the Avatar study ($t(19.2) = 4.511$, $p < 0.001$).

4.4 Discussion

There are several variables that affect distance and affect gaze. In regards to distance, I found that participants increased their interpersonal distance over time **(RQ1)**. At first, this may seem counterintuitive, as participants should become closer by getting to know each other better during this experience. In contrast, I find the reverse because participants adapted to the medium. Hall, in defining proxemic spaces, gives the constraint that the larger end of conversational space is where one can hear another [49]. In these virtual environments, a majority of time was spent without 3D audio enabled. This led to no volume attenuation over distance, and so this restriction is

lifted. This effectively extends conversational space much larger than exists in physical environments. Additionally, participants often needed larger ranges of spaces to complete some of the discussion activities that involved working with 3D models, which shifted the balance in favor of moving away. The opportunity to move away was important, as panoramic spaces also led to greater personal space than constrained spaces **(RQ2)**.

In regards to gaze, I found a pattern that participants tended to look in directions more aligned with the horizontal grid. This was not significantly related to any of the three variables in the context study. I hypothesize that this is either a carry-over from real-world behavior in following rectangularly-aligned seating in a room, or that focal points that draw attention are in or near the centers of walls, rather than corners. In regards to time and social attention, I found that the percentage of time one participant included the other in their field of view increased slightly over time **(RQ4)**. This was different from the results found in Bailenson and Yee [7] that found decreasing attention over time. The most likely explanation I give to this is that headsets are not as heavy as in 2006, and the visual information of looking at another avatar is more useful to the conversation than it was before, due to more fluid tracking and better inverse kinematics. It is worth noting that visual attention was not calculated in the same way (specifically, computing the percentage of participants withing view averaged over time versus the percentage of time at least one participant was in view), so one cannot make absolute comparisons even using the same field-of-view.

I also demonstrate effects due to the pair in each case **(RQ3)**. Variables such as familiarity and liking may have influenced these individual differences, and more follow-up work is necessary. One factor that did affect the distances is the gender composition. In each example, the closest avatars were male-male, followed by female-female, and the farthest pairs were all others. These results do not follow in line with previous VR research [133] and merit further investigation. Future work can also explore the cause of these per-pair effects. For example, it is possible the pair-specific differences in personal space are simply consistency, e.g. participants became comfortable in the relative positions they selected arbitrarily at the beginning of the quarter. It is also possible that certain participants knew each other beforehand and clustered together, certain participants grew to like each other, or consistent factors such as gender could make interpersonal distances predictable beforehand.

Finally, I corroborated previous work [133, 10] on personal distance and mutual gaze. Curiously, both of these values increased over time: participants got farther away while also looking at each other more. Did one cause the other? It is difficult to say. I still believe it is more likely that participants adapted to the medium in both its quirks and its novelty.

Limitations of this study include a lack of preregistration. All of these results are exploratory, as the procedure and analysis were not specified *a priori*. As a field study, there are sacrifices made to control for the sake of realism. Some of those in this dataset included the active/passive motion manipulation and consistent groups and group sizes. Additioanlly, the heterogeneity of the physical settings participants occupied while attending these sessions may have introduced unknown moderators.

While the value of week-scale time was used effectively in these analyses, the value of second- or minute-scale time was not leveraged. Future work ought to explore time-dependent ways to view proxemics, such as models for predicting dynamics or importance of distance in a moment. Future work can also more deeply investigate the inter-dyadic and inter-group effects for interpersonal distance.

4.5 Conclusion

In this work, I report findings regarding the proxemics and gaze of a large longitudinal study in social VR. Participants adapted their interpersonal distances based on affordances in the medium as well as the virtual environments in which they worked. There were also substantial interdyadic and intergroup differences, too. I also found that participants tended to look at each other more over time, contrary to previous longitudinal research. Taken together, these findings encourage future work in understanding adaptations to the medium of VR with more longitudinal studies as well as investigations into the inter-dyad and inter-group differences in these important aspects of human interaction. Finally, understanding human behavior in virtual reality may generalize to human behavior more broadly.

Chapter 5

Nonverbal Synchrony

Synchrony is a construct that has captured much interest since Condon and Ogston's initial observations in the 1960s [26]. The concept has had an impact in psychotherapy [93], creativity [129], education [63], and trust [116]. It has been extended into verbal accommodation, and with the advent of measurements of neural activity, it has been extended there too [47]. It appears to be a fundamental aspect of human communication, but what it *is* is still a subject of debate.

One definition, going back to Bernieri [16], is "the degree to which the behaviors in an interaction are nonrandom, patterned, or synchronized in both timing and form." Naturally, this opens up two follow-up questions: what is a similar timing, and what is a similar form? There are a wide variety of methods previous work has used to operationalize both of these questions (See [31] or [6] for a review).

Importantly, this is not simply a methodological question. As Hale and collaborators point out [48], the characteristics of measures that distinguish synchrony (and distinguish it to different degrees) can be used as springboards for theoretical discussion and empirical validation. For example, some work uses global body motion [39], but some work separates out head from hands, or left hand from right hand [71, 110]. Different body parts likely carry different meanings, especially in the VR context, where one can argue that head movements are used primarily for visual attention and nonverbal signaling while hand movements are more involved in manipulating objects and using hand controller interfaces. Global body motion may register more closely to arousal or energy than individual points. The dimensions on which synchrony occurs (and the relative strength of their synchronization) also matter for similar reasons. Head yaw may indicate watching the speaker role transition from one group member to another [71] whereas head pitch may indicate nodding [48]. Furthermore, how does one define these dimensions within a shared space? If one claims that people synchronize in the way they "move forward," that "forward" could be relative to the environment, similar to following a group, relative to the direction one is facing, which may be more amenable to mirror neurons, or perhaps relative to the position of the other member of the conversation, indicating a maintenance of personal space. Extending into virtual reality, what is the status of

virtual movement? Is synchrony only present in my physical bodies and motions, or can more symbol, abstracted, conscious motions still be synchronous?

Beyond the question of what is moving, there are questions regarding how that motion is perceived. As illustrated by Hale and collaborators [48], the time range and models of lag can give evidence of mechanism: in their case, short time-scale lag (600ms) was consistent with activity of mirror neurons. If this is the sole root of synchrony, then it should be less visible with larger window sizes. On the other hand, many human activities have a wide range of time scales, most famously detailed in Newell's "Time Scales of Human Action" [81, p. 122]. Finally, how are large and small movements integrated together into a measure and a perception of synchrony? On one hand, the use of Pearson correlation asserts motions that are ten times larger are ten times as indicative of synchrony. On the other hand, Spearman (rank) correlation focuses on small, consistent similarities in the stream of motion.

Through this short description of some of the ways operationalizations of synchrony have varied across recent work, it is clear that the mechanisms of synchrony are not agreed upon. However, examination of these operationalizations could begin to tease them apart. I look to content validity and consistency to perform this. Using content validity, I look for measures that effectively distinguish between synchronous and pseudosynchronous interactions, and dismiss measures that do not perform this effectively. Using consistency, I see which measures of synchrony are similar to others, and reduce the search space accordingly.

The contributions of this work are (a) a specification of the many choices required when operationalizing synchrony, along with theoretical significance of these choices, (b) application of a recently developed methodological technique, the multiverse, to the study of synchrony, (c) the introduction of a novel multiverse sampling approach when studying large and computationally expensive multiverses, (d) extensive analysis of the content validity and consistency across this space of measures built upon previous literature, (e) investigation of the variations of branches that show consistency in order to examine psychological mechanisms that might be driving the differences, and (f) recommendations for synchrony computation.

5.1 Related Work

The study of synchrony has had a wide impact in several domains, but its nature is unclear. Works led by Bernieri [16], Delaherche [31] and Ayache [6] detail important distinctions in the literature, specifically between synchrony as an unconscious phenomenon and coordination on a goal-directed task, mimicry and contingency, and the static and the dynamic.

One variable of note, present in Berneri's discussion but highlighted by Ayache, is the degree and nature of coordination, with a distinction drawn between synchrony as an unconscious phenomenon and synchrony as coordination on a goal-directed task. For example, the experimental setting could be so sparse such that the participants are visible to each other but are not supposed to interact (by social or experimental constraints). One step up, the participants may be able to

interact, but have no explicit[1] shared context or goal. Participants could have a more explicit shared goal, such as brainstorming [71]. Continuing on towards more coordination, this shared task could involve repetitive mechanical motion, and presuming participants would like to avoid collision, may synchronize [3]. At the most directly coupled end, participants may in fact directly share mechanical movement in some activity like dance. This example spectrum of coordination is neither comprehensive nor perfectly clear, but it does highlight that the possible causes of coordination can be very different depending on what amount of coordination is taken for granted. my work falls between the shared-task type without much mechanical influence, like a therapy session, and shared-task with some mechanical constraints, because participants were sometimes designing virtual spaces.

A second dimension on which to place my work is between mimicry and contingency, depending on whether the modality of what participants synchronize upon is the same (mimicry) or different (contingency) [14]. The difficulty with this distinction is, again, that it is not a binary but a continuum. One participant's head motion may link with another's hand motion, and it be considered contingency. However, if that motion is collapsed to motion globally, it is now mimicry. Similarly, some pairing that is (on its face) mimicry, intense hand movements at a similar time, may, upon further investigation, be contingency when broken out by directions of motion. The point is that the distinction between contingency and mimicry is not binary *per se*, but is binary as the result of an implied metric of behavior similarity, upon which some threshold is placed. In my work, I largely investigate mimicry due to computation constraints, but I do explore this difference in synchrony between different body parts, i.e., one person's head and another's left hand, in subsection 5.3.4.

Finally, there is the distinction placed between the static and the dynamic. Delaherche draws a distinction between constructs like alignment and mimicry, which tend to focus on static features, and their definition of synchrony, which draws from dynamic features. The example they give is illustrative: "For instance, two people sitting with crossed legs or looking in the same direction are exhibiting either mirroring or the chameleon effect. This behavior becomes a matter of synchrony if they cross or uncross their legs at the same time or gaze in the same direction simultaneously" [31, p. 3]. Again, this dimension is not to be confused to be a binary when it is in fact another continuum. What is implicit in this distinction in some ontology of states and transitions. In the example, the states are whether legs are crossed, or the direction in which one is looking, and correlation of motion speed would be indicative of synchrony. However, if the states were defined as moving and stationary, which might be reasonable in a setting such as a large walkable area like a theme park, then correlation of motion speed is mimicry, because it's related to how often people are in the same state, rather than transitioning together.

[1] This distinction is made because there is always an implicit shared goal and a common understanding of what a conversation is, even if it is due to simply the nature of the experiment being about conversation.

5.1.1 Synchrony Measurement

Beyond the variations in synchrony theory, synchrony has had a history of being difficult to measure. Perhaps best described in a review by Delaherche and collaborators [31], "a method for the objective evaluation of synchrony remains somewhat elusive despite being heavily targeted by researchers." This is not to say the work has been weak: there have been significant strides in methods over this time, including the introduction of comparisons with pseudosynchronous interactions that provide a stronger null hypothesis to test against [16]. Synchrony has been the subject of a plethora of measures, ranging from a holistic report by naive viewers, to relationships on activities annotated by trained coders, correlation, windowed correlation, peak-picking, wavelet analysis, and even information theory methods like cross-recurrence quantification analysis. For a review of these methods, see [83].

Recently, there has been an increased focus on measures of synchrony as an object of study, driven in part by mixed results and unclear influences [6]. Variations in the ability for measures to detect synchrony may give evidence for certain mechanisms of synchrony [48] or indicate facets of synchrony that ought to be distinguished when reviewing past work or designing future work [103]. Work by Novotny and Bente [83] looked to investigate the relationship between observer's judgements of synchrony and a set of algorithms often used to calculate synchrony. They find that, overall, Pearson correlation and phase synchrony related the strongest to observers' judgements. Schoenherr and collaborators [103] find that seven common measures of synchrony did not fit a common factor model, but instead an exploratory analysis showed three factors, which they label as *average strength*, *maximum strength* and *frequency of nonverbal synchrony*. (Note that frequency here is not meant in the technical, wavelike sense, but a more colloquial sense, like commonness.)

5.1.2 Time Scales of Synchrony

The time scales on which people synchronize is valuable as a research topic because it can narrow down mechanisms and correlates of synchrony. It is clear human behavior can be studied on a wide range of time scales [81] and doing so, while not the norm, can provide new insights into behavior [92]. The time scale that is most common in research is relatively short, in the 1-5s range (0.2Hz-1Hz). This time range is easy to observe behavior in, and provides many potential repetitions and synchronizations in short-to-medium length experiments. This time range has even been codified in the development of windowed cross-correlation, which in the initial work by Boker and collaborators, the window size was limited to 4 seconds.

Most investigation of the time scales of synchrony has been performed with a method called "wavelet analysis" [40] that simultaneously breaks down signals by time and frequency. With this method, synchrony is evidenced by *coherence*, a wavelet analysis analog of the coefficient of determination (R^2) that estimates how much of one participant's motion at the specified frequency can be explained by the other participant's motion at that frequency.

Hale and collaborators [48] use wavelet analysis to provide evidence consistent with mirror neuron

mechanisms of synchrony. They found coherence in the 1s-5s period band (0.2Hz-1Hz) and estimated constant-lag offset of 600ms (as opposed to constant-phase). They conclude, based on the constant-lag model, that synchrony is reactive, rather than predicted by the person, which would have been evidenced by a shorter lag time nearer to 0ms, or memory-driven, which would have been evidenced by a larger lag time of up to thirty seconds.

Fujiwara and Daibo [40] investigate the relationship of time scales and synchrony directly and find that that coherence varies by time scale (frequency). In fact, they find a trend that longer time scales show higher coherence, specifically that coherence is strongest from a period of 2s (0.5Hz) onward.

However, these works are largely limited in their ability to detect synchrony on larger time scales due to the short time spans of interaction. Hale uses 90-second sessions, and Fujiwara uses 6-minute sessions. Considering there are rhythms across many time periods [58] and evidence that many behavioral patterns demonstrate patterns across many scales [44], it is possible synchrony occurs on even longer time periods as well. In this work, I investigate this possibility.

5.1.3 Synchrony in Virtual Reality

The use of virtual reality (VR) in the study of synchrony has been growing. The ability of VR to function as an effective tool in the study of social psychology is not new [17], but with the increased power and availability of powerful data analysis methods, VR's ability to both capture behavior at high spatial and temporal resolution and create otherwise impossible situations is valuable to researchers [132].

The earliest work on synchrony in VR was done by Bailenson and collaborators [12], studying the effect of an agent programmed to mimic a user. This was based upon the "chameleon effect" that indicated mimicry could lead to more liking [24]. Follow up work on this concept showed a similar effect with handshakes [13]. Virtual mimicry increased social connection [117], sales [121], trust [48]. Vrijsen and collaborators found that mimicry was lower in populations with social anxiety [120]. Aburumman and collaborators found higher trust with an agent that included mimicry and nodding as opposed to an unreactive agent [1]. However, Zhang and Healey [135] replicated the original study on persuasiveness with a population of 52 participants and did not replicate the effect.

In regards to the spontaneous coordination between two real people, the first work demonstrating this was performed by Sun and collaborators [110] demonstrating synchrony through correlation of head translation over time. They also found negative correlation when relating left hand to left hand and right hand to right hand, which they interpreted as turn taking.

The finding that virtual reality still allowed synchrony was corroborated by Miller et al. [71], who also expanded into the study of triads in VR. In their study of triads performing a design task, they manipulated both environment and task and found that the environment (virtual conference room vs. virtual garage) affected synchrony. They also reported that participants had higher synchrony (defined as the product of normalized speed) when participants were *not* looking at each other

compared to when they were. my work uses virtual reality to continue to explore the data capture possibilities of the medium.

5.1.4 Research Questions

The results section is structured to answer three research questions. These questions use this dataset as a testbed to explore the similarities and differences between synchrony measures and ultimately provoke research and discussion regarding the causes of these similarities and differences.

RQ1: What options, if any, lead to poor content validity? To set up the following analyses, I remove options from branches that on average do not distinguish between synchronous and pseudosynchronous interactions.

RQ2: How much consistency do options have within each branch? Branches with high consistency can be collapsed across for the purposes of this analysis. Branches with low consistency should be specified whenever an analysis is performed, to be distinguished when drawing from related work, and the differences among them can be explained.

RQ3: How much consistency do these measures as a whole have? This provides a better picture of the similarity among measures of synchrony within the literature as a whole. From this, one may expect to see either a degree of consistency or facets of synchrony. This informs how much to consider related work when drawing conclusions from one setting to another.

/draftpar *RQ4: How well do these measures select for synchrony in a face-valid synchronous activity?* One of the manipulations in the period 1 study was the performance of intentionally synchronous actions. This period of time was excluded when calculating synchrony, but it is worth investigating synchrony to see if these measures capture what is defined as a synchronous activity.

RQ5: How much do these measures relate to constructs related to synchrony in previous work? Often, synchrony is the subject of study because it is related to a construct of interest. One additional method I can use in determining which measures are best for synchrony is the measure's ability to predict other constructs.

5.2 Methods

In addition to chapter 3 devoted to this dataset, I report the results from 230 participants because 2 participants each attended only one session and did not partake in that session long enough to meet the minimum threshold of duration for this analysis (5 minutes). The remaining portion of this methods section are approaches taken that are unique to this work.

5.2.1 Multiverse Analysis

In this work, I make significant use of a technique called "multiverse analysis" [109]. This technique aims to increase the transparency of the data analysis step of the research process by reporting many

analyses created from an exhaustive or nearly-exhaustive combination of "all reasonable specifications" [105]. These analysis pathways are the 'universes' within the 'multiverse', which is a metaphor leveraging the many-worlds (multiverse) interpretation of quantum mechanics. This is in contrast to common data analysis techniques, in which a researcher may explore several analyses in a haphazard manner, leading to selective reporting of results and inflation of false positive rates. This is also in contrast to the common preregistered analysis technique, in which only one analysis pathway is proposed ahead of time and performed, and any changes to this pathway are given transparently in preregistration amendments[2].

To describe a multiverse analysis in general as well as the particular analysis I are performing, I use the terminology proposed by Sarma and collaborators in the R package "multiverse" [101]. A multiverse consists of several *branches*, which are points in an analysis where several reasonable *options* could be taken. Each option can have *conditionals* that do not permit certain pairings of options. Ultimately, a *universe* is a selection of one option per branch that is consistent with all options' conditionals. For further discussion of the ontology of a multiverse analysis, I refer the reader to the original paper [109] and the multiverse R package [101].

Because of the recent development of multiverse analysis as a technique, the norms around the framing of the multiverse and the interpretation of its results are still developing. In this work, I consider the multiverse with distinctions drawn by Del Giudice and Gangestad [30] between *equivalent*, *non-equivalent*, and *uncertain* branches. The difference indicates the researcher's assertion of the nature of variation between options within a branch. In an *equivalent* branch, the options are asserted as equivalent: there is no distinction of substance the researcher wishes to make between the options, and any effect of interest ought to be responsive to the options in general. Researchers could also assert a branch to be *non-equivalent*. Del Giudice and Gangestad note that in this case, a multiverse ought not to collapse across these distinctions, and in fact several multiverses should be performed and analyzed separately. Finally, it can be the case that the researcher is *uncertain* as to whether the options are equivalent. In this case, Del Giudice and Gangestad recommend carrying out multiverses as a "deliberately exploratory endeavor" [30].

I note that the assignment to a branch to one of these categories, like many methodological choices, are influenced by the previous literature and the task at hand, and are ultimately the responsibility of the researcher. I begin by asserting all branches to have uncertain equivalence, then use the results I find in the initial analysis to determine whether branches are non-equivalent, equivalent, or remain uncertain. I do this to begin with the broadest possible approach to synchrony, and then use the narrowing process as information to test against theories of synchrony.

5.2.2 Synchrony Measurement Specification

The branches within this multiverse I have performed are *body parts, signed motion, dimension, coordinate system, physical or visible motion, window size,* and *magnitude transformation.*

[2]The operation of a multiverse is not mutually exclusive with preregistration, but the point I wish to make here is that they both approach transparency differently.

Body Parts

The first question to answer when defining a measure of nonverbal synchrony is the points on which to track motion. While some work uses motion all together [39, 45] some work separates head and hand motion [110],or uses only one tracked point [48, 71]. The options for this branch are *head, left hand, right hand, hands,* or *all*. The first three individual points correspond to each tracking device: the headset, left hand controller, and right hand controller, respectively. *Hands* is defined as the sum of motion of both hands, and *all* is defined as the sum of motion of all three tracked points.

Signed Motion

The type of motion is the next branch. In this, the options are either *speed* or *velocity*. Previous work largely uses speed, i.e., motion ambiguous to direction [45], but there are indications that in-phase and anti-phase synchrony is different [41]. Measures of *speed* are guaranteed to be positive, and measures of *velocity* are signed.

Dimension

The previous branch, signed motion, influences the next branch, dimension. Some previous work has used individual dimensions [48, 18] and others have used combined motion [110, 71]. Options for the branch are *translation, horizontal, vertical, forward, rightward, rotation, yaw, pitch,* and *roll*. The first five options use translation (i.e., change in position), and the last four use rotation (i.e, change in orientation). *Vertical, forward,* and *rightward* dimensions are the simplest: they represent either the speed or velocity (depending on the signed motion branch) of the body parts in the analysis along the respective dimension. *Yaw, pitch,* and *roll* work similarly, except for rotation. The remaining options, *translation, horizontal,* and *rotation,* all are only valid with the *speed* option in the signed motion branch. Translation is the *magnitude* of the velocity vector in all three dimensions of Euclidean space, and *horizontal* is the magnitude of the velocity vector projected down to the horizontal plane. *Rotation* is the rotational speed around the axis defining the rotation.

Coordinate System

For the options of *forward* and *rightward*, there must be a further specification of the coordinate system that defines forward and rightward. While most previous work does not distinguish this operationalization, they are often constant and designed around using the setting for the study. The options for this branch are *none, global, self-direction, other-position, group-median,* and *group-mean*. *None* represents all options for dimension other than forward and rightward. *Global* uses the globally defined coordinate system, in which Z is taken to be forward and X is taken to be rightward. *Self-direction* defines forward and rightward relative to the user's head direction, I.e., the forward is defined as the direction of the headset projected to the horizontal plane, and rightward is the remaining orthogonal vector. *Other-position* defines the forward vector as the direction from

the user's head to the other user with whom synchrony is being calculated (again, projected to the horizontal plane), and the rightward vector as the remaining vector orthogonal to vertical and forward. *Group-mean* and *group-median* define forward with respect to the group center, either determined by the 3D mean of others in the group, or the geometric median, which is robust to outliers. Again, this vector is projected to the horizontal plane, and the rightward vector is defined as the vector orthogonal to vertical and forward.

Taken together, the branches of body parts, signed motion, dimension, and coordinate system define the movement under study.

Physical or Visible Motion

As this experiment is run in immersive, headset-based virtual reality, there remains an open question whether to treat virtual motion generated by abstract controls (e.g., teleporting, joystick movement) separate from virtual motion generated by physical motion (e.g., leaning forward, stepping to the side). The *physical* option is concerned only with physical movement: this branch ignores any time point in which teleporting or joystick movement was used. The *visible* option is concerned with all motion that was visible to the participant, including teleporting and joystick movement.

Window Size

The final two branches are concerned more with the perception of motion. The first indicates whether to perform a windowed correlation, and if so, how large the window should be. The window sizes are in terms of samples and are *full* (regular correlation), *10* (1/3s), *100* (3.3s), *1000* (33.3s), and *10000* (5.56 minutes).[3] This is a broad spectrum of window sizes, extending what is often performed in the literature [18, 103]. I do this to explore a representative space of values in between very short and very long.

Magnitude Transformation

Finally, once the window size is established, there can be a transformation performed on the one-dimensional signal before correlation. The six options are *rank* (equivalent to Spearman correlation), *none* (equivalent to Pearson correlation), and *log-plus-p*, where p is one of 0.001, 0.01, 0.1, or 1. The log-plus transforms also preserve the number's sign and always map zero in the input to zero in the output. This allows its use for cases in which the motion is signed (I.e., velocity). Mathematically, this is

$$\textsc{LogPlusP}(x; p) = sign(x)(\log(|x| + p) - \log(p))$$

The final step is to relate these two streams of data using a simple correlation to produce a single value of synchrony for a given pair of participants in the same meeting.

[3] I make a point to report the period of these waves, rather than their frequency, as it is easier to imagine time periods than rates, especially when the rates are almost all below 1.

5.2.3 Sampling from Multiverses

In total, there are 9,300 universes within this multiverse. This, combined with the fact that there were 5341 attendance pairs in the dataset, each of which had approximately 50,000 frames of motion, made it computationally prohibitive to calculate all measures upon all attendance pairs. With an empirically estimated value of one second per analysis (one for each combination of universe and pair) it would have taken over a year and a half of continuous computation for the analysis to complete. Instead, I propose a method of sampling the multiverse such that conclusions can be drawn both based on the space as a whole and on direct comparisons between options. The mechanics of sampling a large multiverse are nontrivial, and are described more fully in Appendix A.

The analyses described in this paper use one of three samples. The first sample is the *sparse* sample. For each pair of the 5341, there were 10 measures selected uniformly at random out of the 9300 possible. This permitted a uniform sample from all measures so that conclusions could be drawn about options, as is done for RQ1.

The second sample is the *edge* sample. In this sample, there is one *path* per pair (produced as described in Appendix A) consisting of 10 to 13 evaluated universes that, when placed in order, changed by a minimal number of branches. Rather than sampling universes, the goal of this sample is to sample edges so that the similarity (a function of pairs, i.e., edges) could be accurately represented.

The third sample is a *dense* sample. This sample is more akin to a multiverse analysis insofar as each universe was applied to each unit of analysis (in this case, a pair). However, its drawback is that only 30 of 9300 universes and only 265 pairs of 5413 were used. Each analysis indicates which sample was used.

5.2.4 Evaluation Criteria

The results I report in this section follow two types of analysis. The first is a study of a measure's content validity, which I define for synchrony as the ability to distinguish true interactions from pseudo-interactions [16]. Within the second type, consistency, correlations are made between synchrony scores given the same data. This is an estimate of how similar two measures are. I also report face validity, the measured synchrony as defined on an interaction designed *a priori* to be synchronous, and predictive validity, i.e., the ability for synchrony measures to predict related constructs.

Content Validity

Based upon Bernieri's definition of synchrony [16], I assert that a good measure of synchrony is in fact one that effectively *distinguishes* between synchronous and pseudosynchronous interactions. To operationalize this, consider one pair and one synchrony measure. I compute a synchrony score on the real data, and also create a pseudosynchronous distribution based upon thirty pseudo-interactions formed by shifting one participant's time series by a random amount (uniformly distributed) and rolling over the extra data from the end to the beginning. The z-score of true synchrony relative

to this distribution of pseudosynchronous values is calculated, which indicates the ability of the measure represented by the selected universe to distinguish between real and pseudosynchronous interactions. I call the score itself *distinguishability*, which I use to then infer the content validity of the measure or set of measures.

There are other methods of defining pseudosynchronous interactions. For example, Hale and collaborators [48] define pseudosynchronous interactions by maintaining a rough time alignment but changing which participants are interacting with each other. In their example, if dyad members A and B have a real interaction, then a pseudosynchronous interaction would be an pseudo-interaction between A and C at the same week. One could also vary session, calculating the synchrony between participants A and B, but it is the motion of A at week 1 and the motion of B at week 2. It is clear that each of these approaches would produce an interaction that does not have the statistical signature of a true interaction. I selected a pseudosynchronous interaction for a shifted time because it maintained variation due to both participants and week, which I judged to be larger sources of variance as opposed to variation due to task.

Consistency

In addition to content validity, I investigate *consistency*, the degree to which these measures of synchrony are capturing similar portions of variance. This is done by finding the correlation between pairs of measures applied to the same data. Many pairs are produced according to the method specified in Appendix A, and the correlation between the measures of synchrony are reported as *similarity*. The difference between *similarity* and *consistency* is that I use similarity to refer to the correlation itself, whereas consistency is the property that I wish to infer from the correlation. The measures may vary only by options in one branch (subsection 5.3.2), or by several (subsection 5.3.3).

5.3 Results

In this work, I begin with the broad multiverse analysis, and then use its results to motivate follow-up analyses. I first report results from the large multiverse, and conclude the options in branches of the multiverse that I find valid. From there, I explore further the dependencies and similarities on each branch in the multiverse that remains of uncertain equivalence.

The first and most basic question in this work is the results of the multiverse analysis. In this work, I aim to continue the threads of work by Schoenherr et al [103], and Novotny et al. [83] that query variations in synchrony measurement. To begin, I perform one multiverse analysis asserting all options within each branch to be of uncertain equivalence. I note that to do this, I must assert *a priori* that synchrony is occurring in this setting, and the question is simply under what measures it occurs. Using this first analysis, I then sort the branches into non-equivalent, equivalent, and uncertain categories (see the section on multiverse analysis), and perform follow-up analyses. I use content validity to establish principled non-equivalence (RQ1), and I use consistency to establish

principled equivalence (RQ2). Values that show neither principled non-equivalence nor principled equivalence are left as uncertain.

In the following sections, I follow this process to conclude the distinction between velocity and speed (signed motion) is non-equivalent and focus on speed only. I also conclude that distinctions in the branches on dimension, coordinate system, and physical or visible motion are equivalent, and so do not investigate these distinctions separately but instead collapse results across the uniformly sampled options. The remaining branches, body parts, window size, and magnitude transformation, remain uncertain, so for each analysis I also provide results broken down by the options in each of these branches.

5.3.1 Content Validity (RQ1)

In this process, I investigate the content validity of options within the multiverse as a whole. To do this, I sample ten universes per pair upon which I can compute synchrony. The aggregation of the distribution of distinguishability z-scores are plotted for each branch, separated out by the options within each branch. This can show the average effect of selecting one option as opposed to another within a given branch, and in turn whether measures of that type can distinguish synchronous and pseudosynchronous interactions, and therefore be an effective measure of synchrony.

Surprisingly, no single option precludes the measurement of synchrony, at least at the power of this dataset I are working with. The weakest distinguishability, selection of velocity in the signed motion branch, is still significant (M = 0.114, SD = 1.41, one-sample t-test with $t(23103) = 12.2$, $p < 10^{-33}$. All other options in all other branches can, on average, distinguish synchrony (all $p < 10^{-188}$).

I interpret these results such that the difference between velocity-based synchrony and speed-based synchrony to be so great as to be non-equivalent. In this work, I ignore measures of synchrony using velocity and only use speed in all further analyses.

Note that these results do not imply *any* measure (other than ones involving velocity) investigated here has content validity: rather, what is show here is that measures involving a certain option, (e.g. head for choice of body parts), *and randomly sampling from the options in other branches, (e.g. equal representation of options for magnitude transform, dimension, window size, etc.)* seem to have content validity. This distinction is like the distinction the ecological inference fallacy makes. It would be ideal to investigate each measure of synchrony individually, but as detailed in the subsection on sampling the multiverse, the exponentially large space of measures prohibits this type of analysis.

I also caution the reader in interpreting these summary statistics as 'more authoritative' or 'more extensive' than individual measures of synchrony. As Del Giudice and Gangestad point out [30], the goal of the multiverse, especially an exploratory one, is to 'deflate' the multiverse, drawing distinctions between types of measures, and (if justified by the results) dismissing part of the space of measures initially explored.

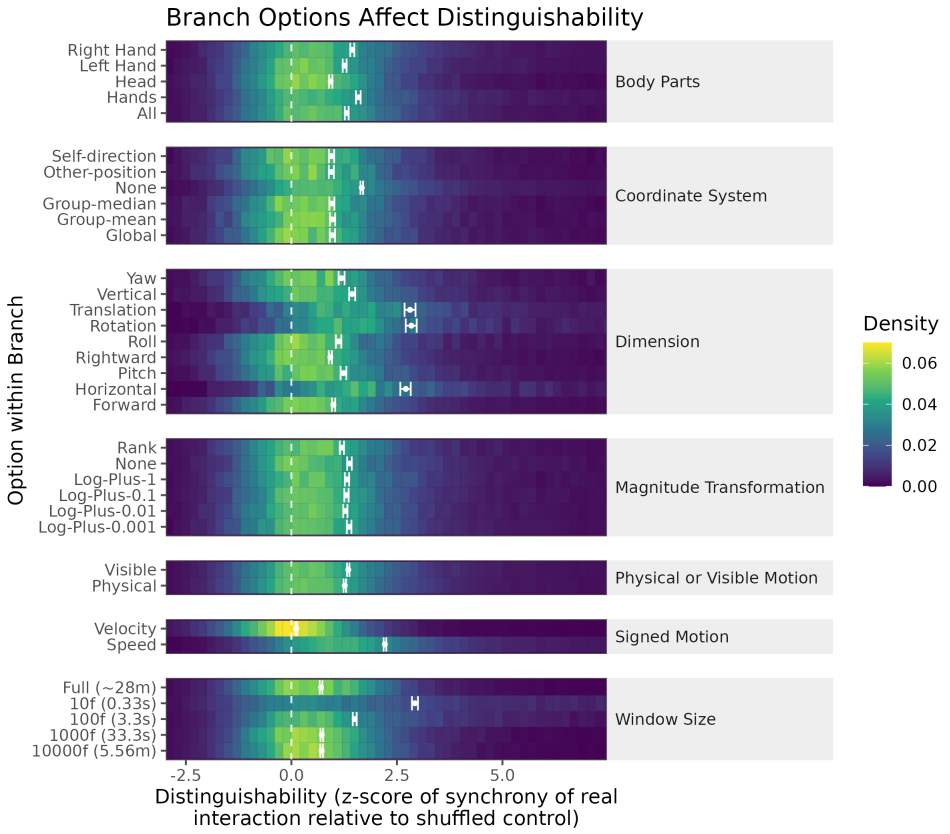

Figure 5.1: Average distinguishability of measures involving a specified option. All show some degree of distinguishability; Velocity in the Signed Motion branch is the weakest. Horizontal position indicates distinguishability, vertical position indicates the options grouped together by a branch. Color represents density, indicating the distribution of distinguishability scores for each option. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals. Each combination of universe and pair sampled for this analysis is represented seven times in this plot, once per branch (facet).

5.3.2 Consistency (RQ2)

In this process, I investigate the consistency between measures that vary only by the option selected in one branch, e.g., on average, how much does a measure using head motion correlate with a measure using hand motion? I do this by pairing the synchrony scores calculated by a pair of measures equivalent in all other branches (e.g., no magnitude transform, vertical dimension, etc.) on the same data (e.g, participants D and F within the meeting of section 5 in week 5). Then, another pair of synchrony scores is calculated, which involves a different selection of other branches shared by the two measures, and a different pair of participants shared by the two measures.

The results in Figure 5.2 indicate which branches have sufficient consistency among their options to be considered equivalent and which branches remain uncertain. If a majority of correlations were above $r_c > 0.80$, I judged the branch to be equivalent for the purposes of this for the purposes of this analysis. These include the branches for dimension, coordinate system, and physical / visible motion. I interpret these results to indicate that the primary underlying factor on which people synchronize naturally within this dataset is the *degree* of motion, not specific to any spatial selection of motion. This does not preclude that dimension, coordinate system, and virtual vs. physical motion matter as a secondary effect, but simply means that amount of motion as a whole dominates the detection of synchrony.

5.3.3 Individual Synchrony Measures (RQ3)

To provide another look into the multiverse, I also sample 30 of the universes from the multiverse (ignoring velocity) and compute the measures of synchrony they represent upon a subsample of randomly selected pair recordings. The set is reported in Table 5.1. This complete set can then be used to estimate the effectiveness of measures in general, not simply the average of measures that include a given option, and the similarity between two randomly chosen measures, not just measures that vary by one option in one branch.

All together, the measures correlate with each other somewhat (M = 0.414, SD = 0.207). This means there is some cohesiveness in the measures - all 30 selected correlated positively. However, the weak correlation between values greatly increases the degrees of freedom for researchers to find spurious correlations. Given thirty participants, measures that correlate at even r = .60 can inflate the false positive rate by approximately 2 if they are both tested and selectively reported. On the other hand, a significant correlation with one measure has a small (7%) chance of also being significantly related to another measure at r=.60. The large number of synchrony measures available in the literature, combined with the weak correlations among them, poses a serious methodological problem for consistency and for false positives.

The following subsections in the results section take a deeper look into the remaining uncertain branches, specifically, body parts, window size, magnitude transform, and physical or visible motion. The options within these branches reliably measure synchrony, I.e., they distinguish between real and pseudo-interactions, but different options capture different portions of synchrony, as indicated

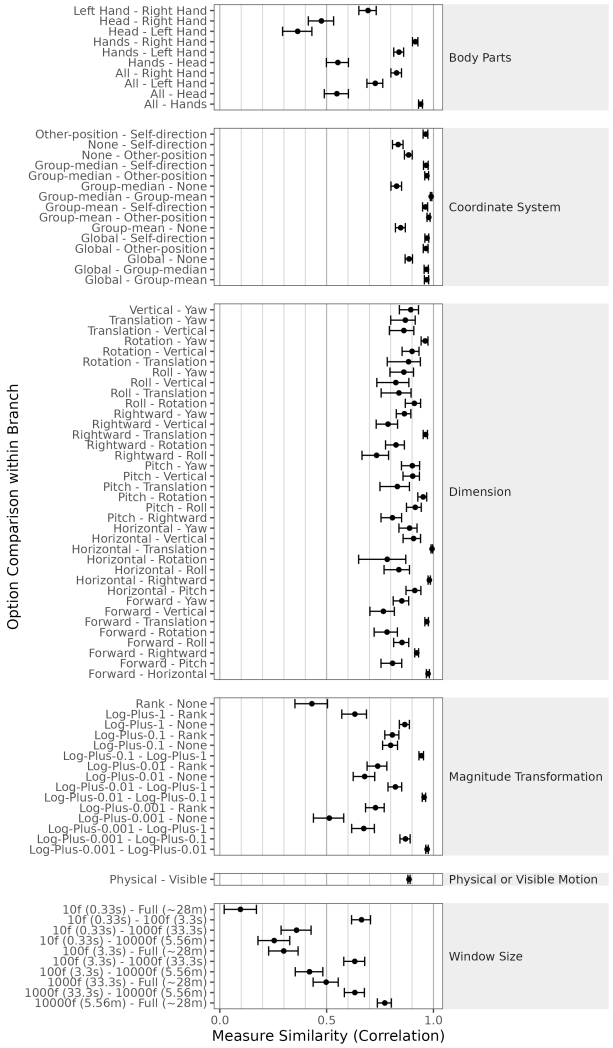

Figure 5.2: Average similarity between measures that vary by one branch. All relationships are positive, but some are quite weak (nine pairs have $r_c < 0.5$). Horizontal position indicates similarity (correlation, domain of -1 to 1). Vertical position indicates the pair of options that are compared, grouped by branch. Error bars indicate 95% confidence intervals of the correlation value.

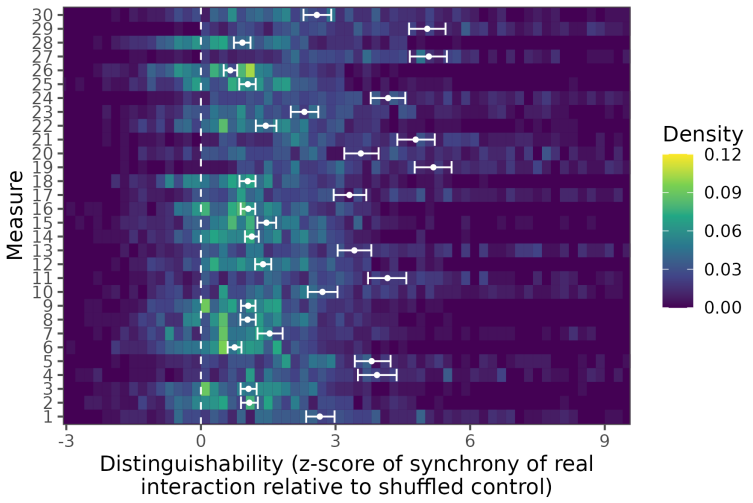

Figure 5.3: Distinguishability of thirty randomly chosen measures of synchrony. All show some degree of distinguishability. Horizontal position indicates distinguishability, vertical position indicates the measure. Color represents density, indicating the distribution of distinguishability scores for each measure. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals. Each pair of the 261 is represented once in each row.

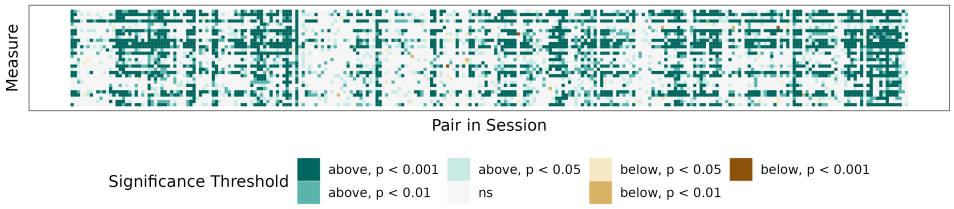

Figure 5.4: Distinguishability for each pair within the dense dataset by each of the 30 selected measures. Strong banding both horizontally and vertically indicates that some measures are more sensitive than others, and some pairs are more easily distinguished than others. The horizontal axis refers to one pair within a session, each of the 261 that have been sampled. The vertical axis refers to a single measure of synchrony, of the 30 sampled. Color indicates the significance threshold in detecting synchrony within that pair's motion.

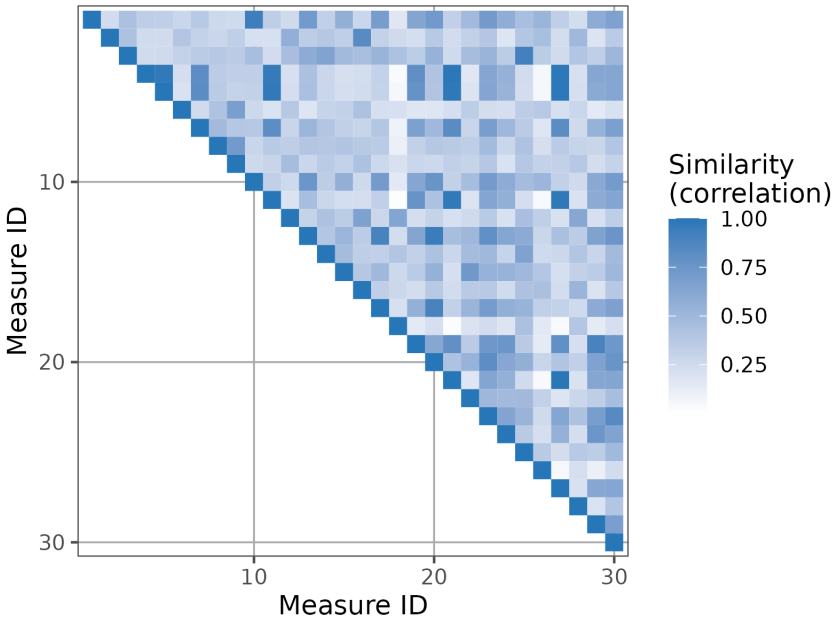

Figure 5.5: Similarity among measures randomly selected from the space of all measures considered. Similarity is positive (blue scale) but is usually weaker than 0.5 and often weaker than 0.25. Horizontal and vertical axes indicate the measure from the 30 sampled. Color indicates similarity, the correlation between measures on the same pairs. The diagonal represents each measure correlating with itself at $r_c = 1$.

Table 5.1: Specification of the 30 measures in the dense sample.

	Body Parts	Coordinate System	Dimension	Window Size	Magnitude Transformation	Physical or Visible Motion
1	Hands	None	Vertical	100f (3.3s)	Rank	Physical
2	Hands	None	Pitch	Full (˜28m)	Log-Plus-1	Visible
3	Left Hand	Self-direction	Forward	1000f (33.3s)	Log-Plus-0.01	Visible
4	Head	Other-position	Forward	10f (0.33s)	Log-Plus-0.01	Visible
5	Head	Self-direction	Forward	10f (0.33s)	Log-Plus-1	Physical
6	Head	None	Roll	Full (˜28m)	Log-Plus-0.001	Visible
7	Head	None	Horizontal	100f (3.3s)	Log-Plus-0.1	Visible
8	Head	None	Yaw	1000f (33.3s)	Log-Plus-0.1	Physical
9	Head	None	Yaw	10000f (5.56m)	Rank	Physical
10	Hands	Global	Rightward	100f (3.3s)	Rank	Visible
11	Head	None	Rotation	10f (0.33s)	Rank	Physical
12	Hands	None	Horizontal	10000f (5.56m)	Log-Plus-0.1	Visible
13	Left Hand	None	Translation	100f (3.3s)	Log-Plus-0.001	Visible
14	All	Group-mean	Rightward	1000f (33.3s)	Log-Plus-1	Visible
15	Right Hand	Group-mean	Rightward	1000f (33.3s)	Rank	Physical
16	Hands	None	Pitch	10000f (5.56m)	Log-Plus-0.1	Physical
17	Left Hand	None	Roll	100f (3.3s)	Log-Plus-0.001	Visible
18	Left Hand	Group-median	Rightward	10000f (5.56m)	Log-Plus-1	Visible
19	All	None	Roll	10f (0.33s)	Log-Plus-1	Visible
20	Hands	Self-direction	Rightward	100f (3.3s)	Log-Plus-0.001	Visible
21	Head	None	Horizontal	10f (0.33s)	Log-Plus-0.001	Visible
22	Right Hand	None	Roll	1000f (33.3s)	Rank	Visible
23	All	None	Horizontal	100f (3.3s)	Log-Plus-0.01	Physical
24	Right Hand	Global	Rightward	10f (0.33s)	Rank	Physical
25	Left Hand	Global	Rightward	1000f (33.3s)	Log-Plus-0.01	Visible
26	Left Hand	None	Vertical	Full (˜28m)	Rank	Physical
27	Head	None	Horizontal	10f (0.33s)	Log-Plus-1	Physical
28	All	Group-mean	Forward	10000f (5.56m)	None	Physical
29	Right Hand	None	Yaw	10f (0.33s)	Log-Plus-1	Visible
30	All	Self-direction	Forward	100f (3.3s)	None	Visible

by their low-to-medium correlations among the options. Within each of the following sections, I dive deeper into the nuances of each branch's options.

5.3.4 Body Parts

The choice of which points to track and synchronize is a varied one. In one part of previous work, largely influenced by the method of motion energy analysis [45], one person's motion is taken together, holistically. In other work, enabled by human pose estimation techniques like OpenPose [38], the Kinect [129], or virtual reality tracking [110, 71], various points are tracked upon the body and can be distinguished in the analysis process.

These two options can be framed using work by Gaziv et al [43] that has broken down the variance in speed of body parts using principal component analysis (PCA). The strongest principal component is roughly the entire body together, capturing 57-97% of variance across the 12 dyadic sessions they captured. The second is motion of the hands and arms, explaining about ten percent of variance. While the third is not significant according to the PCA methods Gaziv and collaborators used, the authors have reproduced the analysis performed by Gaziv upon the same data and have shown that the third PCA for these participants was consistent as the split between the motion between left hand and right hand.

Ultimately, I conclude that a good first approximation to body motion is using the total amount of motion, but the next steps up in approximation quality use the three tracked points (head, left

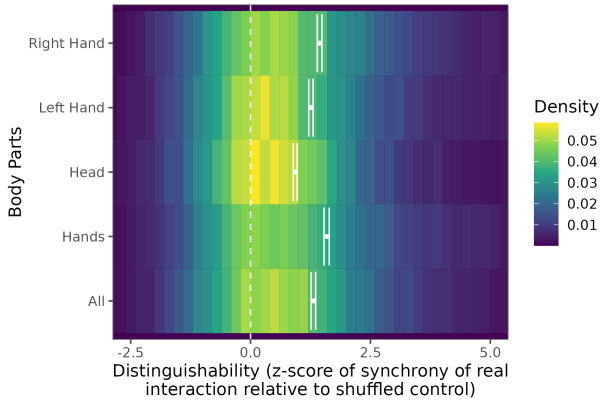

Figure 5.6: Average distinguishability of a sample of measures based upon tracked body part. All show distinguishability; Head is the weakest, Hands is the strongest. Horizontal position indicates distinguishability, vertical position indicates the body part tracked by the measure. Color represents density, indicating the distribution of distinguishability scores for each body part. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals.

hand, right hand) that virtual reality provides.

Like Hale et al. [48] and Sun et al. [110], I investigate synchrony across pairs of body parts (e.g., head and head), but I also investigate the synchrony between different body parts (e.g., head and left hand). The results are displayed in Figure 5.8.

First, it must be noted that all measures do capture synchrony. That is, each of these subsets of measures were able to, on average, distinguish synchronous from pseudosynchronous distributions, one-sided t-test, all $p < 10^{-15}$, corrected with False discovery rate. Second, all measures using the same body part for both participants (along the diagonal) were high. Third, measures that paired one participant's head with the other's hands (together or separately) were significantly weaker. These results indicate that in my setting, there were different processes influencing head and hands motion, though they overlapped somewhat.

5.3.5 Time Scales

The time range upon which people are synchronizing is also of interest. Previous work has indicated trends that longer timescales can have more synchrony, though they largely have not been the direct focus of study. This is in part due to the duration necessary to study this long-term synchrony.

This is also a valuable subject of interest because it touches upon several points of human behavior that tie together different domains of study. Newell's model of human activity [81] consists of several time bands. The smallest is biological, and my data collection rate of 30Hz hits the upper end of this range. I investigate synchrony through the cognitive (0.1s-10s) and rational (minutes to hours), and have collected data over 8 weeks to study the final period, the social (this data are not yet analyzed).

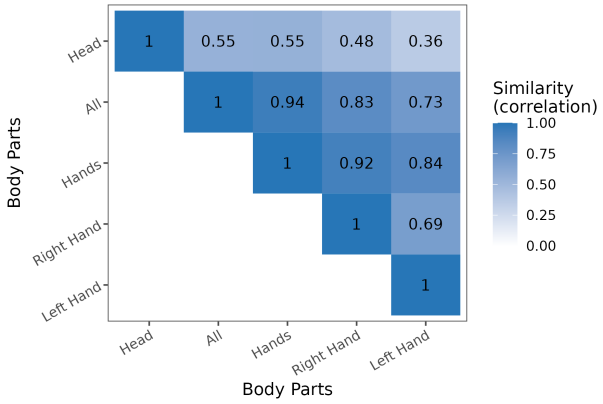

Figure 5.7: Similarity among measures that only vary by body part. Similarity is moderately strong overall, but weaker when compared to head motion. Horizontal and vertical axes indicate which body part is used in the measure. Color indicates similarity, the correlation between measures on the same pairs. Text within each rectangle states similarity to two decimal places. The diagonal represents each measure correlating with itself at r = 1.

In this work, I have found that each window size has some degree of content validity. This is displayed in Figure 5.9. The average distinguishabilities are all different from zero, all $p < 10^{-323}$. By far, the strongest distingushability is the shortest window. I caution this interpretation as the strongest time range, because of artifacts in the tracking data that cause certain frames to be slightly longer than others, affecting everyone present in the same recording equally. Over the course of one third of a second, in which ballistic motion is likely to dominate the signal, this variation can inflate synchrony scores.

In terms of similarity, I find that the difference in window size is related to similiarity such that measures with similar window sizes have similar synchrony values. This is shown in Figure 5.10. I find that all these options specify measures of synchrony that correlate with each other. The weakest is the correlation between 10-frame window and full window, r = 0.0964, t(664) = 2.491, p = 0.0130. This aligns with previous work by Schoenherr who also found small correlations between regular correlation and windowed correlation [103]. The visual pattern of correlation is such that similar window sizes have higher similarity scores. An estimate from this work is increasing the window size by a factor of 10 reduces the similarity by a factor of 0.65 or so.

I also perform a deeper dive into the results of ...

Overall, the results here indicate that while all options are good at distinguishing synchrony, shorter time scale windows appear to be better, and the choice of window influences synchrony. However, it is important to note that window size is *not* the time scale in which synchrony occurs. This discrepancy is further explained in Appendix B and also provides a deeper look at the question of how do people synchronize across time scales.

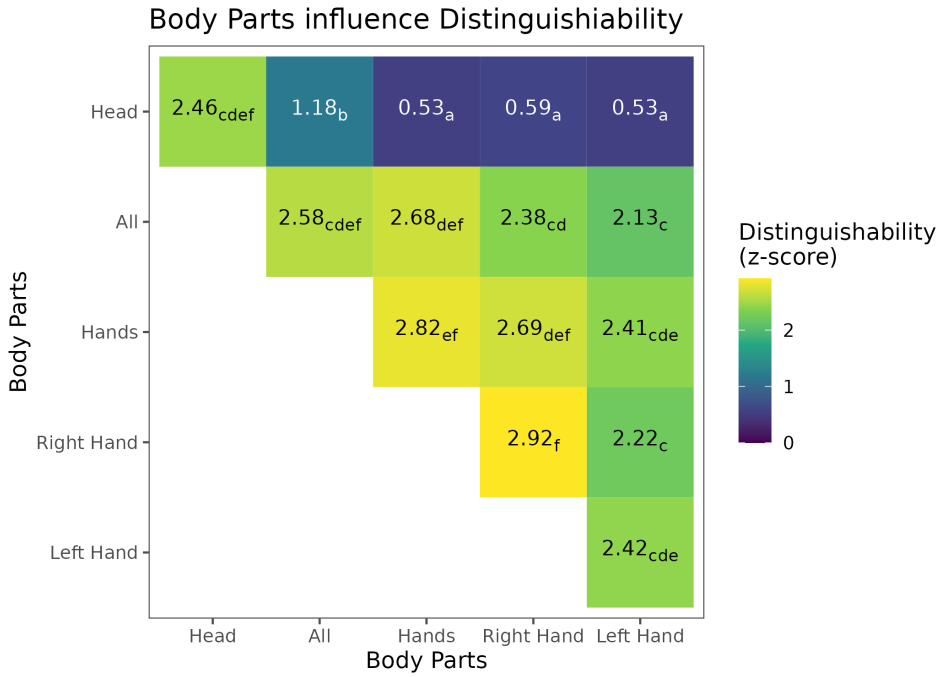

Figure 5.8: Average distinguishability of different measures of synchrony based upon selection of body parts whose motion is tracked, where body parts can vary between each participant in the pair (e.g., A's head syncing with B's hands). Each method of synchron where the measure uses the same body part for both participants has high content validity (i.e., green and yellow are on the diagonal). However, if one participant's head is paired with another participant's hands, distinguishability is much lower. Both horizontal and vertical axes indicate body part, but here they vary by participant. Color and numerals indicate distinguishability. Subscripts on the numerals indicate statistically significant differences such that items that share no letters are different, $p < 0.05$, corrected with False Discovery Rate. Note that in contrast to other figures of this shape, this shows content validity, not consistency.

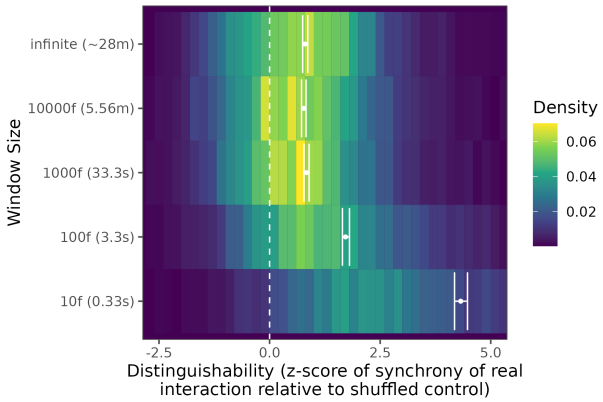

Figure 5.9: Average distinguishability of a sample of measures based upon window size. All show distinguishability; 10f is strongest, full time is weakest. Horizontal position indicates distinguishability, vertical position indicates the window size used by the measure. Color represents density, indicating the distribution of distinguishability scores for each window size. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals. The letter 'f' in the window sizes stands for 'frames', which were collected at approximately 30Hz.

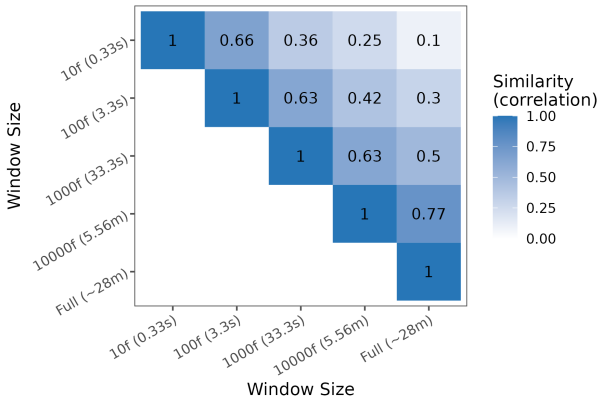

Figure 5.10: Similarity among measures that only vary by window size. Similarity is strong if the window sizes are similar in magnitude, but weak when window size varies by more than 100x. Horizontal and vertical axes indicate which window size is used in the measure. Color indicates similarity, the correlation between measures on the same pairs. Text within each rectangle states similarity to two decimal places. The diagonal represents each measure correlating with itself at r = 1. The letter 'f' in the measures stands for 'frames', which were collected at 30Hz.

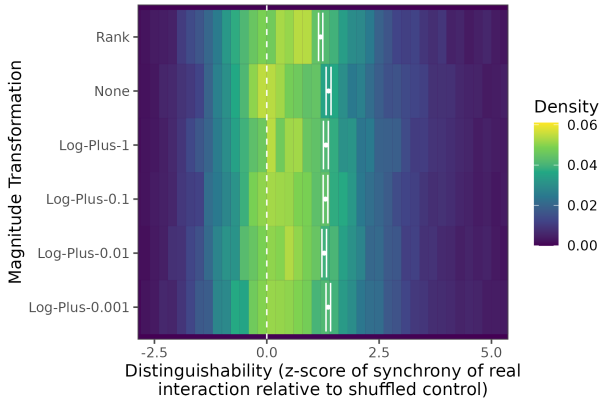

Figure 5.11: Average distinguishability of a sample of measures based upon magnitude transform. All show distinguishability. Horizontal position indicates distinguishability, vertical position indicates the magnitude transform used by the measure. Color represents density, indicating the distribution of distinguishability scores for each magnitude transform. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals.

5.3.6 Perceiving Motion Magnitude

How should a researcher handle the relative importance of large and small motions? On one hand, motion ten times larger may be ten times as important for synchrony. On the other hand, smaller motions may be just as important as larger ones? Once movement is calculated, a choice needs to be in regards for the transformation of its magnitude. Usually, the common usage is Pearson correlation [110] though spearman correlations have been used as well [71]. Speed of points tracked in VR often follows a log distribution [70] so it may be sensible to perform a log-transformation (or log-plus transformation given the possibility of zero motion). I consider all these options when investigating the effect of motion magnitude on content validity and consistency.

As in previous sections, I first examine the ability of the measure to distinguish between real and pseudo-interactions in Figure 5.11. I find that any choice of measure works better than chance, all $p < 10^{-}323$. Variations in distinguishability are rather small, though no transform slightly outperforms rank transform.

There is also the question of which options are most similar to others. With this, I also report similarity scores. I find that measures run on a spectrum from rank through log-plus (small to large p) to no transform. That is, rank transforms are similar to log-transforms with a small positive constant, and no transforms are similar to log-transforms with a large positive constant. Log-plus transforms are similar if their constant is similar. These findings align with the similarity between the different transformation functions.

With high content validity but low consistency, I find evidence that one magnitude transform does not dominate the others but each capture some aspect of how synchronous and pseudo-synchronous interactions are different. Of particular note is that the correlation between rank and no transform

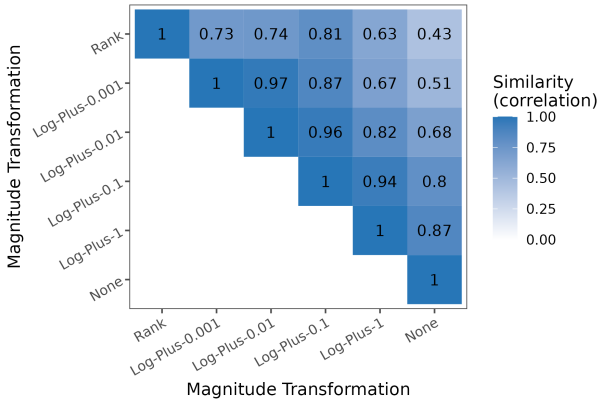

Figure 5.12: Similarity among measures that only vary by magnitude transform. Similarity is strong if the magnitude transforms are similar. Horizontal and vertical axes indicate which magnitude transform is used in the measure. Color indicates similarity, the correlation between measures on the same pairs. Text within each rectangle states similarity to two decimal places. The diagonal represents each measure correlating with itself at r = 1.

(that is, the difference between using pearson and spearman correlation) is 0.43, certainly different enough to be considered two separate constructs in other settings.

5.3.7 Face Validity (RQ4)

There was an activity performed in one of the data collection periods that was designed to be synchronous. The measures of synchrony that register as high during this period have face validity.

One of the data collection periods, period 1, involved an experimental manipulation in which participants first raised and lowered their arms in unison and second pointed towards participants the leader pointed out. This time period was ignored for the other synchrony calculations (as it was selectively applied, being a condition) but is investigated separately here.

Like the content validity analysis, there were a number of measures applied to each pair of participants. This number was larger than in the content validity analysis because there were fewer pairs to work from in order to have a similar level of power - specifically, there were 18 measures applied from the sample of 9300 for each pair of participants that attended the same session. The results are given in Figure 5.13.

Overall, these measures have face validity. Compared to the same analysis performed over the entire time, there are several differences, most of which can be explained by the selection of the synchronous activity. For instance, velocity is not near zero like in subsection 5.3.1 but is notably positive (yet still weaker than speed). This is because one of the synchronous activities involves moving hands up and down, which will have synchronous directions of motion, which is not common otherwise. Furthermore, left and right hand are very similar, but head is notably weaker. This is also because hands were moving in sync but head was not explicitly moved.

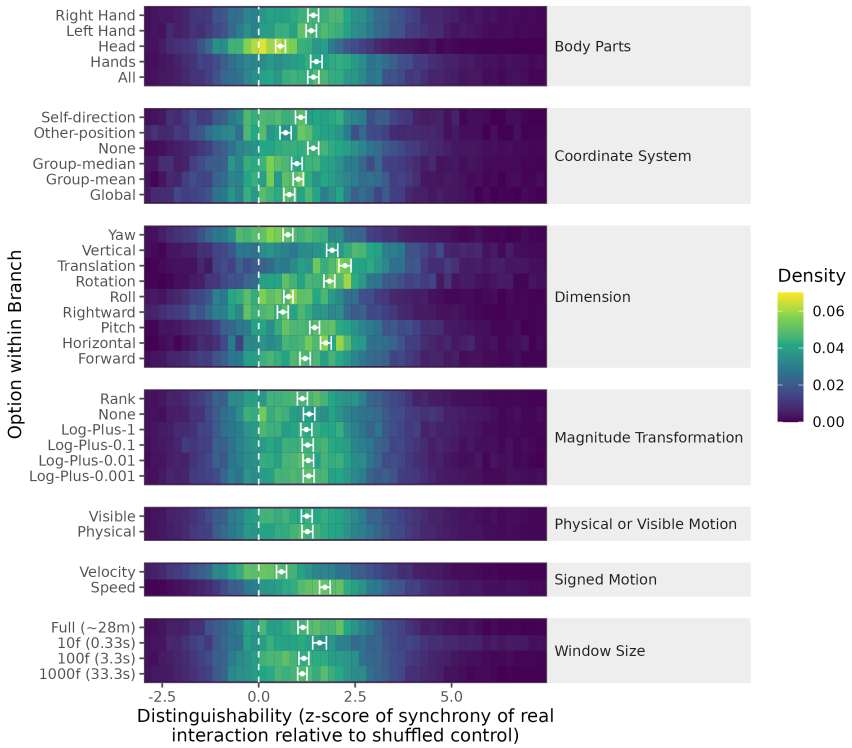

Figure 5.13: Average distinguishability of measures involving a specified option running on time in which participants were intentionally synchronizing. All show some degree of distinguishability; Head in Cody Parts and Velocity in the Signed Motion branch are the weakest. Horizontal position indicates distinguishability, vertical position indicates the options grouped together by a branch. Color represents density, indicating the distribution of distinguishability scores for each option. The white dotted line is where distinguishability is 0 (equal to chance). Error bars indicate 95% confidence intervals. Each combination of universe (measure) and pair sampled for this analysis is represented seven times in this plot, once per branch (facet).

5.3.8 Predictive Validity (RQ5)

Because I have some questionnaire data collected, I also have the ability to test these measures relative to constructs previous related to synchrony. However, my analysis found it difficult to relate any of these measures of synchrony to entitativity, familiarity, and attraction in this dataset.

I performed three types of analyses that vary by the measures selected. In the first, all measures (excluding measures with velocity) were related to the construct of interest. In the second, all measures containing a specific option were selected, as was done for RQ1 and RQ2. For the third, a sample of measures were used, as in RQ3. The full results are available in Appendix C.

In summary, if all analyses are given false-discovery-rate correction, there are no significant results. I do find that 13 of the 30 measures in the dense sample relate liking to synchrony negatively, uncorrected $p < 0.05$, i.e, more familiarity (as reported at the end of the experiment) was related to lower synchrony scores. If false discovery rate correction is applied to that analysis in particular, the number of significant measures reduces to 9 of 30. A specification curve plot indicates that this is true for mid-length synchrony (100 and 1000 frames).

5.4 Discussion

First, I performed an exploratory multiverse analysis on a wide space of synchrony measures. From this analysis, I concluded that in answering RQ1, selecting velocity of a tracked point (as opposed to speed) led to measures sufficiently weak in content validity to be excluded in further analysis, deeming that branch non-equivalent. In regards to RQ2, the branches of coordinate system, dimension, and virtual vs. physical motion showed enough consistency between options to be considered equivalent for the purposes of this analysis, and window size, body parts, and magnitude transformation, with good content validity but low consistency, remained in the analysis as uncertain branches.

The dramatic difference between velocity and speed is surprising given previous work [41], though I judge this difference to be due to the study of spontaneous coordination rather than synchrony induced by directed motion. I leave the possibility open that certain velocity-based measures may in fact measure synchrony. However, I note that the difference in magnitude is substantial: the largest average distinguishability using velocity was 0.601 (limiting to measures with the vertical dimension) and the smallest average distinguishability was 1.11 (limiting to measures with full window size).

It is perhaps surprising that the dimension upon which people synchronized does not seem to matter that much. Considering when the signed motion is speed, one can interpret these results to indicate that synchrony is largely determined by whether participants are moving at the same time, regardless of the direction in which they move. This result echoes why Motion Energy Analysis [45] has been an effective measure of synchrony, despite its drawbacks. Because differences in dimension did not matter much, the question of which coordinate system (for forward and rightward) is not very relevant for the same reason, that ultimately total motion seems to be considered.

The question of virtual vs. physical motion is also interesting. On one hand, one might judge that

synchrony is about perceived movement, and including it may be helpful. On the other hand, one may focus on the motion generation and only include physical motion. However, the interpretation I give to these results is such that both are signals upon which participants may synchronize, or happen to work with at the same time.

The options within the branches of window size, body parts, and magnitude transformation each have content validity, but lack consistency. Most dramatic is the window size, where correlation between my largest window (28min) and smallest window (0.3s) was $r_c = 0.1$, corroborating a similar number by Schoenherr and collaborators [103] in their study of motion using MEA. This is especially concerning for interpreting the literature given that there are substantial portions of research in synchrony for both windowed correlation and full correlation.

I interpret the results regarding the low consistency in body parts to be the result of the type of action one is doing. For example, talking likely moves the hands more than the head, looking moves the head more than the hands, and moving around moves both. This aligns with explanations given in previous work that shows varying effects on synchrony from body parts, such as Yilu and collaborators [110].

Finally, there is magnitude transformation. This branches shows low consistency, with .43 correlation between rank transform aka Spearman correlation and regular Pearson correlation. It is unclear which method provides a better approach, and further work that investigates predictive validity, e.g. matching this to an outcome like client-therapist rapport [93] can help distinguish measures in this branch.

I also performed an analysis indicating the overall consistency of randomly selected measures. I take it to be a fair approximation such that if two measures vary along multiple branches, the correlation between them is about the product of the correlations of the different differences. That is, each branch seems to work independently of the others and the change of options in one branch is not likely to be made up by a change of options in another branch. Overall, I found that the median correlation was $r_c = 0.39$. The synchrony scores are in the same direction, but there is extensive flexibility for a researcher to leverage degrees of freedom in measurement when looking for a significant relationship to another value. Therefore, I strongly encourage researchers of synchrony to preregister their pathway (or multiverse) ahead of time, or transparently report the explored measures after the fact.

While the weak consistency causes concerns for synchrony as it relates to some construct of interest, the high content validity continues to show synchrony happens in interactions, including interactions in virtual reality. According to the results of sampled universes, it is clear that the vast majority of measures can significantly distinguish between synchrony and pseudosynchrony, not just the average ones. All thirty measures were able to distinguish between synchronous and pseudosynchronous interactions.

This result does not mean all measures are equal. I find that in my dataset, measures using a shorter window size showed higher distinguishability, but I do believe this is due to recording timing errors.Hands-hands had particularly high distinguishability in my dataset, because of the tasks being

performed, but any paired body parts seemed to have good distinguishability.

5.4.1 Implications

The largest theoretical question is simply: why do so many variations of this measure work? What is it that causes similarities across so many time scales, body parts, and definitions of motion? One option I propose is that perhaps what is being synchronized upon is the total "motion energy" [45] of the interaction. As the conversation progresses, there are high moments of energy, and low moments of energy, and the ability for participants to be "in-tune" with others indicates some combination of focus on the conversation and ability to communicate.

Another option I propose is that spontaneous coordination is very difficult to distinguish between arbitrary (non-intentional, non-task-based) mimicry or the explanation of a *common response* to a *common stimulus* given *common ground*. For example, two people may both look down at a similar time because they are mimicking each other separate from any intention or appropriateness. Alternatively, they may hear a third person point out something on the table, and because both participants have common ground with each other regarding the task at hand and both receive the common stimulus of their third partner's words, they look down. This would also explain why in previous work in triadic synchrony there was higher synchrony when participants were *not* looking at each other compared to when they were [71].

These proposals are difficult to distinguish when studying spontaneous coordination in natural settings, due to weaker experimental control. The more the scientist influences behavior for another person to mimic, either by task or physical constraints, the less realism in the experimental setting. There are some exceptions to this rule, for example, the use of virtual reality in a transformed social interaction paradigm [9], which I discuss in the future work section.

Both of these explanations also indicate why measures of synchrony have varied so immensely. A wide range of stimuli cause synchrony, and no single measure within the would dominate others at distinguishing synchronous from pseudo-synchronous interactions. The proliferation of synchrony measurements may not in fact be a methodological problem but rather something teasing at the nature of synchrony. I encourage further theoretical work into why so many of these measures are effective at distinguishing between real and pseudo-interaction.

In regards to the measurement of synchrony, I give some practical recommendations. First, there is a sharp difference between synchrony in speed and synchrony in velocity. Most spontaneous work uses speed in regards to spontaneous nonverbal synchrony, so I encourage its use to be continued. Second, selections of body parts should be symmetric, i.e., both participants are tracked in the same way. The selection of which point or points to focus on is dependent on the task, and I suggest selecting the body parts that have the most meaningful movement. Third, selection of magnitude transformations are all acceptable, and there does not appear to be easily visible differences between them. The risk here is inflation of false positives due to unreported exploratory analyses, which can be addressed by preregistration. Fourth, the window size for windowed correlation should be

specified ahead of time. The use of correlation (i.e., a window size of the full recording) is not acceptable unless the time periods are very consistent. However, the most informational approach is to break down synchrony by time scale, if possible (see Appendix B). Finally and most importantly, transparently report exploratory and confirmatory analyses.

To give a very specific recommendation for a synchrony measurement procedure, track all body parts summed together, track total motion in terms of speed, preregister a window size based on previous work and perform an exploratory analysis across frequency, ignore virtual-only motion if relevant and possible, and perform no magnitude transformation upon the motion.

5.4.2 Limitations and Future Work

While one contribution of this work is that it explores many different branches in measurement of synchrony, it does not cover all the variations evidenced in the literature. For example, all these analyses use Pearson correlation as a final step. The addition of ranking before correlating also effectively produces Spearman correlation, but there are far more ways to link two streams of information, as Novotny and collaborators point out [83]. These other methods include wavelet analysis, cross-recurrence quantification analysis, peak picking, and cross-correlation. Each of these methods has its own parameters and thresholds, too, adding in more options and branches to the multiverse. There is also the question of pseudosynchrony measures. In this work, I shifted the full time by a random offset, ie., comparing participant A and participant B (synchrony) to A to B ten minutes later (pseudosynchrony). This preserves the distribution of motion each person has per recording, but can be sensitive to long-scale fluctuations. If there are multiple recordings, as I have, one could also shift by a recording, so A week 1 interacts with B week 2, or person, where A week 1 pseudo-interacts with C week 1. I also applied the same measure to both participants in an interaction (except for the body parts follow-up analysis). It is possible to expand the multiverse by separating the measures of motion so that, for example, A's left hand vertical motion is synchronized with B's head pitch.

There are also several factors that make this dataset unique in the study of synchrony. This provided benefits to new questions, but may make it difficult to extend conclusions back to previous work. This data was collected in group interactions, and pairwise synchrony was calculated. It remains unclear how two-person synchrony relates, numerically and theoretically, to synchrony in larger groups [69]. The task structure, group design activities in VR alongside conversation, may make the synchrony signatures unique to my work. Finally, the whole process was performed in VR, and while there is evidence synchrony occurs in VR [48, 117, 110, 71] and theory explaining why many social psychology effects carry over from real life to VR [133], there is a possibility some of these effects are VR unique.

To further discuss the limitations and future work in regards to groups size, it is important to note that almost all previous work in synchrony has focused on dyads. When groups larger than two are studied, though, the largest span of work is on exceedingly large groups in obviously

synchronous activities like marching. The mid-size group - especially small enough when individual relationships become important - is relatively less-studied. In this context, one of the key features that differentiates mid-size groups is attention. In a dyad, attention would largely be focused on the other in the interaction. In a massive group, is is nearly impossible to focus on one other person. What is the liminal space between these points? Is is more accurate to say that everyone follows a single 'mean' signal in these groups? Do we follow everyone individually? Or do we follow different people at different times? If the last option is likely, there are further questions as to how often that synchronization-attention shifts. If we synchronize, on what scale does *who* we synchronize with change? Is it something that changes more slowly, like identifiability or liking? Or is it something that changes more quickly, like turn-taking or attention? Model comparisons and descriptive analyses of interactions may be able to elicit evidence useful for these questions.

It is also true that the measures of synchrony in this analysis are quite low-level. For example, analysis is performed on a stream of numerical values of head yaw as opposed to some point at which the participant is directing their attention. The former has many potential causes, some of which may align with social behavior and some that may not, but the latter has a more directly interpretable social signal. First, it ought to be mentioned that this is the state of the art in the automated analysis of synchrony [48, 110, 71]. Put simply, there is a lot of context that still needs to be taken into account if a person is truly looking at an object or simply is facing their head (and eyes) in that direction. Second, there is some degree to which this would be captured by the currently captured signals like yaw. For example, if two participants shift their attention from one member of a group to another at a similar time, their synchronized movement will appear in a speed-based measure of head yaw. The bias towards positive synchrony even in this rare case will bias the measure of synchrony positive, even if there are many other causes of head motion. Nevertheless, it is clear that future work can more directly leverage known social signals like head gaze.

The activities performed also vary by week and even by minute within each group. These results collapse across these heterogeneous settings, but some measures could be much better for particular kinds of activities. Given that the face validity of measures shifted given the task of intentional synchrony, it is not unreasonable to consider that speaking may have different statistical signatures of synchrony compared to working with VR objects.

One unique opportunity for future work in VR is to use transformed social interaction [9] to directly manipulate motion in the study of spontaneous coordination. The virtual representation of one participant that is displayed to the other participant can have its motion modified in real time, possibly evidencing spontaneous coordination. This can work through some of the difficulties in the realism-control tradeoff and provide a way to demonstrate spontaneous coordination in realistic settings.

5.5 Conclusion

In this study, I investigated the content validity and consistency of 9300 measures of synchrony produced by a multiverse analysis. I found that most measures of synchrony could distinguish between synchronous and pseudosynchronous interactions. Measures that included velocity distinguished at a far worse rate than those that used speed. From this, I provoke questions as to why so many measures have content validity, and suggest that either global motion is a good approximation to the signal on which people synchronize, or that people synchronize by having the same response to a shared cause, such as both being attuned to the co-constructed tone of a conversation, or performing similar movements in similar tasks when prompted.

I found low consistency between options when selecting window size for windowed correlation, magnitude transformation, and body part. The selection of windowed correlation functions as a high-pass filter, erasing most synchronization that occurs on a scale larger than the window size. Instead, I suggest synchrony should be broken apart by time scale. Body parts synchronized differently, which I judge to be the effect of the task performed.

Future work can extend the domains in which this work has been studied, with other populations, settings, tasks, contexts, and expanding the multiverse as computationally feasible. Using this work and future work, I can better understand synchrony and human interaction broadly.

Chapter 6

Identifiability

6.1 Introduction

Recently, social virtual reality (VR) has been increasing in popularity. If it becomes a mainstay in the consumer space, it will be important to discover, understand, and address the risks associated with its use [125]. Because the full status of all objects in a virtual environment is stored and maintained by a computer, and this status is shared broadly in social virtual reality settings, activity that occurs within a virtual world as part of "the metaverse"[36, 32] is equivalent to all virtual spaces being outfit with security cameras and live microphones are running at all positions at all times. It is clear there is a dramatic risk to privacy in this setting. One type of risk is the risk to privacy in these social spaces through re-identification attacks enabled by the rich nonverbal behavior traces [132] that VR captures, from which behavioral biometrics can be inferred.

Re-identification attacks work by aligning several types of partially identifying information. For example, knowing only one of a target's ZIP code, gender, or date of birth is not likely to identify that target out of the entire United States population. However, these three data points together do identify about 87% of United States residents enumerated in the 1990 census [114]. A similar pattern holds for web browsers given several browser identifiers including version, operating system, language, and timezone [35]. In VR, this re-identification has been demonstrated by leveraging behavioral biometrics [88, 76, 69, 79].

In judging this threat, it is important to understand how long these identifying characteristics of individuals last. Some previous work [75] indicates a reduction of identifiability over the long term (e.g., several months), but this work focused on a very short activity chosen to produce identifying data, and sessions were not taken regularly to compare identifiability across many different moments in time. I have collected a large sample size (232 participants) of a long-duration activity (about 20 minutes per session) over a long timespan (8 weeks), which has enabled several contributions to the state of the art, in order of appearance in the paper:

- a proposal of *body-space coordinates*, a refinement of the feature space so that the two horizontal

dimensions is not based upon an arbitrary global coordinate system but rather relative to a person's "forward" direction (subsection *Feature Engineering* > *Body-space Coordinates*)

- a motivation, selection, and justification of a classification model evaluation metric, *multiclass AUC* [53] that is invariant to the number of classes (i.e., individuals) being identified, producing more effective comparisons across disparate datasets, activities, and participant pool sizes (subsection *Evaluation* > *Multiclass AUC*)
- results indicating short samples taken over several sessions are more identifying than longer samples in fewer sessions (subsection *Identification over time* > *Duration*)
- results indicating the delay between training data and testing data affects identifiability in the range from one to seven weeks (subsection *Identification over time* > *Delay*)
- corroboration with previous work [76] that identifiability is higher within a session than between separate sessions (subsection *Identification*)
- demonstration of an inference of gender and ethnicity from VR pose tracking data, producing small to medium gains in accuracy over baseline models (subsection *Inferred personal attributes*)

6.2 Related Work

First, I give a short description of the risks across VR. Then, I focus on VR pose tracking identifiability, first in works that presume a willingly participating user, and second on works performing re-identification without any influence on behavior. The degree of willingness and awareness gives substantially different design constraints for the eliciting and capturing of identifiable motions.

6.2.1 Risks of Social Virtual Reality

The growth of social VR has come with a commensurate growth in others urging caution. Many papers aim to anticipate the risks of VR while there is time to set norms and design new solutions. Slater and collaborators [107] describe several risks including *identity hacking*, an iteration on catfishing and identity theft where one user poses as another in a virtual environment.

There are risks to the use of single devices as well. Virtual and augmented reality devices collect substantial amounts of user data in order to operate. Modern headsets track motion of several body parts at about 90 times per second, capturing both deliberate and automatic behavior [8]. Biometric data itself is available in several ways. For example, height can be deduced from VR pose tracking data through simply the vertical position of the headset [69] and inter-pupillary distance is available through the device API [56].

Data collection is not limited to the user, though. There are ethical considerations for bystanders who have their privacy compromised by virtual or augmented reality systems. In the case of devices

tracking environments, such as always-on cameras in augmented reality systems, individuals who are not the user often lack data privacy safeguards and the ability to opt-out of data collection [96].

The operation of virtual and augmented reality requires the collection and processing of this pose data in order to render spatially coherent scenes. The transmission of high-fidelity nonverbal behavior - a key affordance of VR over other communication media - also requires this behavioral data to come off the device and be sent to other users in the same shared world.

6.2.2 Identification of willing users

There is a fair amount of work on use of VR pose information as a behavioral biometric. One subset of this work queries what kinds of situations and motions may produce reliably identifiable data. In this work, the research contribution is a setting, action, and architecture that show this reliable identifiablity, often justifying elevated privileges in the use of a device. For example, works use certain identifiable movements such as throwing a ball [62] or one's natural walking pattern [104] that can be captured by the headset and controllers. The ability to utilize these characteristics as a method for verification has been demonstrated by several works [66, 84, 124]. For example, Wang and collaborators [124] leverage identifiable head gestures (nodding, turning, and tilting) in order to develop a biometric verification method. Similarly, Li and collaborators [66] utilize head movement patterns that are formed by a user in response to listening to music for a similar verification task.

What makes this set of work different from mine is that the person being identified wishes to be identified. This difference allows the application designer to ask the user to remember an action (e.g., how to nod one's head in response to music) or to perform a specific activity (e.g., throwing a ball). In the identification problem I pose, this is unrealistic, as I are investigating identifiability given no interaction with the target at all.

However, it is worth highlighting that these works often are framed as a type of continuous authentication (CA) scheme [57], a type of security measure, in which users are authenticated continuously and unobtrusively by a trusted program. Both CA methods and re-identification methods are concerned with the accuracy of a given situation, motion, and model, but the design values of a continuous authentication scheme and re-identification attack are different.

6.2.3 Identification of unaware or unwilling users

The primary question regarding identification is twofold: what makes certain situations more amenable to identification than others, and what about those situations can be remedied?

In this work, I focus specifically on VR pose identification. There are other types of previous work, such as gait recognition [123], that have similar questions and findings. However, one of the strongest differences is that most situations in which VR is used do not allow the user to have enough space to permit walking for more than a few seconds at a time, so findings unique to the activity of walking are not likely to transfer to the VR setting. Based upon the work so far in VR pose identification, there are some threads of work beginning to appear.

First, the number of users to identify between, which I refer to as the *classification size*, affects accuracy, but the relationship between classification size and accuracy is not clear. Three separate works explicitly analyze the relationship of classification size to accuracy [88, 69, 124], and all show effects that larger problems tend to have lower accuracy, even with the same training algorithm and underlying dataset. This is simply due to the number of classes increasing the likelihood of data points of another class sufficiently close to data of the class being queried. However, it is not clear how to compare simple accuracy measures across two studies of difference sizes. For example, is 90% accuracy on 10 participants better or worse than 50% accuracy on 50 participants? Without access to the datasets and the models to make an intermediate comparison, it is impossible to conclude which.

Second, the fit of the featurization to the activity also affects accuracy. Activities that are less unlikely to be organically encountered while spending time in social VR seem to elicit higher identifiability accuracies. For example, training using the trajectories taken when placing blocks in specific places produces 98.6% accuracy [84], standing and watching 360° video produces 95% accuracy [69], throwing a ball at a target produces between 85% and 91% accuracy [75], pointing, grabbing, walking, and typing produce between 48% and 63% accuracy [88], and walking and interacting with virtual objects elicits 37-42% accuracy [76].

While some activities have been known to be identifying, e.g., walking, what is being discussed here is a fit between the featurization and the activity. Moore and collaborators intentionally did not innovate on the featurization in their work in order to make a direct comparison to previous work [69], and the feature set selected in that work was not good for walking as it was primarily on static biometric values. On the other hand, that same feature set was unreasonably effective for 360 videos because participants mostly stood in the same spot or simply looked around, allowing reliable measurements of height and hand controller orientation. One may naively think the most salient and most active measures for the activity (e.g., head direction in terms of pitch and yaw) would be the most identifying, but in fact it is the seemingly ancillary data of height and controller positioning. Though this 'fit' is not clearly defined *a priori*, what is clear is that some tasks are more identifying than others, some tasks are more commonly performed than others, and those are not often found together.

Outside of these three threads on identifiability and its confounds, there are other works to highlight. Nair and collaborators [79] use the identifying characteristics from the VR data collected by a client-type attacker to infer demographic information about each user. Compared to my work, they use a wider array of features, while I focused specifically on pose over time. Olade and collaborators [84] perform whitebox penetration testing of their identification method (i.e., describing the inner workings of their verification method so that it is known to the attacker), and find that intentional imitation of others, even if the attacker can watch the target, is not sufficient to fool their identification process. Falk and collaborators [37] demonstrate identifiability across real-world video and VR-world video by leveraging available skeleton identification pipelines and innocuous social VR activities. Sabra and collaborators [100] leverage identified smartphone motion data to

link identities to users in VR. Finally, R. Miller and collaborators [75] indicate that moving from one VR headset to another can reduce identifiability.

Some work has been done in demonstrating how much more identifiable actions can become if a minor degree of influence in the virtual environment can be permitted. Falk and collaborators demonstrate identifiability on set a of N=5 when an unprivileged attacker (appearing to the target as simply another social VR user) asks the target to perform an innocuous activity like throwing a ball or eliciting an automatic reciprocal response like waving back [37]. This influence is even greater in cases where the attacker can design a gamut of activities for the target to perform [79].

In contrast to the attack space, there is less work about defense mechanisms for this data. M. Miller and collaborators [69] reduce the training data streams from 18DOF (head and hands position and rotation) to 3DOF (head rotation only) and reduce accuracy from 95% to 20% on a set of 511. Moore and collaborators [76] reduce accuracy from 89% to 32% on one set of data and 42% to 13% on a second by switching from position-based to velocity-based feature vectors. Nair, Gonzalo, and Song [80] use differential privacy methods on the biometric features they lay out in [79]. Differential privacy methods incorporate a type of noise to each data point within a dataset so that even when the entirety of the dataset is compromised save for one point, the relative likelihood between the data point being the true value and the data point being any other value is bounded above by the privacy parameter. For a formal mathematical definition, see [34]. There are other types of privacy guarantees, such as k-anonymity and plausible deniability. While to my knowledge these approaches have not been taken on sample-level data, there has been work on protecting eye-tracking data [29].

6.2.4 Identification Over Time

In this work, I focus specifically on identification over time. The delay in time between a user's training data (i.e., enrollment) and a user's testing data (i.e., input) seems to affect accuracy. Works that have a minimal delay between training and testing have a higher accuracy than works with longer delays, e.g., a delay of 30 seconds between sessions and data collected over the span of about an hour with 98% accuracy [124], no delay and a span of 30 minutes with 98.6% accuracy [84], no delay and a span of 10-15 minutes with 95% accuracy [69], no delay and a span of 60 minutes with 89-95% accuracy [76], no delay and a span of 60 minutes with nearly 100% accuracy [102], sessions recorded on "different days" with 90% accuracy [67], "at least three days [between]" with 63% accuracy [88], and one week later with 42% accuracy [76]. This effect of time delay is explicitly studied by R. Miller and collaborators [75] by combining two sets of data collected up to 18 months apart. They find no effect of delay on short-scale separations (within 24 hours) or medium-scale separations (comparing delays shorter than 3 days and longer than 3 days in one analysis, and the same but for 10 days in a second analysis). On long timescales, which in their work goes from 7 to 18 months, there were changes in behavior and a reduction in accuracy, which was not the case in the short- and medium-term delays. However, the delays were not regularly spaced and some varied widely in magnitude. It is still an open question how much and in what situations identifiability

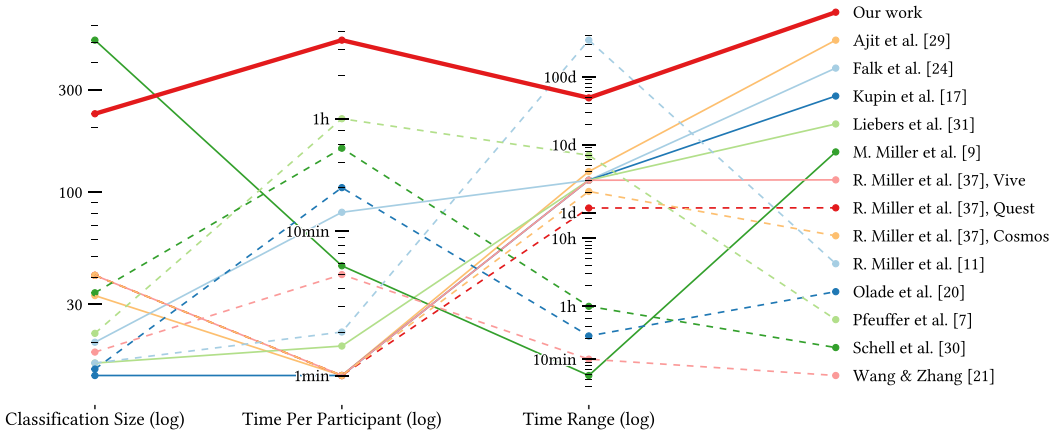

Figure 6.1: Parallel coordinates plot of classification size, span of time in which data was collected, and total duration of data collected per participant (average or estimate if there is variation across participants). The current work is the largest or the second-largest on all dimensions. Note all dimensions are log-scaled in order to better scale the variation. Papers are given alphabetically by the last name of the first author.

changes over time.

In contrast to previous work, I focus on an extremely common social VR activity, group discussion. This way, an attacker need not elicit a response nor manipulate the environment. As shown in Figure 6.1, this work is also on a far larger sample size than most, has more collected data than any other, and collected that data for a longer duration than most. Additionally, there is regular spacing of data collection periods. This dataset is uniquely posed to answer the next round of questions about identifiability.

6.3 Methods

6.3.1 Threat model

It is important to establish the kind of threat under study. In previous work on VR and identifiability, there are two dimensions upon which researchers have categorized threats. First, there is the question of what data is available to the attacker. Nair and collaborators [79] delineate between hardware-level attackers that have access to firmware, client-level attackers that have access to the headset APIs, server-level attackers that have access to the telemetry data sent to the servers and 'unprivileged user' attacker which is another VR system partaking in the same social virtual world. Along this dimension, I focus on the unprivileged user.

The second aspect of space of threat models is the capability of the attacker to influence the behavior of the participant, and the extent to which this can be done. For example, is the attacker designing a virtual world [79], are they another user that is interacting with the user [37], or do they

wish not to interact with the target entirely? In my work, I focus on no interaction at all. This may occur because the attacker is working with previously-collected data, does not want to be vulnerable in the virtual world, or has data collected at scale and cannot interact with each user.

This threat model is selected because it is the least privileged attacker. Therefore, findings based on this work are likely to be applicable to all attacks leveraging VR pose tracking data. It also sets a baseline on threat for all these other conditions. Finally, there are some cases in which this may be the mode of an attacker, e.g., large-scale surveillance where individuals are not queried directly, re-identification attacks where actions are stored for a period of time before being queried, or any other situations in which the attacker does not which to have any direct interaction with the target. Note that this threat model is quite different from the traditional authentication threat model in which a user attempts to gain unauthorized access by posing as another user.

I have also selected the activity, group social discussion, to be representative of activities one may encounter in social VR. This is in contrast to activities like throwing a ball that may not be encountered regularly. This activity also makes the findings more ecologically valid.

6.3.2 Feature Engineering

The feature set I used consisted of 840 *streams* that were subset and summarized in various ways. Some of these streams were defined in terms of body-space coordinates, which is described below.

Body-space Coordinates

The registration of a coordinate system is often not amenable to moving, flexible, and diverse human bodies. Over time, different coordinate systems have been developed for specific purposes, such as the anatomical planes (coronal, sagittal, transverse) for medical terminology. For the purposes of my work specifically and of VR more generally, I propose a coordinate system that synthesizes the global vertical axis with horizontal axes relative to the headset's forward direction.

In this section ssec:proxemicsand next, I designate $\mathbf{p}_\alpha[f]$ a 3×1 vector specifying the position of object α at frame f in global coordinates, where $\alpha = h$ for head, l for left controller, and r for right controller. Each dimension can individually be accessed as $px_\alpha[f]$, $py_\alpha[f]$, and $pz_\alpha[f]$. I also define $\mathbf{R}_\alpha[f]$ as a 3×3 matrix specifying the rotation of object α at frame f. Additionally, I define $yaw(\mathbf{R})$, $pitch(\mathbf{R})$, and $roll(\mathbf{R})$ as functions receiving a rotation $\mathbf{R}$ and returning a single real number indicating the corresponding Tait-Bryan angles using the Unity game engine rotation conventions. All these expressions implicitly refer to one session.

The construction of the body-space coordinate system is illustrated in Figure 6.2 and described as follows: while the vertical (Y) direction remains vertical in this coordinate system, one of the horizontal directions (in my convention, Z) is defined to be forward relative to the participant's body. I operationalize the forward direction as the horizontal direction of the mean of the head's forward vector over the course of a span Δf of frames F centered on a query frame f. In my work, the span is $\Delta f = \pm 3s$. The projection onto the horizontal plane downweights directions facing upward

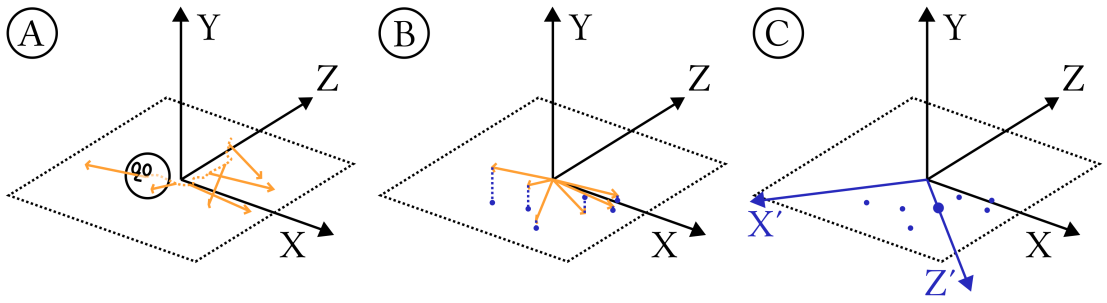

Figure 6.2: Body-space coordinates. In Panel A, a span of one person's motion is shown in space. The orange arrow represents the direction the headset is pointing. In Panel B, all these direction vectors are translated with initial points at the origin. The blue dots are the projection of the orange forward vectors onto the XZ (horizontal) plane. Panel C denotes these blue dots, the large blue dot showing the mean of these points, and the Z' and X' axis denoting forward and rightward, respectively.

or downward, which I judge is appropriate. The rightward direction is then defined as the cross product of upwards with forwards so that it is orthogonal to both. Mathematically speaking, if F is the set of frames over which to calculate a forward vector, then the transformation $\mathbf{R}_{bsc}$ to the body-space coordinate system is calculated as follows:

$$F[f] = \{f' : f - \Delta f \leq f' \leq f + \Delta f\}$$

$$\mathbf{v}[f] = \frac{1}{|F[f]|} \sum_{f' \in F} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{R}_h[f'] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\theta[f] = atan2(v_x[f], v_z[f])$$

$$\mathbf{R}_{bsc}[f] = \begin{bmatrix} \cos(\theta[f]) & 0 & -\sin(\theta[f]) \\ 0 & 1 & 0 \\ \sin(\theta[f]) & 0 & \cos(\theta[f]) \end{bmatrix}$$

The intuition behind the value of body-space coordinates straightforward: often the identifiable features of one's pose are invariant to rotations within the horizontal plane. For example, when one crosses their arms while facing northwest, then ten minutes later does the same pose but facing south, the relative position to each other will be similar relative to the body-space coordinate system, but not relative to the global coordinate system. In regards to the question at hand, this would mean the body-space coordinate system is likely to be more effective at separating one's pose from another's than the global coordinate system would be. To my knowledge, this has not been done in the study of identifiability of VR data, because in previous work there is often a clear 'forward' direction, and that

forward direction can be either enforced or maintained in the course of an experiment [69, 75, 84]. The value of body-space coordinates is further discussion in the subsection on identification features.

Features

In all, there were 42 streams (positions and speeds) each summarized in 20 ways for each sample, leading to a total of 840 features.

The first nine of the 42 streams were kept from my starting model, which was originally used by [69] and re-used in [76]. These were vertical position, roll, and pitch for each of the head and the left and right hand controllers. The subsection on identification features details the value of dropping the other three data streams, specifically, yaw, X position, and Z position. In the notation given above, the first nine streams were

$$py_h[f],\ py_l[f],\ py_r[f],\ pitch(\mathbf{R}_h[f]),\ pitch(\mathbf{R}_l[f]),$$
$$pitch(\mathbf{R}_r[f]),\ roll(\mathbf{R}_h[f]),\ roll(\mathbf{R}_l[f]),\ roll(\mathbf{R}_r[f])$$

The next nine streams are defined in terms of body-space coordinates. First, each point is calculated as the difference between tracked objects α and β by $\mathbf{d}_{\alpha\beta}[f] = \mathbf{p}_\alpha[f] - \mathbf{p}_\beta[f]$ in a method similar to [74]. Then, this difference is transformed to body-space coordinates $\mathbf{b}_{\alpha\beta}[f] = \mathbf{R}_{bsc}[f]\mathbf{d}_{\alpha\beta}[f]$. In this body-space coordinate system, y refers to difference upward, x to difference rightward, and z to difference forward. All three dimensions are used in all three pairs, producing nine streams:

$$bx_hl[f],\ bx_hr[f],\ bx_rl[f],\ by_hl[f],\ by_hr[f],\ by_rl[f],\ bz_hl[f],\ bz_hr[f],\ bz_rl[f]$$

The third set is also nine streams, but they are defined on speed rather than position as before. Specifically, one-frame changes in position are denoted $\mathbf{v}_\alpha[f] = \mathbf{p}_\alpha[f] - \mathbf{p}_\alpha[f-1]$. Three types of speeds are derived, one that indicates total motion $|\mathbf{v}_\alpha[f]|$, one that indicates horizontal motion $|\mathbf{v}_\alpha[f]|_H = \sqrt{vx_\alpha[f]^2 + vz_\alpha[f]^2}$, and one that indicates vertical motion $|\mathbf{v}_\alpha[f]|_V = |vy_\alpha[f]|$. This results in nine streams, specifically:

$$|\mathbf{v}_h[f]|,\ |\mathbf{v}_l[f]|,\ |\mathbf{v}_r[f]|,\ |\mathbf{v}_h[f]|_H,\ |\mathbf{v}_l[f]|_H,\ |\mathbf{v}_r[f]_H,\ |\mathbf{v}_h[f]|_V,\ |\mathbf{v}_l[f]|_V,\ |\mathbf{v}_r[f]|_V$$

Finally, on the speed of the body-space difference vectors $\mathbf{v}_{\alpha\beta}[f] = \mathbf{b}_{\alpha\beta}[f] - \mathbf{b}_{\alpha\beta}[f-1]$, one can compute the same total, horizontal, and vertical motions, but it is meaningful to also compute the difference along both the forward $|\mathbf{v}_{\alpha\beta}[f]|_F = |vz_{\alpha\beta}[f]|$ and rightward $|\mathbf{v}_{\alpha\beta}[f]|_R = |vx_{\alpha\beta}[f]|$ directions. Therefore, there are a total of 15 streams of this type:

$$|\mathbf{v}_{hl}[f]|,\ |\mathbf{v}_{hr}[f]|,\ |\mathbf{v}_{lr}[f]|,\ |\mathbf{v}_{hl}[f]|_H,\ |\mathbf{v}_{hr}[f]|_H,\ |\mathbf{v}_{lr}[f]|_H,\ |\mathbf{v}_{hl}[f]|_V,\ |\mathbf{v}_{hr}[f]|_V,\ |\mathbf{v}_{lr}[f]|_V,$$
$$|\mathbf{v}_{hl}[f]|_F,\ |\mathbf{v}_{hr}[f]|_F,\ |\mathbf{v}_{lr}[f]|_F,\ |\mathbf{v}_{hl}[f]|_R,\ |\mathbf{v}_{hr}[f]|_R,\ |\mathbf{v}_{lr}[f]|_R$$

The featurization for any given session is a collection of 840-entry vectors computed for a set of frames regularly spaced at one second intervals. The 840 features are computed by selecting each one of the 42 streams and computing each of five different summary statistics (mean, median, maximum, minimum, standard deviation) of each of four varying window sizes (1s, 3s, 10s, 30s) centered on the selected frame. On the edges, where a full window frame was not available, the frame was not used.

6.3.3 Model

The model's task is to identify a user based upon their motion. More formally, the model is given a feature vector of motion as described in the featurization subsection, consisting of summaries of several window sizes up to 30s of data. With that vector, the model ought to predict the participant who generated that motion, represented as a value of a categorical variable. In the smallest analysis I perform (1 minute of training data) there is 60 samples per class, but in most analyses there are 500-1000 samples per class (i.e., per user).

The model I have selected is random forest implemented in R [91], version 4.1.2 (https://www.r-project.org/) with the package `ranger`[1], [130], version 0.14.1. The choice of random forest was to balance simplicity with expressiveness and was used in previous work [69, 76]. Most settings for the creation of the random forest were the defaults, in particular, no limit to node depth, a minimum node size of one, number of variables to try as the square root of total variables. The one custom parameter was training only 30 trees on a sample of 100,000 entries from the entire database, with this resample-and-grow process repeated 20 times.

The predictions were made per session by taking the entire session of pose tracking data, computing features, and then aggregating the votes across all 600 trees across all samples. I interpreted this distribution as a probability estimation for the classification of the session as a whole, in line with previous work [69] and consistent with other uses of random forests [82].

The problem type is classification rather than a ranking problem, along the lines of similar work [69, 76, 84]. While this is a simple setting for the problem and has been effective for related work, it may bias towards increasing identifiability of nearly-identifiable participants rather than all participants in general, especially as the classification size increases. This method can be contrasted with multiclass AUC, which increases with any improvement in identifiability, not simply when a sample is correctly classified against all other samples.

6.3.4 Evaluation

I provide several evaluation metrics for these models. my primary metric is multiclass AUC [53] which addresses the dependency of accuracy as the number of classes in the classification problem varies. For the sake of interpretability, I include accuracy, and for comparison and synthesis with previous work, I use accuracy limited to a N-class testing set.

[1] https://cran.r-project.org/web/packages/ranger/index.html

The split into data was simply train-test with several Monte-Carlo cross-validations, as is used in several previous works [69, 76, 72]. The test here performed the role of validating the random forest hyperparameters.

Multiclass AUC

To my knowledge, no work in the space of user identification with VR data has used multiclass AUC. Because it addresses the effect of classification size on accuracy, I give a short description and justification of its use in enabling future comparisons across studies with varying numbers of classes.

Identification-focused works [88, 69, 76, 75] almost exclusively use accuracy for the model's evaluation metric. The benefits of accuracy as a metric include its ease of interpretation and its directness to the question at hand - a less accuracy model is obviously less identifiable, and vice versa. However, accuracy does vary significantly as the number of classes varies, even for the same data distributions and identification processes, as evidenced by multiple works [88, 69, 124]. Intuitively, this is true - it is easier to guess who is walking up the stairs in an apartment with two other people than a house of ten. This effect of the number of classes on accuracy can make synthesis of findings across works difficult, as the classification can vary as much as two orders of magnitude (e.g., 5 in [124] to 511 in [69]).

My criteria for an evaluation metric that addresses this issue is that it produces the same value regardless if it is computed upon the full set of classes, or computed as the average of randomly chosen subsets of classes of any size. A formal mathematical description of this problem is given in Appendix E.

To solve this problem and enable comparisons across analyses with varying numbers of classes, I choose my primary evaluation metric to be *multiclass AUC*, defined by Hand and Till [53]. Multiclass AUC can be described as the average of the pairwise separability between all pairs of classes. More specifically, for each ordered pair of classes, several pairs of instances are produced that contain one instance from each of the two classes. Given the multiclass prediction that provides probabilities to all potential classes, each pair of instances are compared by their probabilities of being a member of the first class. If the instance of the first class has a greater probability of being a member of the first class than the instance of the second class, then a value of 1 is returned, otherwise a value of 0 is returned. Multiclass AUC is the average of this binary variable across all pairs of instances across all pairs of classes. Note that the asymmetry of only comparing upon the first class becomes symmetric because pair selections are ordered.

Put another way, it is the chance that, after randomly selecting a true class, and a fooling class, and an instance from each of those classes, that the member of the true class will have a stronger association (as estimated by the model) than the member of the fooling class does. This process has the benefit of having a baseline of 50% regardless of class size. Furthermore, unlike a binary classification between "matching" or "non-matching" samples, this method uses all pairwise classification information and does not require a manual selection for the relative cost of false negative to false positive to avoid unreasonably high baseline scores (e.g., always predicting "no match" in

an unweighted setting would lead to an accuracy of $\frac{N-1}{N}$).

Hand and Till note that this metric weights the separability of each pair of classes equally regardless of the number of samples in the classes, which may not be appropriate if priors are to be taken into account. Additionally, this is not an estimate of the accuracy attained by the same training process upon a smaller data set constructed in the same class-reduction process, but is instead an estimate based upon the model after training.

Accuracy limited to an N-class testing set

While multiclass AUC is a good multiclass evaluation metric for future work, there are no works in this space that currently use it. In order to allow comparisons to be drawn from this work to previous work, I define accuracy limited to N-classes. This metric may be narrated as a prediction task in which there is a model and a set of N potential classifications, a subset of all the classifications the model could make. First, the model proposes its classification, and if the classification is outside this subset, the model is asked to provide its next best classification. This process only ends when the model gives a predicted classification within the set of potential classifications. A formal description of this process and metric is given in Appendix F. This accuracy is then comparable to accuracy of a similar size identification set.

6.4 Results

There are several analyses performed. First, I estimate the identifiability of participants in this dataset using both within- and between-session methods. From there, I study the effect of time on identification, first through the duration of the training data, and second through the delay between training and testing. Finally, I report identifiability by feature set and preliminary results on inference of individual characteristics such as gender.

6.4.1 Identification

The simplest and most prominent question is identification. As previous work has indicated that the delay between training and testing data sets can influence identification [75, 76], I report two separate analyses within this section: the first is a *between*-session split, and the second is a *within*-session split. The *between*-session split makes the train/test split at the scale of weeks such that the first six weeks of data are the training set and the final two weeks are the testing set. The second analysis within this section, the *within*-session split, splits training and testing segments within each session, such that 80% of each session is used for training data, and 20% is used for testing data, except for one minute as a buffer between segments. Results are given in Table 6.1 that show each accuracy metric for the two splits across the dataset 1, dataset 2, or the combined dataset.

The most dramatic difference for all three evaluation metrics is whether the split is within-session or between-session. This corroborates previous work [76, 75] that the delay between training and

Table 6.1: Evaluation of identification models by train-test split and dataset.

Split	Dataset	Accuracy	Multiclass AUC	30-Class Accuracy
Between	Combined (C=232)	31.68%	85.48%	51.71%
	Dataset 1 (C=86)	45.19%	86.86%	55.34%
	Dataset 2 (C=146)	32.39%	86.37%	50.02%
Within	Combined (C=232)	67.15%	98.43%	85.92%
	Dataset 1 (C=86)	82.30%	98.60%	89.29%
	Dataset 2 (C=146)	69.36%	98.28%	84.46%

testing influences identifiability. Second, in both cases the accuracy for dataset 1 is substantially larger. While multiclass AUC also has slight increases in conditions with dataset 1, it is much less pronounced, both in an absolute scale and in logit units. This result highlights the importance of accounting for classification size. However, it should be noted that identification size is not the only determinant of accuracy. There is a much smaller difference in accuracy between the dataset 2 and the combined data (+60% size difference) than there is between dataset 1 and dataset 2 (+70% size difference).

With this setting being a classroom field study, there are other variables that may influence the degree of identifiability of participants, such as group size or headset used. The group size varied across sessions and dataset. However, in a logistic mixed effects model of group size predicting correct identification with random intercepts for dataset, section, and participant, the effect of group size on identification was not significant according to a Type II Wald test ($\chi^2(1) = 1.023$, $p = 0.312$). Another variable that may influence identifiability is differing VR systems, and there is previous work [72] indicating this to be the case. In my analysis, due to only 2 of 232 participants using headsets that were not the Meta Quest 2, I did not find an effect of headset on identifiability. In a logistic mixed effects model of headset type predicting correct identification with random intercepts for dataset, section, and participant, the effect of headset type was not significant, Type II Wald test ($\chi^2(1) = 2.9743$, $p = 0.085$). I conclude the effect of headset on identifiability in this dataset is negligible and include all participants' data in estimating identifiability.

6.4.2 Identification over time

Motivated by R. Miller et al. [75] and the finding in the subsection above, I investigated the effect of duration and delay on identification. To study duration, I vary the number of separate sessions in the training set and the training time per session to investigate its effect on multiclass AUC for both within and between session splits. To study delay, I vary the weeks upon which a model is trained and tested while keeping the duration the same.

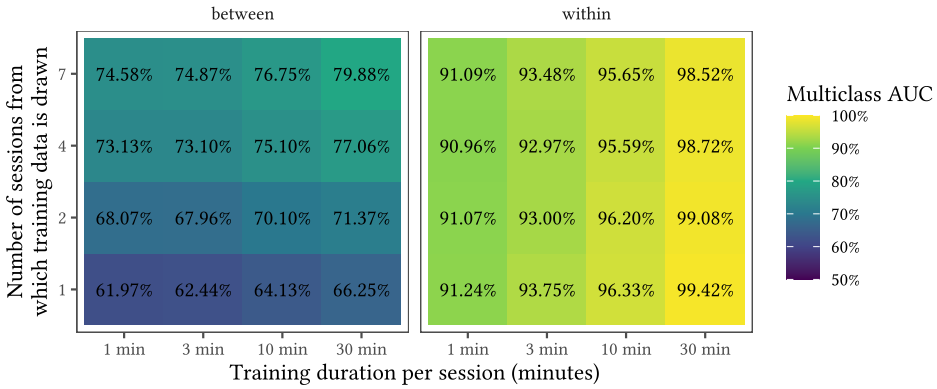

Figure 6.3: Number of sessions and duration of each session affect identifiability, operationalized as multiclass AUC. Two panels shown side-to-side indicate whether the comparison is drawn between sessions or within the same session. The number of sessions is the y-axis, and the training duration per session is the x-axis. The panels are colored indicating identifiability, with yellow as a higher accuracy. Figure produced with ggplot2[127], version 3.3.6.

Duration

In the first analysis, the train-test split was performed by first randomly selecting a set of training sessions of size at most 1, 2, 4, or 7 for each participant, presuming at least one session remains for testing. For example, in the case seven sessions were requested but a participant only took part in six, five of those six were used for training and one was held out for between-sessions testing. Of the selected training sessions, spans of time for training and within-sessions testing were chosen. First, a five-minute span was marked out for testing, and one-minute spans adjacent to the testing span were marked off as buffers. Then, if there was enough space remaining on either side of the testing span, the training span was one continuous block. Otherwise, the training span was two separate blocks that totaled the requesting time. In the case there was not enough time in a session, the entire length of time other than testing and buffer was used for training. For example, if there was a 32-minute recording with 30 minutes requested, there would be 25 minutes for training, five minutes for testing, and one minute on each side of the testing data as a buffer. This means the average training span for each of the 1, 3, 10, and 30 minute conditions were durations of 1:00, 2:59, 9:39, and 22:32 respectively. Sessions shorter than eight minutes total were dropped from this analysis.

The reported result is the average multiclass AUC across 10 Monte Carlo resamplings. For the sake of training time, I reduced the number of trees to 5 and the number of sub-samples to 3. With this change in the training process, the reduction in the training set, and the reduction in the test set to five minutes, I expect to account for the differences in multiclass AUC between the 30 minute, 7 session (top right corner, between panel) training and 5 minute testing (i.e, the results in Figure 6.3) compared to the full duration, 6 session training set with a full session testing set (i.e., the results in table 6.1). Results are given in Figure 6.3.

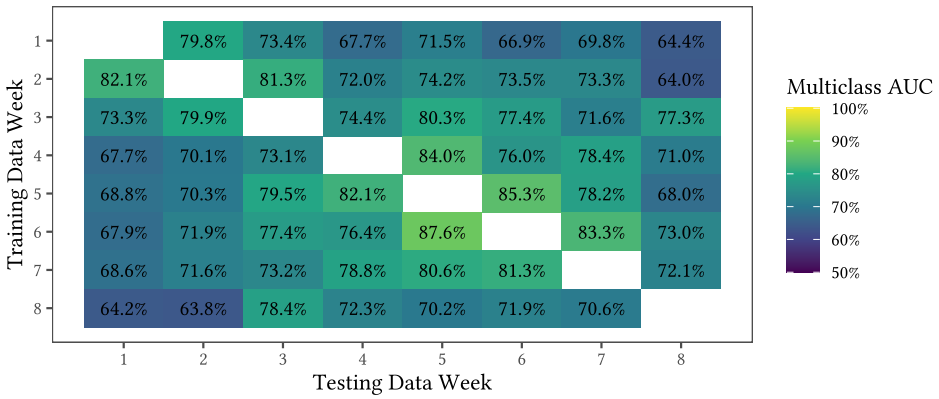

Figure 6.4: Separating the training and testing sets by larger time reduces accuracy. The x-axis and y-axis are the testing and training weeks, respectively. The panels are colored indicating identifiability (operationalized as multiclass AUC), with yellow as a higher accuracy. Note a trend that higher multiclass AUC is along the diagonal (i.e., minimal delay). Figure produced with ggplot2[127], version 3.3.6.

Again, the most dramatic difference in multiclass AUC is in the delay (between vs. within) distinction, even so far that training on simply one minute from the same session as the testing set provides a better multiclass AUC than thirty minutes on all other sessions available to the model. Looking specifically at the between-sessions case, both more sessions and more training data per session produces higher multiclass AUC scores. Of the two, the number of sessions more dramatically affects the multiclass AUC, considering the ranges chosen in this study. In particular, training on one minute from four sessions produces a higher multiclass AUC than training on thirty minutes from only two sessions ($t(15.06) = 3.194$, $p = 0.006$, 95% CI = [0.5%, 2.9%]). In the within-session data, more time is beneficial to multiclass AUC, but more sessions very mildly reduces multiclass AUC. This may be because the variation between sessions does not aid the model but rather misleads it.

Delay

The second analysis keeps the training and testing durations the same but varies the delay between these moments. The multiclass AUCs reported in Figure 6.4 are produced by training a random forest upon one week's worth of data and testing it on a different week's worth of data. In total, there are $8 \times 7 = 56$ entries. The random forests had 30 trees like in the original analysis but only had 3 separate 100,000-sample draws due to the smaller sample set (one week, about 350,000 samples). All data for the selected training session is used, and all testing data matching a participant in the training set is tested with. Note that multiclass AUC is reported both for pairs where training week happens before testing week, as would be expected for an attacker, but also in pairs where testing week happens before training week, which is relevant to pose re-identification well after data collection.

The results in Figure 6.4 show a pattern that multiclass AUC is higher when training and testing sessions have less delay (i.e, at the near-diagonals) than when there is more delay (near bottom left and top right). This effect varies somewhat across weeks, but is still easily visible.

To confirm this effect, a mixed-effect model was fit using the software package "lmerTest" to the 56 data points of multiclass AUC shown in Figure 6.4. This model fit multiclass AUC based upon delay (the difference in number of weeks between training and testing) with random intercepts for training and testing weeks. The effect of delay upon multiclass AUC is highly significant ($t(24.53) = -5.59$, $p = 8.6 \times 10^{-6}$), with an intercept of 81.1% and a slope of -2.3 percentage points per additional week offset, predicting a prototypical one-week delay to have a multiclass AUC of 78.8% and a prototypical seven-week delay to have a multiclass AUC of 65.1%.

6.4.3 Identification features

The features in this model were developed relative to [69] but have significant changes. In order to shed light on the relative value of these change, I provide both a ranking of these features as well as a series of intermediate models and the relative changes in identifiability for each model, allowing the reader to differentiate the contributions of each change to the feature set.

The ranking of features is studied within the setting where 30 minutes of training data are collected from each of the 7 sessions. Then, importance itself is defined using the 'permutation' option from `ranger`, which tests the training data after permuting one of the features, comparing to accuracy on un-permuted data, and returning the reduction or increase in accuracy.

Considering the features in Figure 6.5, what is most visible in the distribution of feature types is that the first section of features, from rank 1 to approximately rank 100, are disproportionately vertical position (YPosition) features. It is not entirely limited to height, as other features like pitch and roll are also present, but few speed-based features are used. Furthermore, most of the top features are with both head and vertical position, as evidenced in Table 6.2 in more detail.

Following that, from ranks about 100 to about 300, there are more pitch and roll values. After that, the other measures are sorted by duration, decreasing in length form 30s down to 1s. While these are all features with low importance, longer-range features appear more likely to be effective.

Finally, there are some clusters of features at the weak end. It appears the combination of speed and minimum is quite weak, which I infer to be reasonable, as they are very likely not stable across participants.

Focusing in specifically on the top 30 features as detailed in Table 6.2, one can see just how much vertical position dominates the feature set. The most important feature that doesn't use height is median head pitch, at 27th, and the only one in the top 15 that doesn't use head is 5th, the combination of left hand and vertical position. Notably, all except one of these features use head and left hand. I judge this to be due to most participants in this set being right-handed. Being right handed means that the motion from the right hand is more due to task and content and less due to idiosyncrasies in motion. Additionally, the featurization biases towards static features and does not

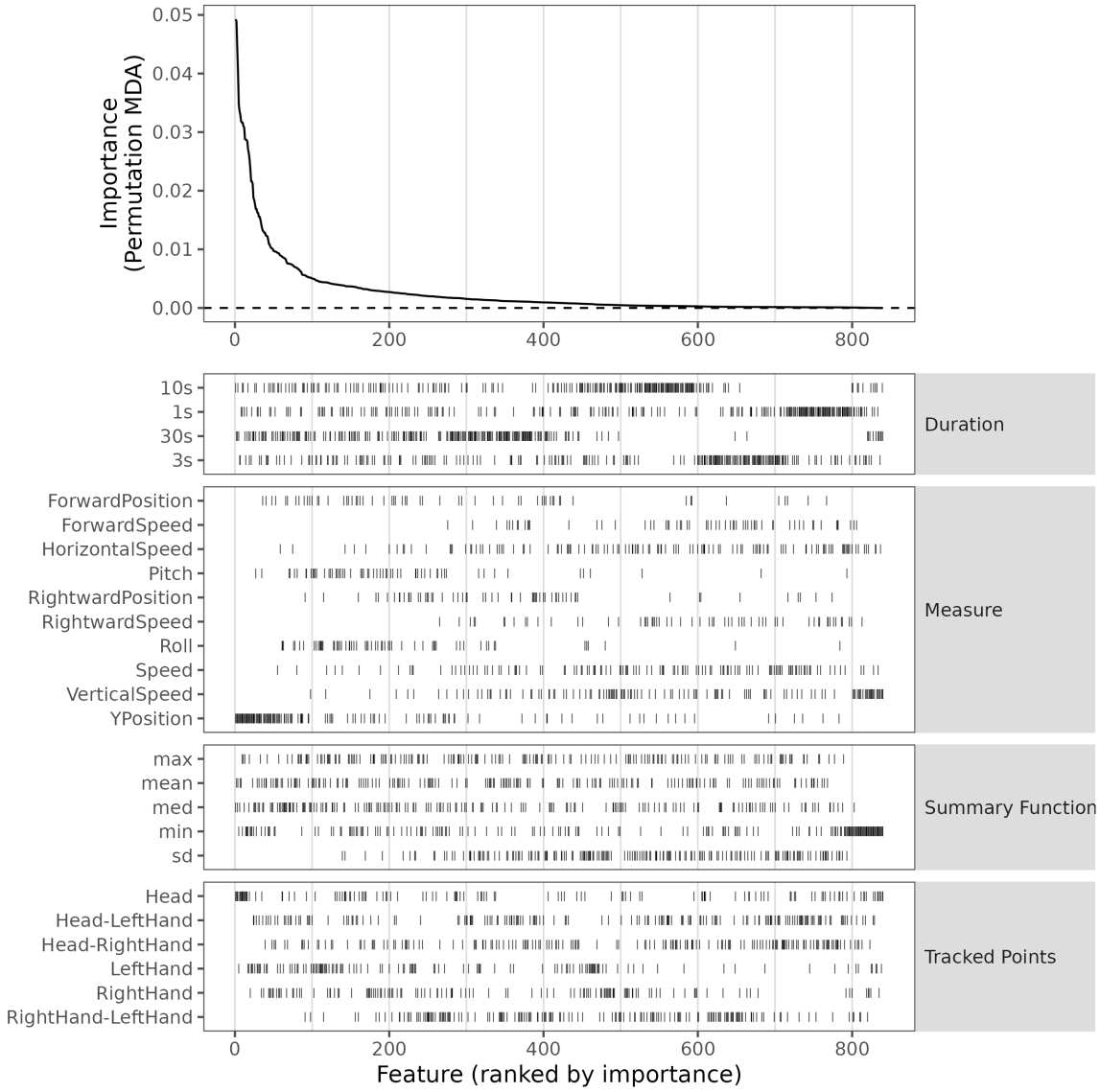

Figure 6.5: Summary view of relative importance of 840 features used in the 7-session, 30-minute, between-sessions model. The horizontal axis represents each feature, which is ranked in order of importance. The top panel indicates the importance score, which is produced by permuting all values of that variable in the training set and calculating the mean decrease in accuracy (MDA). The units are change in accuracy (e.g., 0.04 meant a drop of 4 percentage points of accuracy when permuting the feature). The remaining plots show the specifications, one per facet, of the feature according to duration, measure, and summary function as described in subsection 6.3.2.

adequately represent idiosyncrasies, if they exist at all.

Table 6.2: Top 30 features by feature importance

	Tracked Points	Measure	Summary Function	Duration	Importance (Permutation MDA)
1	Head	YPosition	med	10s	0.0493
2	Head	YPosition	mean	30s	0.0489
3	Head	YPosition	med	30s	0.0443
4	Head	YPosition	mean	10s	0.0402
5	LeftHand	YPosition	min	30s	0.0347
6	Head	YPosition	med	3s	0.0336
7	Head	YPosition	mean	3s	0.033
8	Head	YPosition	mean	1s	0.0319
9	Head	YPosition	max	1s	0.0316
10	Head	YPosition	min	10s	0.0316
11	Head	YPosition	max	10s	0.031
12	Head	YPosition	med	1s	0.0308
13	Head	YPosition	max	30s	0.0288
14	Head	YPosition	min	3s	0.0287
15	Head	YPosition	min	1s	0.0287
16	Head	YPosition	min	30s	0.0285
17	LeftHand	YPosition	min	10s	0.0271
18	LeftHand	YPosition	med	30s	0.0264
19	Head	YPosition	max	3s	0.0253
20	RightHand	YPosition	min	30s	0.0236
21	LeftHand	YPosition	min	3s	0.0216
22	LeftHand	YPosition	min	1s	0.0216
23	LeftHand	YPosition	mean	30s	0.0211
24	Head-LeftHand	YPosition	min	30s	0.0188
25	Head-LeftHand	YPosition	med	30s	0.0184
26	LeftHand	YPosition	med	10s	0.0177
27	Head	Pitch	med	30s	0.0169
28	Head-LeftHand	YPosition	min	10s	0.0169
29	LeftHand	YPosition	mean	1s	0.0163
30	LeftHand	YPosition	med	1s	0.0162

Regarding the comparison of models, I provide stepwise comparisons (e.g., M1 to M2, M2 to M3) using McNemar's test based on either binary variables of correct/incorrect for accuracy, or binary variables of correct/incorrect for each pairwise comparison for multiclass AUC. These are given in table 6.3.

Model M1 used the same feature set as used in [69]. Note that this is the lowest accuracy of all six models here. Models M2 and M3 use a subset of features available to M1, specifically, dropping the two horizontal positions and then dropping horizontal orientation. These changes improve each evaluation metric substantially. Note that the samples within the same session are not independent, and in fact the samples highly correlated upon these features in particular. Domain knowledge in leading these discussions and activities indicates that the horizontal locations of a participant from

Table 6.3: Model comparisons, breaking down changes in the feature selection and their impacts on three measures of identifiability. C = classification size. Asterisks indicate significant change in measure relative to the model in the row above as indicated by McNemar's test, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. C=30 Accuracy is provided for reference and is not tested.

Model	Features	Accuracy (C=232)	Multiclass AUC	C=30 Accuracy
M1	Miller et al. [69]	7.76%	75.18%	21.20%
M2	M1 without horizontal positions	***25.29%	***81.91%	43.10%
M3	M2 without yaw	27.87%	***82.73%	47.51%
M4	M3 with windows of 3s, 10s, 30s	**31.61%	*82.71%	49.22%
M5	M4 with body-space displacements	32.76%	***84.43%	51.54%
M6	M5 with speed values	32.47%	***85.59%	52.59%

session to session varies almost arbitrarily. Disallowing these values allows other features to be considered more strongly.

Model M4 quadruples the feature space by adding three new window sizes. As the original work [69] used only thirty-second clips, while my work has on average about thirty minutes, I can increase my window size without suffering the same loss in sample size. The difference between M3 and M4 is puzzling, as accuracy goes up substantially but multiclass AUC goes slightly down. I hypothesize this is due to the larger feature space adding variance to predictions across the board, but the larger time window providing an important distinguishing factor when there are differences between just a handful of plausible values.

M5 and M6 apply features from previous work. M5 adds features computed from the body-coordinate displacement vectors $\mathbf{b}_{\alpha\beta}[f]$ described in the subsection on body-space coordinates, building on [74]. This reintroduces the horizontal information lost in M2 and M3, but this featurization is invariant to horizontal translations and rotations of the participant. M6 adds velocity features for each stream, inspired by [76] that shows velocity is still identifiable, though less so than position.

In summary, this work aligns with previous work [69] finding that static features like height provide the bulk of identifying power. However, the use of body-space coordinates increases identification both by removing poorly parameterized data and by reintroducing a better parameterization. Additionally, I corroborate the value of other featurizations on the current dataset.

6.4.4 Inferred personal attributes

In this work, I also have a large sample of participants such that inference of individual characteristics is plausible. I perform this analysis to provide some empirical grounding of the potential for VR to be used to infer personal attributes such as gender and ethnicity. I do this work in the judgement that its release will encourage the development of defenses or large-scale more than make bad actors aware of its potential. Biometrics has a history of being used for drawing racial lines as well as giving those lines an objective, scientific backing[21]. I do this work not to continue to essentialize these

Table 6.4: Statistics on inferred personal attributes. Note that multiclass AUC reduces to AUC when classification size is 2.

Inference	Baseline	Accuracy	Multiclass AUC
Gender (C=2)	58.60%	80.26%	71.17%
Ethnicity (C=10)	35.98%	40.01%	63.10%

contingent and constructed characteristics of race, but to demonstrate their effectiveness in inferring an already-essentialized concept of race, and therefore being an opportunity for discrimination and abuse.

To estimate the ability of inference of these characteristics from pose tracking data, I produced models predicting gender and ethnicity from this data. In each setting, I ran 10 Monte Carlo splits of participants such that 80% of participants were in the training set and 20% were in the testing set. Note that no participant's data was in both the training and testing set. Here, each participant's data across all the eight weeks was collapsed into a single prediction. The baseline accuracy reported is from a naive classifier that simply predicts the most common entry within the training set regardless of the testing data. These results are given in table 6.4.

There is indeed gains in accuracy over the baseline model. While the gains in accuracy are mild and mainly rely upon height, they still demonstrate that it is possible to infer personal attributes from VR pose tracking data alone. This deepens the privacy risks to social VR.

6.5 Discussion

6.5.1 Summary of Results

First, I show, in accord with other work [75, 76], that the largest difference in identifiability is whether the enrollment (training) and input (testing) sessions are delayed or adjacent. This work has demonstrated this fact on the largest dataset to date in terms of total recorded time. This finding is echoed in the timing analysis, in which one minute from the same session is more identifying than seven sessions of thirty minutes each. Overall, I infer that there are several identifying variables at play, and some may work on short time scales and some work on long time scales. Future work ought not to look at one time scale but many.

Second, in response to previous work with varying identification sizes, I select and justify the Multiclass AUC evaluation metric to evaluate identifiability across sample sizes. A review of previous work indicated a trend that simple accuracy-based measures of identifiability show lower accuracy with larger sets of users to identify from. Removing this variation can let future work clarify other important trends in accuracy, such as time, feature selection, or activity.

Third, I investigate identifiability as participants have discussions in VR. This activity is a very common activity that occurs in social VR even without prompting, and it is far more available to attackers because of its ubiquity. This obviates the need to have targets perform an action

based on trust of the identification method (e.g., [62, 66] or co-present social engineering in the VR environment (e.g., [37]).

Finally, to better understand the risk of VR pose tracking data, I develop body-space coordinates to allow for the usage of horizontal position data. Both removing the reference to global coordinates and introducing the reference to body-space coordinates increase identifiability. This work takes a step forward in the arms race between identification and de-identification.

This work has relatively low accuracy in raw terms compared to related work [69, 88, 124, 84, 76], but as is shown by the section on multiclass AUC, this is an inappropriate measure to compare without considering the baseline classification size. With this said, the accuracy is much lower than work by Miller and collaborators[69] given the same algorithm. I attribute this to the drop in in within vs. between-section accuracy as shown by Moore and collaborators[76]. This is part of a larger thread continued in the current work indicating an important factor of identifiability being the delay between enrollment and testing. Furthermore, the focus of this work is not the best accuracy *per se* but a scientific comparison of several factors influencing accuracy. I encourage further development of models to better represent the state of the art attacks to defend against.

6.5.2 Implications for Privacy

This work continues to survey the risks that VR poses to privacy. The most important question in this space is how identifying various data sources, situations, and activities are, what makes these identifying, and what can be done about it. By understanding what influences the accuracy of de-anonymization techniques, researchers can develop more effective and more efficient ways to limit risk to end users.

The bulk of these findings are directed towards researchers that understand the extent to which this data can be identifying. In this vein, I encourage future researchers to continue to investigate the effect of delay on identifiability in their own datasets. This includes focusing on between-session identification, as is also highlighted by [76]. Within-session identification can lead to unrealistically high accuracies. Second, I encourage other researchers to report not simply accuracy but also multiclass AUC so that model performance can be adequately compared across classification sizes.

As an order of magnitude larger in terms of classification size than most previous work, this work also has a unique insight into extension of behavioral biometrics in the setting of the metaverse, in which there may be years of interactions with millions or billions of people. It is clear increasing classification size reduces re-identification risk, but it does not appear as if behavioral biometrics are reached a ceiling in the identifiability they can perform. The difference between this work and a purported metaverse is about four to seven orders of magnitude, but it is certainly clear that this data can be used alongside others to re-identify data, as with other identifiers such as birthdate or zip code [114].

Considering the state of the field as a whole, there are several steps to take. First, developers ought to protect this data with standard practices for personally identifying data [125]. When this

data needs to be shared with others, it can be helpful to reduce the time span available, minimize variation in activities, or modify data to produce security guarantees like k-anonymity, plausible deniability, or differential privacy [80, 29]. Furthermore, there are developments in law that need to be made to clarify the legal status of this data based on its risks to privacy [55].

Ultimately I aim to see the users, designers, and developers of social VR to understand and account for the risks of this medium so that people can safely and consciously use this new and powerful medium.

6.5.3 Limitations and Future Work

Some limitations of this work include that while participants knew their pose tracking data was collected, they were not aware what features of their data would be most identifying so that they could change their behavior to avoid being tracked, e.g. vary their height week-to-week to fool the model. All participants used the same headset for the entire duration of the study, which according to previous work [72, 73] can make identification easier. On the other hand, almost all participants used a Meta Quest 2 headset, so idiosyncrasies across headsets (e.g., tracking errors that appear differently in the Meta Quest 2 headset vs. HTC Vive headset) could not be leveraged as partially identifying.

The activity performed (discussion) and the context (a college course) was selected due to the authors' personal experiences with activities in social VR, but a more formal study of social VR may indicate other activities that commonly take place that an attacker may leverage. The classroom setting for the discussions in this study adds realism in terms of ecological validity, but does increase the natural variance of most factors compared to a more controlled lab study. This, in turn, may lead to less sensitive comparisons.

This work only uses a single model - random forest - and so the results may be limited to this type of model. However, the focus of the work is not attaining the highest accuracy possible given a dataset, but exploring some of the conditions on which the dataset achieves relatively higher and lower accuracy. Therefore, I judged it out of scope to select several models and compare results. Further, I refer the reader to other related work[69, 76, 79] for a deeper discussion of the feature importance beyond what is covered in the Identification Features subsection. Also, the sampling process used to minimize discontinuities (see the subsection on duration) does introduce a bias in the sampling such that the middle of the recorded time is more likely to be part of the testing set than the training set.

The formatting of the task as classification, in contrast to ranking, implies that the value of identification is when there is one prediction from the model and it is correct. However, this biometric identification could be done alongside other mechanisms, in which a rank classification method may better express the value of the identification task.

Regarding attack models, some avenues for future work include demonstrating effective attacks beyond biometrics. For example, depending on what is transmitted, almost all of a target's visual

and auditory experience can be recorded or inferred by a VR lurker. This includes inferences about the target's attention to objects, content, or people in the virtual world coming from both conscious and unconscious mechanisms. These features will likely involve models different from position-based random forest, such as neural networks [73]. On the defensive side, there is still important work to be done on methods of obfuscating pose tracking data to minimize identifiability. Additionally, it may be plausible, given a user's preferences, to disconnect real-world biometrics like height and arm length from a user's virtual avatar entirely, or use transformed social interaction [9] so that gestures that might otherwise be identifiable can come from another person but still be communicative.

One direction for future work is to draw more upon work in the space of gait recognition [123] for attacks, defenses, and models that may carry over into questions of VR pose re-identification. While often VR users do not have enough space to walk for significant periods of time, the whole-body nature of VR pose tracking may provide sufficient similarities.

6.6 Conclusion

With the rising interest in social virtual reality and the broader metaverse, understanding risks to and defenses of privacy becomes an increasingly urgent priority. In this work, I study the ability to identify users based simply upon 18DOF data, position and rotation of head and hands, as is available with consumer VR headsets and common social VR platforms. This setting is unique as I are able to study the effect of delay, duration, and . Ultimately, I find that the number of sessions recorded greatly increases identifiability, and the duration per session used in training also increases identifiability, though to a lesser extent. I also introduce *body-space coordinates* as a mechanism likely to increase the effectiveness of identification. All together, I pursue understanding how motion and pose tracked by VR can be identifiable, and I hope this knowledge can be useful for future defenses of privacy in social VR.

Chapter 7

Discussion

In this section, I recount and synthesize these middle chapters, chapter 4, chapter 5, and chapter 6, through a summary of results, implications for theory and practice, and extended future work section.

7.1 Summary of Results

The first work I presented was on proxemics and gaze. One of the opportunities of this work was the study of these factors over time in large groups. I found that participants increased their interpersonal distance over time and looked at each other more often. I interpret this result on personal space to be caused by the participants' adaptations to the medium: participants also needed more space for the virtual designs, and could still hear each other from father away, and so could still maintain conversations with each other.

The amount of time participants looked at each other increased over time, which was the opposite effect from the results in [7] that found decreasing attention over time. I give the explanation that nonverbal behavior in the current system is more valuable as a social signal than the systems and graphics used to perform the study performed in the mid-2000s.

Expanding to questions beyond time, I found that panoramic spaces also led to greater personal space than constrained spaces, and participant pairs maintained some distance over time, i.e., participants that were closer than average one week were also likely to be closer than average on another week. Variables such as familiarity and liking may have influenced these individual differences, and more follow-up work is necessary.

The next chapter was work on synchrony. Given the wide range of measures of synchrony, I performed an exploratory multiverse analysis of this space of measures. The conclusions I make from these results are that velocity of a tracked point should be excluded from further analysis due to weak content validity, while coordinate system, dimension, and virtual vs. physical motion are equivalent and window size, body parts, and magnitude transformation remain uncertain.

I found a dramatic difference between velocity and speed in measuring synchrony, with speed

measures having far more content validity. It is possible velocity-based measures can capture synchrony, but they are substantially weaker than speed-based measures. After restricting the space of measures to speed, dimension, virtual vs. physical, and coordinate system showed high enough similarity for these branches to be deemed equivalent. I note that this may be due to gestalt motion being the subject of synchrony. I also found that the options within the branches of window size, body parts, and magnitude transformation have content validity but lack consistency. Low consistency in window size may be due to synchrony occurring across time scales and low consistency in body parts may be due to the type of action being performed, but low consistency in magnitude transformation is difficult to interpret.

Correlation between randomly selected measures varied independently across different branches, with a median correlation of 0.39, providing the possibility that statistically significant results could be due to researchers selecting a measure to fit results, rather than selecting a measure ahead of time.

The contrast of high content validity, mixed similarity, and low predictive validity poses a problem for theories of synchrony.

In the third work, the study of identifiability, I show that the effects of time are important with re-identification. The largest factor influencing accuracy is the duration between the testing and training sessions, with the highest accuracy being when the training and testing sessions are adjacent. This is so dominant that one minute of data from the same session is worth more to accuracy than 30 minutes of data from 7 other sessions, given my dataset and algorithm.

I have also contributed to feature selection by specifying body-space coordinates, defining the horizontal plane in terms of the participant's average horizontal head direction for a given time frame, and permit usage of horizontal motion that if left in the global coordinate system would significantly decrease accuracy. I also encourage the use of multiclass AUC in future studies to better enable comparisons between widely varying identification set class sizes.

7.2 Theory

The use of VR motion data in terms of science has been a rich source of insights in this dissertation. One of the findings in the study on proxemics and was that the change relative to real-world interaction depended on both a *push*, as participants needed more space for their dioramas, and a *pull*, as participants could spread out more distant while still maintaining communication. This is in contrast to a simpler model that would elide the necessity of either the push or the pull. This theory aligns with previous work in the study of media, specifically the importing of real-world norms into social environments [133]. Changes in behavior that are simply attributed to VR are ambiguous as to its boundary conditions; making a claim specific to one of the affordances of VR makes that claim more falsifiable.

It is also beneficial to estimate the differences due to pair in interpersonal distance. The effects of some dyad-variables have been explored, e.g., gender composition of pairs, but much variance

remains to be explained in these interpersonal distance settings. One interesting path forward is distinguishing between path-dependent variance and maintenance of distance. In the former, perhaps participants found a location in the room in the first weeks that became familiar and comfortable, and in the latter, perhaps participants are in fact not comfortable near each other and are increasing their personal distance.

The second study, the study of synchrony, posed a strong theoretical challenge: explaining the contrast of high content validity, mixed consistency, and low predictive validity. This is a challenge because it goes against the commonly held theory that synchrony is simply motor coordination due to mirror neurons, and specifically that coordination would only be expected on short time scales.

A potential explanation I propose is that spontaneous coordination is more often a *common response* to a *common stimulus* given *common ground*. If we reduce human behavior to an information processing machine, and assert that better coordination requires more similarity (which is not true in the extreme but may be true in many cases) then the degree to which people have shared responses indicates both a shared attention and a shared way of turning stimuli into behavior, i.e., a shared processing system. This would explain the wide range of time and without requiring a particular biological mechanism; the assertions are simply in the realm of data-processing and social cognition.

However, one cannot downplay the potential explanation that synchrony is having replication issues. With such broad definitions that center around 'doing a similar thing at a similar time' and such a wide range of interesting, relevant, and therefore publishable outcome measures as the many related to social cohesion, rapport, creativity, and others, it is quite possible a sizable portion of work performed on synchrony is due to selective publishing of positive results and post-hoc selections of measurement. The high content validity, mixed similarity, and would seem to be the ideal case for production of non-replicable effects.

This is not to say synchrony is not a valid subject of study. It is clear from the high content validity that interactants *do* do similar things at similar times, and why that is the case is still a meaningful question. Furthermore, a failure of replication for a topic of study more often means that effects are usually smaller or more restricted than previously though, and less often that there is no effect to speak of.

The study of identifiability found that the time delay between training and testing formation was a strong factor in identifiability. Why this is the case exactly remains to be seen. What types of identifying behaviors are consistent on these different time scales but not on others? Understanding this behavior can provide better mechanisms to anonymize this motion data at the right level of fidelity.

If other aspects of human behavior are any indication, it is likely there are many, many processes happening fractally. Changes may happen slowly and persist for a while, or may happen quickly and fade. The time scale at which these behaviors occur is likely to be wide.

7.3 Design

Many virtual reality experiences are being designed to support and encourage person-to-person connection. The findings of the studies presented here, in particular the study on gaze and proxemics, provides several implications for designers and new avenues for exploration. Both the virtual space and the audio can influence interpersonal distance, which in turn is known to influence connection.

Recall that in subsection 4.3.1, there was an effect of environment such that places that had wider views (panoramic condition) had larger interpersonal distances. Because interpersonal distances are linked to constructs like intimacy, the larger interpersonal distance would likely signal lower intimacy and connectedness, all else being equal, which would indicate that larger, less-populated virtual spaces may seem lonely.

Interestingly, this aligns with a popular critique of social VR that describes it as an "empty mall." The comparison is apt: the spaces are highly built up and quite large, but finding other people there is rare. There is almost a sense of eerieness in virtual spaces that the user is left saying "there should be people here, but there aren't." Taking this a step forward, if this sense arrives as the confluence of much space, few people, and an artificial environment, then work in any of these domains may reduce this feeling of eerieness.

Most prominently, there is the question of space. One of the constraints that often exists in the real world that does not in VR is the constraint of the price of space. Anecdotally, I have noticed that spaces in VR are often luxuriously large. In the context of connection, though, this can contribute to the sense of emptiness and eerieness.

Instead of always getting larger, one direction for future design exploration is getting smaller: sharing spaces that were not possible to share before. One instance of this is a co-embodied avatar [61] in which a trainer and trainee occupy the same virtual avatar, and the avatar's pose is a mathematical blend of the trainer and trainee's poses. This approach has its risks as well, certainly visible in this instance as violations of personal space in VR. Regular usage of this technique or other interactions will establish a norm that, in some situations, what constitutes a violation of personal space in real life does not do so in VR. This norm would risk exacerbating the dismissal of users' concerns regarding personal space violations in VR. Having strong social and technical norms around this unphysically-close type of shared space can prioritize users' safety over the small (but tempting) benefits in terms of performance and other features when they come into conflict.

Additionally, spaces are not constrained to be static in VR. It is possible for a space - say, VR live comedy venue - to expand and contract based on the number of people that show up. The smaller space can be cozy; a larger space can bring energy. The mechanics of that expansion and contraction have several open questions, like the risk of disorientation, the salience of the change from explicit to subtle, and the conditions for expansion and contraction.

Further beyond current capabilities, it may be possible to simulate artificial agents that populate a virtual space. Initial work by Latoschik and collaborators [64] indicate that a more populated space may increase social presence. However, interactions with current virtual agents are limited,

and low fidelity social behavior can backfire and in fact decrease comfort and increase eerieness.

Interestingly, in [51], my colleagues and I found that this increased interpersonal distance was also linked to higher entitativity. However, it is not clear why this is the case - it may because people are interacting farther apart in an absolute sense, but relative to the space available to them, they are closer. This may signal that what is not perceived is space in an absolute sense, but space relative to the social norms. For example, it is more acceptable to select a seat next to someone on a bus when the bus is crowded than when it is empty. This question of entitativity and distance is currently under further study. Expectancy violations theory [22] may be useful here.

The second result of note is the effect of audio on interpersonal distances. Hall's work in proxemics [49] defined the outer boundary of conversational space to be the edge of where conversations can take place, and so I judge one of the necessary conditions to find an increase in interpersonal distance is the increase of that boundary through audio modulation that drops off less over distance compared to real-world interactions.

Like most things imported into VR from real life, there is no technical necessity for this to be the case. There are many ways to modulate sound from person-to-person. The current state of the art in the ENGAGE social VR system provided two options: either everyone heard everyone equally save for options for muting (like contemporary teleconferencing software), or there was a volume drop-off to zero at a given distance threshold.

The current design space for modulating sound from one person to another is broader in real-world settings. The default is sound modulated by air, with signal energy dropping off roughly with the square of distance. This allows maintenance of conversations in groups. In settings such as a nightclub, with loud ambient volume, can decrease conversational space, perhaps increasing arousal and liking. Designs like amphitheaters increase one-to-many broadcasting. Technical affordances of handheld transceivers ("walkie-talkies") allow members of a team (tuned to the same frequency and located reasonably nearby) to communicate one-to-many. Cell phones provide one-to-one communication with a single selected other, and with the right technology in support, that other can be located nearly anywhere on earth.

Open directions for audio modulation include audio defined less by distance and more by direction, producing a sort of "spotlight sound" for one-to-few communication. Variables other than affiliation and distance can be used to define modulations, and effects can go beyond simply volume. For example, suppose that in a competitive game, teammates are heard loud and clear, but the opposition is only heard when nearby. Additionally, it may be a opportunity for functioning as part of a team for the audio modulation to only pass information to the nearest or N nearest players, requiring a team to pass along messages in order to coordinate.

When designing for social connection, it must be noted that there are many ways to express connection, and a simple system manipulation to increase intimacy one one channel may only lead to the user leveraging other signals of intimacy to maintain equilibrium. An elevator being small does not encourage people to get to know each other, it encourages them to go to the edges of the space or not face each other. The corollary to this is that each potential drawback - increased

distances, for example - may be offset by other intimacy signals, if they are properly supported. One approach like this is the use of "large heads" for distant display of facial expressions [25].

7.4 Future Work

Individually, each of these works has inspired follow-up work. These proposal vary in scope: some are narrow to be simply add-on analyses, some are large to be papers or whole topics of study. Taken together, these works also indicate further paths of work that extend beyond just one of these projects.

7.4.1 Proxemics and Gaze

There remain questions to ask in regards to this dataset itself. For example, what is the effect of starting interpersonal distance or familiarity on continuing interpersonal distance over time? It is reasonable to expect that people will be closer with people they know, which can be validated in this dataset. Familiarity may have also moderated the rate at which people disperse over time.

I would also like to explore ways to automatically distinguish between "tones" of personal space approach. Based upon my experience in these discussion sections, there are instances of several (unintentional) violations of personal space. Are there statistical signatures visible as to the "intention" of the violation? This would provide bubbles that do not simply maintain distance, but permit passable bubbles when performing work in close proximity.

One measure that has been recently in use in concurrent work in this dataset is what I and other researchers have termed “social attention", the proportion of time in which one person is near the center of another’s field of view. Given a distribution of head positions, what patterns are there in head direction and social attention can be inferred? If this is performed, social attention can be broken into attention due to position effects and to otherwise personal effects. Given this setup, one may find that, for example, people with a particular personality type are more or less likely to look at someone.

7.4.2 Synchrony

The most fascinating question to me that comes from this work is using VR in the paradigm of transformed social interaction to distinguish between context-based coordination and spontaneous coordination. That is, we can synthesize some gesture in some random moment of an interaction, and display that gesture to another participant. The research question becomes whether the person synchronizes more with the real gesture or the generated gesture. If it’s the generated gesture, we have relatively more synchrony due to what we see. If it’s the real gesture, then more synchrony is due to the context and shared information.

This development requires subtle manipulation of participant movement, involving good synthesis of motion, predictive abilities, and interpolation between movements, all in real-time.

A question this dataset can answer directly is the nature of synchrony in groups. Who is synchronizing with whom, how, and why? Of course, an analysis like this presumes synchrony is occurring, and so difficulties in understanding synchrony may carry over to this dataset itself. Nevertheless, the area of work exploring synchrony in groups larger than two is very limited. While there is some work indicating pairwise and group can be equivalent for a certain correlative method of calculating synchrony [71], it remains to be see whether that method is theoretically meaningful.

Another simple analysis to perform on this data is a replication of [71] that synchrony was higher when participants were *not* looking at each other compared to when they were. Is it related to task? If so, how?

Additionally, there can be clearer ways to demonstrate some of the claims made in chapter 5. For example, it appears in the data that head and hands synchrony are different. However, part of the problem is that when one's head is moving, often one's hands are moving too. It is possible to model this as a set of correlations and see the relative influence of several factors, perhaps distinguishing between body parts even more. The same can apply to teleporting, smooth movement, and physical movement.

To distinguish between the two proposed explanations for high content validity and mixed similarity, it would be possible to discuss whether synchrony would be noticeably higher when participants were tracked in the same way compared to different ways. For example, if measures where yaw was correlated with yaw, vertical-vertical, roll-roll, etc, were significantly higher than, then it would not make sense to argue that gestalt motion is the source of synchrony. If those heterogeneous measures were related, though, it could be (with the right modeling) appropriate to dismiss that mimicry explains all (or even most) of spontaneous coordination in nonverbal synchrony.

Finally, I note that the multiverse proposed in subsection 5.2.2 can be expanded, including various ways to link two streams of data other than Pearson correlation. These are covered quite well by Novotny and Bente [83].

7.4.3 Identifiability

Based on the current work in identifiability, there are a handful of directions for future work. One spin-off project is an investigation of the distribution of how things are misclassified. What is the real relationship between group size and identifiablility? For example, it is not as if every pair is independently distinguishable - then we'd see a binomial distribution of errors, which do not match up with observations.

There are other approaches to defending this data. One simple option is creating a standard library of low-grade motion if participants are outside of a certain radius of the person, or are within a certain number of people of the person, based upon who is likely being paid attention to.

7.4.4 Nonverbal Behavior in Communication

The work across these three domains allowed a more generic position on the use of motion data in general. One theme across these three studies is the potential to dive further into the temporal resolution of the data. To avoid endless tuning and open-ended analyses, the study of dynamics was kept very simple. However, I believe this will be fertile ground for further discoveries. The tension will be between data-driven approaches that may be effective but theoretically meaningless and approaches built upon vague or overextended theory that prove ineffective. Explainable AI may provide avenues forward for this scientific tension, or current models that capture much of the motion participants have already made do provide a starting point for a lower-dimensional representation of motion.

Chapter 8

Conclusion

In this work, I have demonstrated that the collection of head and hands pose data in current consumer VR, an instantiation of the paradigm of VR as an observatory, enables collection of body motion data in high temporal and spatial fidelity in large groups which is valuable for quantitative analysis regarding the study of nonverbal communication and poses risks to immersant privacy through biometrics.

In the first work, I studied gaze and proxemics. The large-group data led to findings that participants adapted to the medium and maintained their interpersonal distance over time. This would have been much more difficult to study with more obtrusive or lower spatial or temporal resolution in traditional settings.

In the second work, I studied synchrony. By using this data, it was more expedient to investigate synchrony in groups. The spatial and temporal resolution was necessary for this study. Furthermore, this motion data provided a expressive bases from which to generate many measures of motion and therefore many measures of synchrony. The results highlighted high content validity, mixed consistency, and low predictive validity. Furthermore, these results are not due to a lack of spatial or temporal precision. The results together pose opportunities for researchers to adapt and develop theories of synchrony.

In the final work, I show that this motion data is also a risk to user privacy. I demonstrate that the time between recording and inference is very important for this identifiability. This poses a risk to several parts of the VR experience, but most prominently in social VR when one's motion is available to other server members.

To close, virtual reality is a fascinating and powerful concept: the immersant is in a computer-controlled world, whether physically operated as Sutherland had expressed or generated to the senses as Lanier expressed. Because this world is interactive, totalizing VR as a display implies totalizing VR as an observatory, and steps toward VR as a better display require steps toward VR as an observatory. More data from this observatory is beneficial to researchers who perform the scientific process, certainly, but it also poses a risk to user privacy.

It remains to be seen how we - designers, researchers, programmers, investors, businesspeople, purchasers, VR regulars, and many more communities of people - decide how to design and re-design this technology of VR in the coming years and decades.

Appendix A

Sampling from a Multiverse

In total, there are 9,300 universes within this multiverse. This, combined with the fact that there were 5341 attendance pairs in the dataset, each of which had approximately 50,000 frames of motion, made it computationally prohibitive to calculate all measures upon all attendance pairs. With an empirically estimated value of one second per analysis (one for each combination of universe and pair) it would have taken over a year and a half of continuous computation for the analysis to complete. Instead, I propose a method of sampling the multiverse such that conclusions can be drawn both based on the space as a whole and on direct comparisons between options.

A.1 Sampling a Universe from a Multiverse

The simplest approach for sampling a multiverse is sampling each branch individually. However, this can lead to biases depending on the order in which branches are sampled if conditions exclude some values. The next simplest approach is to sample over the entire space, and then apply conditions, but the time it takes to generate a successful universe increases exponentially with respect to the number of conditions (more rules mean less of the space is a valid option). It is also possible to specify all possible combinations, then filter by all conditions, but this explicit specification grows exponentially by multiverse size, which became prohibitive when I explored separate choices of most these branches for each member in the pair (increasing the total number of analyses to 2 million). Instead, I exploit the fact that the conditions in this multiverse are in different disconnected clusters. For example, while signed motion, dimension, and coordinates have conditions together, they are independent from all other measures. With this decomposability, it is possible to break the multiverse into several collections of branches independent from each other. Each of these can be sampled independently to produce a universe sampled from the full space of valid universes.

A.2 Sampling Pairs of Universes

This method provides individual random universes, but if universes were sampled randomly, it would be exceedingly difficult to pair multiverses to estimate the similarity between two options within the same branch. For example, to estimate the effect of switching a measure using head to one that uses hands, all else being equal, one can take the correlation of many pairs of evaluated universes that differ only within that variable. However, the probability of both universes in the multiverse being evaluated would be the square of the proportion of universes sampled. Considering I have just above 0.11% of the space sampled (10 per 9300 for each pair of participants), there would be almost no pairs on which to calculate similarity. The approach I take when investigating the similarity of synchrony measures is to instead compute all synchrony measures along a random walk from two universes selected uniformly at random. This process can be approximated by first selecting a random universe, then randomly ordering the branches. For each branch, select a random valid universe that varies from the current one only on that given branch. This produces a sequence of valid universes for which adjacent universes differ by only one branch. Repeatedly sampling and calculating synchrony in these ways ensures sufficient coverage for pairwise comparisons.

Appendix B

Wavelet Analysis Illuminating Window Size Effects

B.1 Wavelet Analysis by Time Scales

Overall, the results in subsection 5.3.5 indicate that while all options are good at distinguishing synchrony, shorter time scale windows appear to be better, and the choice of window influences synchrony. However, it is important to note that window size is not the time scale in which synchrony occurs. Rather, windowing works like a high-pass filter. Providing a window to the motion signal cuts away the low-frequency (long-period) factors at play, and focuses on the short time scale. I argue there may be long-period factors of synchrony, such as the level of energy in a conversation rising as one friend tells another a exciting story about recent travel but then falling as the topic shifts to a more serious, somber tone at the passing of a mutual friend. Second, this filtering provides synchrony relative to the variation present in that range of frequencies, meaning one cannot necessarily infer synchrony at a time scale through the increase or decrease in synchrony when the window size expands to include the given time scale. This fact prompted a deeper look at the question: how do people synchronize across time scales?

The approach we take to answer this question uses wavelet analysis. In wavelet analysis, a one-dimensional signal in the time domain is decomposed into the 2D time-frequency domain, specifying for each frequency at each moment in time the amplitude and phase of a wavelet that best fits the signal around that point. This provides a benefit over a simple Fourier transform when studying signals that change amplitude and phase over time. A Fourier transform assumes that oscillations leading to the signal are largely consistent in phase and amplitude over time. This assumption does not hold in with complex, adaptive nonverbal behavior. Mathematically, the wavelet transform $z = [W_\psi]m(t, T)$ of the function of motion $f(x)$ given a wavelet $\psi(x)$ receives a time t and a period T and returns a complex number z for which its absolute value $A = |z|$ represents amplitude and

$\theta = arg(z)$ phase of a wavelet with period T around the given signal of motion $m(x)$ around time t. This is done with the equation

$$[W_\psi f]m(t,T) = \frac{1}{\sqrt{T}} \int_{-\infty}^{\infty} \psi\Big(\frac{x-t}{T}\Big) m(x)dx$$

Applying this wavelet transform to each participant within a pair produces two amplitudes and phases for a given time and frequency. From these values, a researcher is able to both isolate motion to a specific frequency and determine whether or not this motion is in-phase or out of phase. Then, by computing an average of the amount of motion and its phase alignment over time, the researcher can provide a far richer picture of synchronization than simply manipulating window size.

To compute the amount of motion (energy), denote the amplitudes A_1 and A_2 for participants 1 and 2 within a pair, θ_1 and θ_2 for phase in the same way, and t_0 and t_Ω for the beginning and end of the recorded time, respectively. For any period T, we compute two values: first, the energy carried at that frequency between the two participants, which is the time-averaged product of the wavelet-estimated amplitudes:

$$E_s(T) = \frac{1}{t_\Omega - t_0} \int_{t_0}^{t_\Omega} A_1(t,T)A_2(t,T)dt$$

This value represents the total variation in this frequency band to which participants in a pair could be synchronizing with, and is akin to the product of standard deviations that functions as the denominator in Pearson correlation.

Using the same denotations, phase alignment $S(T)$ is given by

$$S(T) = \frac{1}{t_\Omega - t_0} \int_{t_0}^{t_\Omega} A_1(t,T)A_2(t,T)cos(\theta_2 - \theta_1)dt$$

When participants are perfectly in phase, the phase alignment is 1, when they are completely opposite, phase alignment is -1. This represents an average of similarity weighted by the energy (variance) available at the given moment, and is akin to the numerator in Pearson correlation.

B.2 Comparison to Previous Wavelet Analyses

It is important to note that this is a different measure from previous work that leverages wavelet analysis. We take this different approach for a theoretical reason considering synchrony. Usually the value from a wavelet analysis that is reported in the synchrony literature [39, 48] is coherence. This value in a traditional analysis is most similar to R^2, the proportion of variance of one variable one can explain given another variable. However, what makes coherence different from the classic definition of R^2 is that coherence is not the result of fitting a linear model but a model with two parameters, amplitude and phase. With these two parameters, and considering that wavelet analysis allows relationships to vary by time and frequency, any wavelet can be fit perfectly to any other

wavelet, as these are the other two parameters that define the wavelet once its time and period are set.

Taking this into the domain at hand, it is as if a researcher pointed to an arbitrary point in the wavelet transform, calculated the wavelets around that point for that time period, and asserted the wavelets were perfectly "in sync." When the enthusiastic researcher is pressed upon their strange assertion about wavelets that look substantially different, the researcher replies - yes, they are different in amplitude, but you wouldn't want to ignore effects simply because one person moved less, would you? And yes, they are different in phase, but you won't want to ignore an effect simply because one is a time-lagged version of the other? However, our researcher is in a bind, because there are no degrees of freedom left in the model. If one permits any change in amplitude or phase, any two signals are synchronized at all time points.

The solution to this problem in domains that use wavelet transforms (e.g. geophysical applications [46]), is to assert there is some consistency over time in terms of phase and amplitude in any relationship of interest. This is done by smoothing both the combined wavelet value and computing the coherence upon the smoothed values. Bringing this back into the domain of synchrony theory, this smoothing is an assertion that the time-lag (relative phase) of participants does not abruptly change but instead changes smoothly according to some filter, and that the amplitude of synchrony also changes smoothly across time.

This assertion is certainly appropriate when the duration for which two signals would be coherent on a given period is much larger than the period itself. For example, in Grinsted's example, their two signals (the Arctic Oscillation and Baltic Maximum Ice extent) had coherence that extended for about 80 years at the period of 12-14 years. In the study of synchrony, it is not clear whether synchrony ought to be limited in this way.

However, the greater concern is that the rate at which participants synchronize and de-synchronize is not even discussed: it is left implicit in the parameters of smoothing, which are often left as default values based upon the software used. This concern is compounded by the fact that Grinsted et al. point out that smoothing values are highly influential on what degree of coherence is considered significant [46, section 3.4, figure 4]. Together, these make coherence difficult to interpret, especially when coherence is positive or negative but still near zero.

Our resolution to this dilemma is to assert phase alignment is necessary for synchrony. This permits arbitrary and instantaneous changes in the amount of motion (wavelet amplitude) and time offset (wavelet phase offset). In response to the enthusiastic synchrony researcher, we will ignore effects where one signal is a time-lagged version of another, to the extent that the lag is of similar size to the period in question. The upshot to this is that the analysis performed is analogous to Pearson correlation of the signal only available on a certain frequency band. This is an easily interpretable result: it has broken down a common measure of synchrony, Pearson correlation, and applied it to several frequency bands.

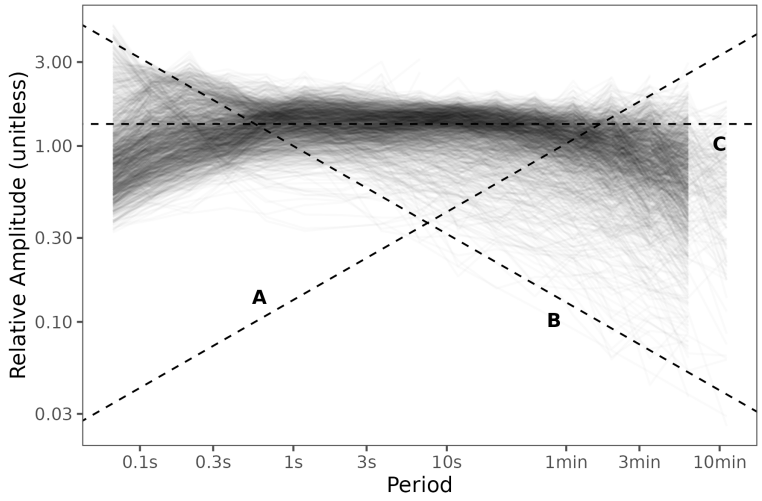

Figure B.1: Distribution of relative amplitudes of the motion signals across periods. The horizontal line indicates pink noise. The horizontal axis is the period of the signal, the vertical axis is the relative amplitude of the signal at the given period. Each solid translucent line represents the power spectrum of a single participant. Black dotted lines are labeled and give reference values for types of colored noise: A is Brownian noise (also known as red noise or integrated white noise), B is white noise, and C is pink noise.

B.3 Synchrony by Period

With this method established, we now report results in two plots. The first is the amplitude of participant motion for a given period, $E_s(T)$. The second is the proportion of that energy (amplitude squared) that is synchronized, $S(T)$, i.e., the Pearson correlation for signals around that frequency. In the first plot, Figure B.1, we find that the amplitude is relatively constant across a range of scales. This is a characteristic of "pink noise", which is a signal where there is approximately equal energy per octave. In terms more common outside of wavelet analysis, there is equal variance in the signal due to oscillations with periods 1-2s, 2-4s, 4-8s, etc. What this means for synchrony is that the denominator of our measure, analogous to the product of standard deviations, is relatively constant across the spectrum: that is, long-period fluctuations, for most participants, have a similar weighting as small-period fluctuations, especially in contrast to Brownian or white noise.

The tendency for signals of a certain to sync up - i.e., their average phase alignment - is given in Figure B.2. We find that very short and very long signals have a stronger tendency to synchronize. We believe the short-scale synchrony is due to artifacts in the recording process (See section XX) and the longer time scale synchrony is due to the task structure, which encouraged certain types of movements at certain times (e. g., hands during making, head during conversation). Nevertheless, the degree of motion was positive at every period, i.e., the average phase alignment for each period was positive, not negative or zero. This indicates that across the whole spectrum, participants were

on average moving similar amounts at similar times.

There is also a trend that variance in average phase alignment increases with period. This trend is due to the fact that given the same duration, longer periods mean fewer cycles, and the cycles are proportional to degrees of freedom. Nevertheless, the increase in average phase alignment outpaces this increase in variance.

The product of the two values under study, the amount of motion and the average phase alignment, lead together to the final value, the relative contribution of each period to measured synchrony (i.e., correlation). This is plotted in Figure B.3. In this plot, the contribution to synchrony scores are more positive (on average) in the 10s to 5min range than other bands, but all bands have some degree of positive contribution to synchrony.

Ultimately, how should the effect of window size on synchrony be understood? First, the selection of a window for windowed correlation should be viewed as a threshold beyond which longer periods are to be ignored. This is not simply a practical concern, as if it is the removal of minor portions of noise. It is a theoretical consideration, that any synchrony happening at time scales beyond the window is ultimately asserted to be not of interest. This is brought to the fore by showing that in our dataset (aligning with many analyses of other aspects of human perception and behavior [4, 33]) activity occurs across a wide range of temporal scales, evidenced by the pink noise distribution of motion over time.

Then, we show the distribution of contributions to the synchrony score by period. There is an indication of synchronization across the whole spectrum, though there are particularly high synchronizations at the very short end (due to artifacts) and the long end (due to task coordination). This helps to explain the impact of window size on synchrony measures. If the time period varies between recordings by a significant proportion, then there may be different aspects of behavior being captured. Therefore, our recommendation is to set a window size explicitly (rather than the entire length of the session) or better yet, use wavelet analysis or Pearson correlation on several band-pass ranges to better describe synchrony across time.

Empirically, we find evidence that participants synchronized across a wide range of time scales. This can be on very short recording spans, and while we were not able to be sensitive in this range because of autocorrelation, we are confident in previous work (cite) showing tendencies to sync at this scale. This corresponds to the upper end of Newell's biological band which can be attributed to mirror neurons. We investigate synchrony throughout the cognitive band and into the rational band, with task-level synchronization. Continuing this thread, we also investigate the social development of synchrony in the weekly band, as people worked into weeks, in section XX.

When studying time scales, we found that largely, each octave of time (each range in which period doubled) provided a similar amount of energy (variance). This is characteristic of pink noise. This also means that correlative synchrony is influenced by the entire range of an interaction. This motivated a much more explicit and intentional selection of window size, or even better, a breakdown of synchrony by period (or conversely, frequency).

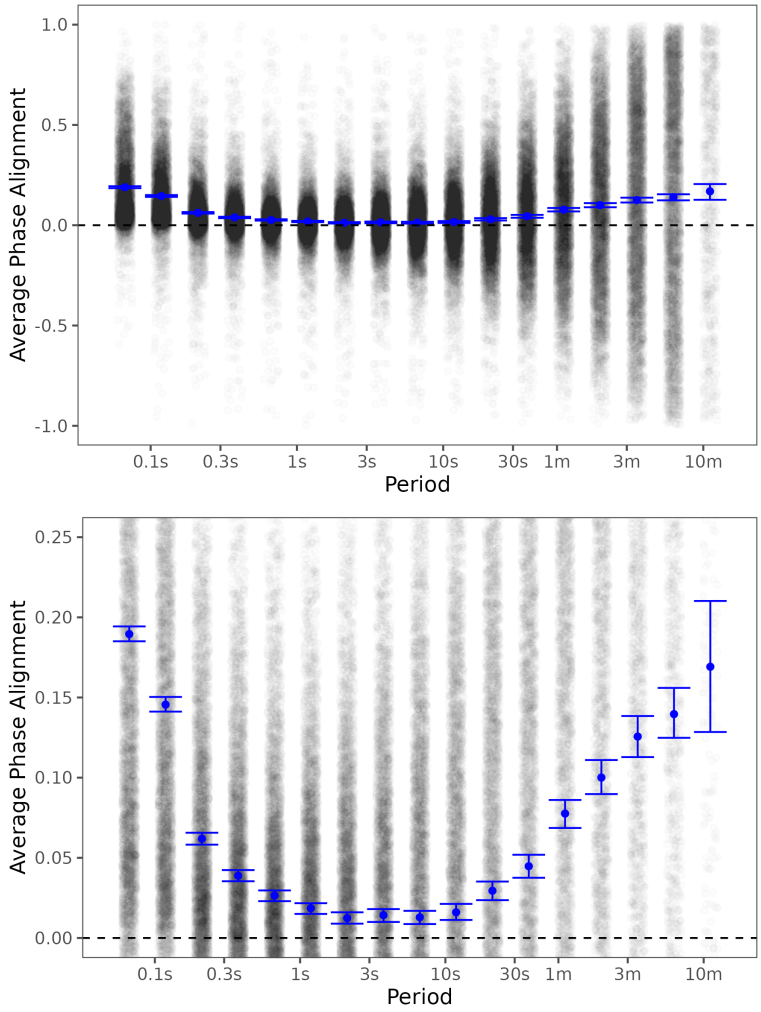

Figure B.2: Phase alignment by period, shown in two panels. Phase alignment is particularly high for short periods (less than 0.15s) and long periods (>30s), but was on average positive for all periods. The horizontal axis indicates the period, and the vertical axis indicates phase alignment of the signal at that period as given by the wavelet analysis. Blue dots and error bars indicate means and 95% confidence intervals of the mean. The second panel is the same data as the first panel, just with an expanded y-axis.

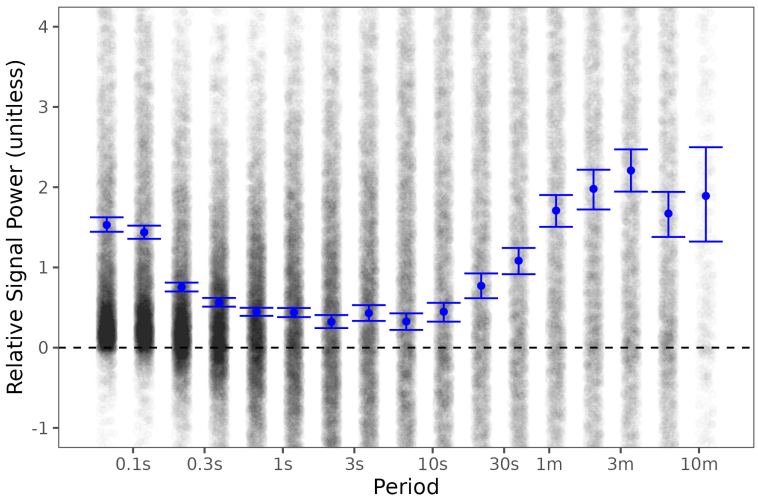

Figure B.3: Relative synchronized signal power by period. Contributions to positive synchrony are high for short periods (less than 0.15s) and long periods (>15s), but are on average positive for all periods. The horizontal axis indicates the period, and the vertical axis indicates power of the synchronized signal at that period as given by the wavelet analysis. Blue dots and error bars indicate means and 95% confidence intervals of the mean.

Appendix C

Predictive Validity

Because we have some questionnaire data collected, we also have the ability to test these measures relative to constructs previous related to synchrony. These are done using methods traditionally found when relating one construct to another - in our case, a mixed-effect model. When interpreting these results, we include both uncorrected p-values and false-discovery-rate corrected p-values. Because the point of an exploratory multiverse analysis is to reduce the measure space, we assert *a priori* the question is not whether these measures are related to the construct of interest, but instead which measures are related to the construct of interest. Therefore, we caution against the common interpretation to dismiss uncorrected p-values, but instead encourage researchers (to the extent they believe there is a relationship between synchrony and these constructs) to find which measures may be most fitting.

In addition to the investigations of content validity and consistency, we also explore the predictive validity of synchrony. Predictive validity can be less reliable, because a lack of predictive validity is indistinguishable from problems with the experimental design from the results only. On the other hand, using a measure to predict other constructs is often the end goal of the study of a construct, so when predictive validity is particularly valuable.

There are three outcome variables we use for predictive validity that have previously been linked to synchrony: entitativity, familiarity, and attraction. Each of the following analyses use a mixed-effect model, but the structure varies because the unit of analysis at which these values were collected were different.

Then, for each of the three outcome variables, we perform analyses on three measure types. The first analysis is on the whole dataset, i.e., any measure of synchrony except for those that use velocity. However, the goal of this process is to distinguish between measures of synchrony, i.e., we are expecting that some measures of synchrony will relate and some will not. For this goal, we perform two more analyses: the second analysis is repeated one for each option within branch, like is done in the other analyses on content validity, and the third is performed on the dense sample, so that added variance is minimized by removing variations due to different measures.

C.1 Entitativity

Entitativity was measured by seven items adapted from Rydell and McConnell [99] using a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree). Sample items include "My discussion group is important to its members" and "Members of my discussion group are affected by the behaviors of other members." Weekly entitativity scores were calculated as the mean of the seven items (Cronbach's $\alpha = 0.9$ in course 1, α=0.86 in course 2), with higher scores indicating greater entitativity.

Synchrony is often related to variables of interest such as group collaboration. In this analysis, we fitted a mixed-effects model using lmerTest upon the combined dataset. The fixed effects included all used in the analysis of time (see Appendix D) but also included synchrony. The random effects were the same as the analysis of time. The model was predicting the entitativity averaged between the two participants within the pair. The effect of synchrony on entitativity, a decrease of 0.00476 points (on a seven-point scale) per standard deviation, was not significant according to a t-test with Satterthwaite's method of estimating degrees of freedom, $t(4440) = -1.350$, $p = 0.177$.

In the second type of analysis, we investigate the relationship between entitativity and synchrony where the measures are conditioned on one option. We then have 36 analyses. Of these 36, only 2 showed a significant relationship with entitativity, and neither was statistically significant given false-discovery-rate correction. The first was measures that limited the tracked body parts to hands. In this, the relationship of synchrony to entitativity was a decrease of 0.0094 points (on a seven-point scale) per standard deviation of synchrony, $t(2497.04) = -2.946$, $p = 0.00325$, $p_{FDR} = 0.117$. The second was measures that used no magnitude transform, which had a decrease of 0.0077 points (on a seven-point scale) per standard deviation of synchrony, $t(2278.37) = -2.237$, $p = 0.0254$, $p_{FDR} = 0.457$. Because both were in a range expected by chance considering false discovery rate, we do not give them much weight in assessing predictive validity.

In the third analysis, using the 30 randomly selected universes in the dense sample, no measure had a significant relationship to entitativity. The result closest to significance was a decrease of 0.038 on the entitativity scale per standard deviation of synchrony, with an uncorrected p-value of 0.07, and its FDR-corrected value was 0.993. Again, we do not find any evidence for predictive validity distinguishing between measures.

C.2 Familiarity

We also collected data in one of the data collection periods, the summer, on the degree to which synchrony could predict two ratings of liking we collected from participants at the end of the experiment. We used a mixed effects model but with pair (rather than pair within meeting) as the unit of analysis, so data was collapsed across synchrony measures and weeks.

Familiarity was measured by a single-item question by each participant for each other participant within the same section. The question text was "Consider the members of your discussion section.

How well do you know this member?" The response distribution was "Not familiar at all", 39.8%; "Slightly familiar", 30.2%; "Moderately familiar", 17.0%; "Very familiar", 7.06%; and "Extremely Familiar", 5.90%.

In the first model, there was only the fixed effect of synchrony with a random effect of section. The relationship of synchrony to familiarity was an increase of 0.0157 points on a five-point scale per standard deviation of synchrony. This relationship was not significant according to a t-test with Satterthwaite's method of estimating degrees of freedom, $t(401.0) = 0.328$, $p = 0.743$.

The second set of models, akin to the one on entitativity, was repeated 36 times, one for each option within each branch. In this analysis, only one analysis produced a significant result, which was not significant when controlling for false-discovery rate. By limiting to measures that select a window size of 10000 frames (5.5 minutes), synchrony was positively related to knowing such that the knowing score increase .138 points on the five-point scale per SD of synchrony, $t(350.26) = 2.525$, $p = 0.012$, $p_{FDR} = 0.433$.

The third set of models used the 30 measures in the dense sample. Of these models, 13 of 30 showed a significant *negative* relationship between synchrony and familiarity, which only fell to 9 of 30 when correcting within this analysis with false discovery rate. The small model set size allows the production of a specification curve, which matches each model to the choices that were made in its production. This curve is plotted in Figure C.1.

The specification curve plot should indicate if measures that take some branch cluster at the higher or lower end of the range of the effects. In Figure C.1, observe that the 1000f (yellow) and 100f (brown) options in the Window Size facet (bottom) are biased towards the left side, which is, for the most part, where the relationship is. From this, I conclude that synchrony on shorter time scales seems to be driving this relationship. However, it should be noted that this is not mirrored in the second set of analyses.

C.3 Attraction

Attraction was measured by a single-item question by each participant for each other participant within the same section. The question text was "Please describe the extent to which you agree with the following statement: I would be excited to get to know this member better." The response distribution was "Strongly disagree", 1.69%; "Disagree", 3.27%; "Somewhat disagree", 1.05%; "Neither agree nor disagree", 36.2%; "Somewhat agree", 19.1%; "Agree", 29.6%; "Strongly agree", 9.06%.

We also performed an analysis of the same form but with the dependent variable being attraction. The relationship between synchrony and attraction was a decrease of 0.0296 points on a seven-point scale per standard deviation of synchrony, and was not significant according to a t-test with Satterthwaite's method of estimating degrees of freedom, $t(400.8) = -0.626$, $p = 0.532$.

The second set of analyses, like the others, was conditioned upon one option in each branch. Of these, two models were significant, neither when correcting for false discovery rate. First, by limiting to measures that select a full window size, synchrony was positively related to attraction

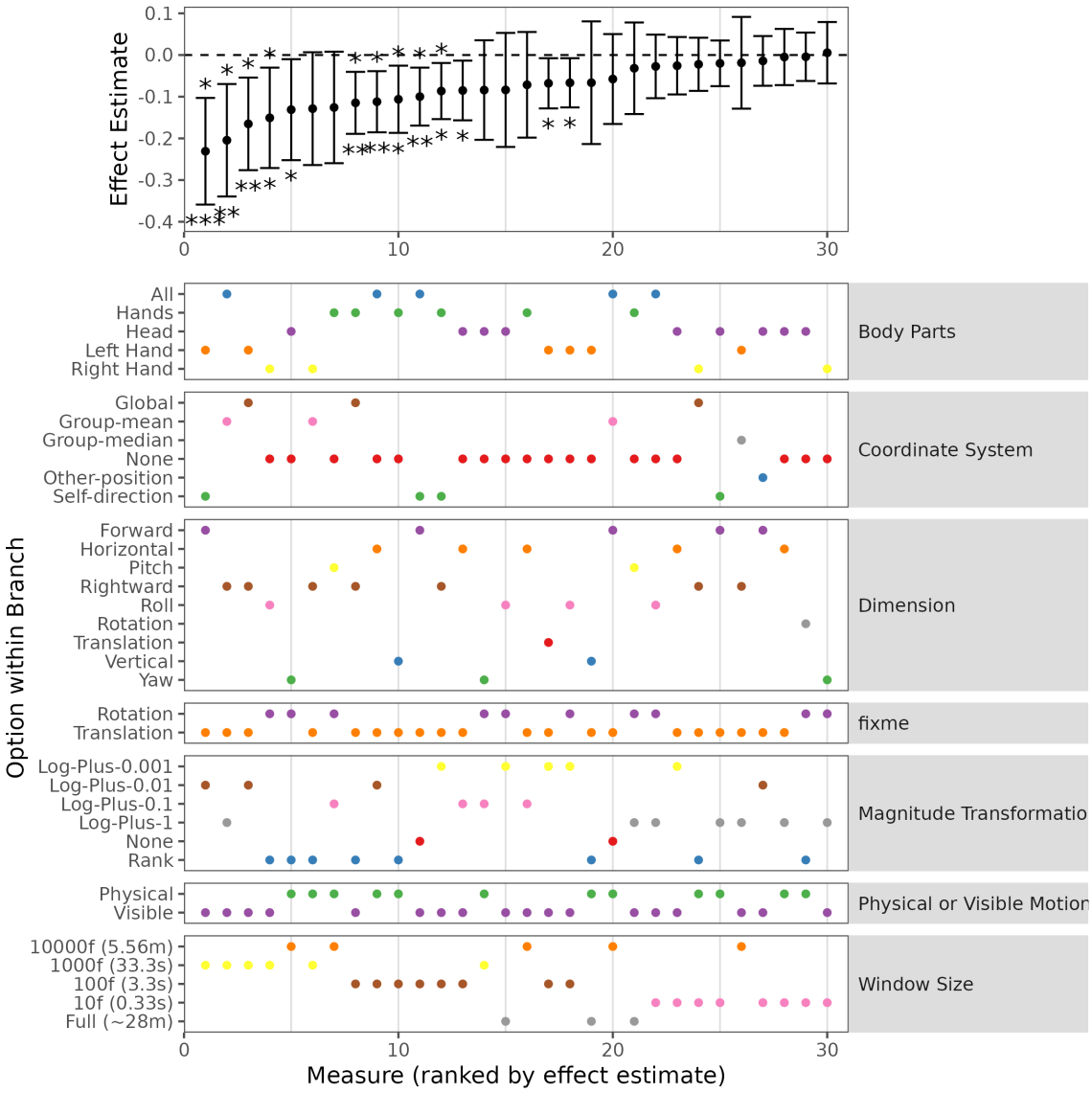

Figure C.1: Effect of synchrony and familiarity, conditioned on the 30 synchrony measures from the dense sample. The top plot shows the 30 universes ordered by effect estimate, largest negative effect to largest positive effect. The faceted plots below show which options were taken in each branch in the construction of that universe (measure). Color redundantly encodes option within branch (y-axis) to permit easier perception of patterns within a branch.

such that attraction increased 0.115 points on the seven-point scale per unit (SD) of synchrony, $t(349.15) = 2.87$, $p = 0.00435$, $p_{FDR} = 0.115$. Second, by limiting to measures that select a window size of 10000 frames (5.5 minutes), synchrony was positivite related to attraction such that attraction increased 0.149 points on the seven-point scale per unit (SD) of synchrony, $t(350.02) = 2.743$, $p = 0.0064$, $p_{FDR} = 0.115$.

In the third set of analyses, the ones that use the dense sample of measures and participants, there was no significant relationships, not even relationships without correction.

C.4 Conclusions

In addition to the content validity and consistency measures, I report the results of predictive validity. This was done through three measures: entitativity, liking, and attraction, and done with three types of models, the first with all measures together, the second using the sparse sample from the main chapter, and the third using the dense sample from the main chapter. Within each analysis type, only the liking in the dense sample had significant relationships with synchrony after controlling for false discovery rate. Specifically, nine of the thirty models showed a relationship, with almost all with 100- or 1000-frame window sizes (3.3 or 33.3 seconds). However, this was not reproduced in the sparse sample, leaving the overall results unclear. Recent works [6, 103] indicate difficulties in linking behavioral measures of sync to constructs of interest, which may indicate problems in theory or in measurement of sync, which is part of the motivation of this work.

It is worth noting that if a researcher were to object to this analysis on the grounds that all $3 \times (1 + 30 + 36) = 201$ models ought to be put into the same false-discovery rate test, then no results would be significant. This might be justified, for example, if it still needed to be proved that synchrony was related to these constructs. However, we caution against such a conservative reading of these tests because the goal of an exploratory multiverse is to in fact "deflate the multiverse" [109], which should bias interpretations toward selecting measures.

Appendix D

Change in Synchrony over Time

One of the unique opportunities of this dataset is the ability to investigate synchrony over time. We collected interactions over the course of eight weeks and so can track the development of synchrony over time of pairs of users. In this analysis, we fitted a mixed-effects model using lmerTest upon the combined dataset. The fixed effects were week (numeric) and the independent variables in the summer and fall settings. There were five random effects: four were *pair* nested within *section* nested within *section leader* nested within *course*, and the fifth was the interaction of week and course (i.e., the task the participants performed). The model was predicting the z-score of synchrony averaged across ten randomly selected measures (excluding velocity as an option) for each pair. The effect of week on synchrony, an increase of 0.0607 standard deviations per week, was not significant according to a t-test with Satterthwaite's method of estimating degrees of freedom, $t(12.83) = 1.612$, $p = 0.131$. From this, we did not see any significant change in synchrony over time.

Appendix E

Multiclass AUC

To our knowledge, no work in the space of user identification with VR data has used multiclass AUC. Because it addresses the effect of classification size on accuracy, we give a short description and justification of its use in enabling future comparisons across studies with varying numbers of classes.

Identification-focused works [88, 69, 76, 75] almost exclusively use accuracy for the model's evaluation metric. The benefits of accuracy as a metric include its ease of interpretation and its directness to the question at hand - a less accuracy model is obviously less identifiable, and vice versa. However, accuracy does vary significantly as the number of classes varies, even for the same data distributions and identification processes, as evidenced by multiple works [88, 69, 124]. Intuitively, this is true - it is easier to guess who is walking up the stairs in an apartment with two other people than a house of ten. This effect of the number of classes on accuracy can make synthesis of findings across works difficult, as the classification can vary as much as two orders of magnitude (e.g., 5 in [124] to 511 in [69]).

Our criteria for an evaluation metric that addresses this issue is that it produces the same value regardless if it is computed upon the full set of classes, or computed as the average of randomly chosen subsets of classes of any size. More formally, let $\mathcal{C}$ represent a classification problem whose elements $C \in \mathcal{C}$ are sets containing individual members of the class C. We define an ideal evaluation metric $\mathcal{M}$ such that the evaluation $\mathcal{M}(f, \mathcal{C})$ computed from the prediction function f and the classification problem $\mathcal{C}$ is equal to the expected value of the evaluation $\mathcal{M}(f, \mathcal{C}')$ for a randomly chosen combination of classes $\mathcal{C}'$ of a given size N, uniformly randomly selected from the classes in $\mathcal{C}$. Numerically, this is:

$$\mathcal{M}(f, \mathcal{C}) = \binom{|\mathcal{C}|}{N}^{-1} \sum_{\mathcal{C}' \subseteq \mathcal{C}, |\mathcal{C}'|=N} \mathcal{M}(f, \mathcal{C}')$$

To solve this problem and enable comparisons across analyses with varying numbers of classes, we choose our primary evaluation metric to be *multiclass AUC*, defined by Hand and Till [53].

Multiclass AUC can be described as the average of the pairwise separability between classes. In the original work, Hand and Till extend area-under-the-curve (AUC), the well-known measure of separability, to the multiclass case. AUC can be expressed as the probability that a randomly selected member a of class A will be larger than a randomly selected member b of class B according to the value of the binary prediction function f_{binary} meant to separate the two. This can be easily computed in closed form as

$$AUC = \frac{1}{|A||B|} \sum_{a \in A,\ b \in B} \mathbf{1}[f_{binary}(a) > f_{binary}(b)]$$

where $\mathbf{1}$ is the indicator function. Multiclass AUC extends this definition provided a multiclass prediction function f that specifies values $f(m, C)$ for each combination of member m and class C. From this, Hand and Till [53] define the multiclass AUC for a given prediction function f and set of classes $\mathcal{C}$ to be the average of separabilities of one class from another for all pairs of classes in the model:

$$\frac{1}{|\mathcal{C}|(|\mathcal{C}| - 1)} \sum_{A,B \in \mathcal{C},\ A \neq B} \frac{1}{|A||B|} \sum_{a \in A,\ b \in B} f(a, A) > f(b, A)$$

As a sketch for the proof that metric fills the criteria above, consider that this metric produces a separability value for each ordered pair of classes independent of the other classes present. Due to the symmetry of classes in being selected within the final set, each class and class pair is weighted similarly in the averaging process. By linearity, the average of the final values for a given class size can be understood as the average of all pairwise values, which produces the same value as evaluating the full set of classes.

Hand and Till note that this metric weights the separability of each pair of classes equally regardless of the number of samples in the classes, which may not be appropriate if priors are to be taken into account. Additionally, this is not an estimate of the accuracy attained by the same training process upon a smaller data set constructed in the same class-reduction process, but is instead an estimate based upon the model after training.

Appendix F

Accuracy limited to an N-class testing set

While multiclass AUC is a good multiclass evaluation metric for future work, there are no works in this space that currently use it. In order to allow comparisons to be drawn from this work to previous work, we define accuracy limited to N-classes. This metric may be narrated as a prediction task in which there is a model and a set of N potential classifications, a subset of all the classifications the model could make. First, the model proposes its classification, and if the classification is outside this subset, the model is asked to provide its next best classification. This process only ends when the model gives a predicted classification within the set of potential classifications.

To derive this formula, consider the probability $P[\arg\max_{C\in\mathcal{C}'} f(a,C) = A]$ for a randomly selected subclassification $\mathcal{C}' \subset \mathcal{C}$, $|\mathcal{C}'| = N < |\mathcal{C}|$, that sample a known to be from class A is predicted correctly. Explained simply, for a sample to be correctly classified, the random selection of classes within this set of N potential classifications must avoid all classes that would trip up the prediction for a given sample a whose true class is A. The number of these 'error classes' is $N_{error} = \sum_{C\in\mathcal{C}} \mathbf{1}[f(a,C) > f(a,A)]$. The general expression for a sample a to be correctly classified in an N-class testing set is a simple combinatorics expression:

$$P[\arg\max_{C\in\mathcal{C}'} f(a,C) = A] = \frac{\binom{|\mathcal{C}|-N_{error}}{N}}{\binom{|\mathcal{C}|}{N}}$$

Then, by linearity of expectation, the accuracy for the whole model across all selections of $\mathcal{C}' \subset \mathcal{C}$ is equal to the mean of each sample's accuracy, and so the end result is simply the mean of the expression above across all sessions.

Bibliography

[1] Nadine Aburumman, Marco Gillies, Jamie A. Ward, and Antonia F.de C. Hamilton. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. 164:102819.

[2] Ashwin Ajit, Natasha Kholgade Banerjee, and Sean Banerjee. Combining pairwise feature matches from device trajectories for biometric authentication in virtual reality environments. *Proceedings - 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2019*, pages 9–16, 2019.

[3] Jamie S. Allsop, Tomas Vaitkus, Dannette Marie, and Lynden K. Miles. Coordination and collective performance: Cooperative goals boost interpersonal synchrony and task outcomes. 7:1–11.

[4] Camila Alviar, Rick Dale, and Christopher Kello. The fractal structure of extended communicative performance.

[5] Michael Argyle and Janet Dean. Eye-Contact , Distance and Affiliation. *Sociometry*, 28(3):289–304, 1963.

[6] Julia Ayache, Andy Connor, Stefan Marks, Daria J. Kuss, Darren Rhodes, Alexander Sumich, and Nadja Heym. Exploring the "dark matter" of social interaction: Systematic review of a decade of research in spontaneous interpersonal coordination. 12:718237.

[7] J. N. Bailenson and N. Yee. A Longitudinal Study of Task Performance, Head Movements, Subjective Report, Simulator Sickness, and Transformed Social Interaction in Collaborative Virtual Environments. 15(6):699–716, 2006.

[8] Jeremy Bailenson. Protecting Nonverbal Data Tracked in Virtual Reality. *JAMA Pediatrics*, 172(10):905–906, 2018.

[9] Jeremy N. Bailenson, Andrew C. Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators and Virtual Environments*, 13(4):428–441, 2004.

[10] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 10(6):583–598, 2001.

[11] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall, and Jack M. Loomis. Interpersonal Distances in Virtual Environments. *Personality and Social Psychology Bulletin*, 29(7):819–833, 2003.

[12] Jeremy N. Bailenson and Nick Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. 16(10):814–819.

[13] Jeremy N. Bailenson and Nick Yee. Virtual interpersonal touch and digital chameleons. 31(4):225–242.

[14] Jeremy N. Bailenson, Nick Yee, Kayur Patel, and Andrew C. Beall. Detecting digital chameleons. 24(1):66–87.

[15] Catherine Mj Beaulieu. Intercultural study of personal space: A case study. *Journal of Applied Social Psychology*, 34(4):794–805, 2004.

[16] Frank J. Bernieri and Robert Rosenthal. Interpersonal coordination: Behavior matching and interactional synchrony. In *Fundamentals of nonverbal behavior*, pages 401–431.

[17] Jim Blascovich, Jack Loomis, Andrew C. Beall, Kimberly R. Swinth, Crystal L. Hoyt, and Jeremy N. Bailenson. Immersive virtual environment technology as a methodological tool for social psychology jim. 13(2):103–124.

[18] Steven M. Boker, Jennifer L. Rotondo, Minquan Xu, and Kadijah King. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. 7(3):338–355. ISBN: 9780805862805.

[19] Andrea Bonsch, Sina Radke, Heiko Overath, Laura M. Asche, Jonathan Wendt, Tom Vierjahn, Ute Habel, and Torsten W. Kuhlen. Social VR: How Personal Space is Affected by Virtual Agents' Emotions. *25th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2018 - Proceedings*, pages 199–206, 2018.

[20] Susan E Brennan, Joy E Hanna, Gregory J. Zelinsky, and Kelly J. Savietta. Eye gaze cues for coordination in collaborative tasks. *DUET Workshop, CSCW'12, Seattle, Washington, USA*, pages 11–15, 2012.

[21] Simone Browne. *Dark matters: On the surveillance of blackness*. Duke University Press, 2015.

[22] Judee K. Burgoon. Expectancy violations theory. In *The International Encyclopedia of Interpersonal Communication*, pages 1–9. John Wiley & Sons, Ltd. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118540190.wbeic102.

[23] Thomas P. Caudell and David W. Mizell. Augmented reality: An application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*.

[24] Tanya L. Chartrand and John A. Bargh. The chameleon effect: The perception–behavior link and social interaction. 76(6):893. Publisher: US: American Psychological Association.

[25] Zubin Choudhary, Matthew Gottsacker, Kangsoo Kim, Ryan Schubert, Jeanine Stefanucci, Gerd Bruder, and Gregory F. Welch. Revisiting distance perception with scaled embodied cues in social virtual reality. *Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2021*, pages 788–797, 2021.

[26] W. S. Condon and W. D. Ogston. Sound film analysis of normal and pathological behavior patterns. ISBN: 0022-3018 ISSN: 0022-3018 Pages: 338–347 Publication Title: The Journal of Nervous and Mental Disease Volume: 143.

[27] Carolina Cruz-Neira, Daniel J. Sandin, and Thomas A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the CAVE. pages 135–142. ISBN: 0897916018.

[28] James J. Cummings and Jeremy N. Bailenson. How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. 19(2):272–309. ISBN: 10.1080/15213269.2015.1015740.

[29] Brendan David-John, Kevin Butler, and Eakta Jain. Privacy-preserving datasets of eye-tracking samples with applications in XR. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2774–2784, 2023. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[30] Marco Del Giudice and Steven W. Gangestad. A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. 4(1):2515245920954925. Publisher: SAGE Publications Inc.

[31] Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines.

[32] Roberto Di Pietro and Stefano Cresci. Metaverse: Security and Privacy Issues. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 281–288, December 2021.

[33] Sidney D'Mello, Rick Dale, and Art Graesser. Disequilibrium in the mind, disharmony in the body. 26(2):362–374.

[34] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3):211–407, 2013.

[35] Peter Eckersley. How unique is your web browser? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6205 LNCS:1–18, 2010.

[36] Ben Falchuk, Shoshana Loeb, and Ralph Neff. The Social Metaverse: Battle for Privacy. *IEEE Technology and Society Magazine*, 37(2):52–61, June 2018. Conference Name: IEEE Technology and Society Magazine.

[37] Brandon Falk, Yan Meng, Yuxia Zhan, and Haojin Zhu. POSTER: ReAvatar: Virtual Reality De-anonymization Attack through Correlating Movement Signatures. *Proceedings of the ACM Conference on Computer and Communications Security*, pages 2405–2407, 2021.

[38] K. Fujiwara and K. Yokomitsu. Video-based tracking approach for nonverbal synchrony: A comparison of motion energy analysis and OpenPose. 53(6):2700–2711.

[39] Ken Fujiwara. Triadic synchrony: Application of multiple wavelet coherence to a small group conversation. 07(14):1477–1483.

[40] Ken Fujiwara and Ikuo Daibo. Evaluating interpersonal synchrony: Wavelet transform toward an unstructured conversation. 7.

[41] Ken Fujiwara, Kunihiko Nomura, and Miki Eto. Antiphase synchrony increases perceived entitativity and uniqueness: A joint hand-clapping task. 14.

[42] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. *Conference on Human Factors in Computing Systems - Proceedings*, (5):529–536, 2003.

[43] Guy Gaziv, Lior Noy, Yuvalal Liron, and Uri Alon. A reduced-dimensionality approach to uncovering dyadic modes of body motion in conversations. 12(1). ISBN: 1111111111.

[44] K.-I Goh, A.-L Barabási, and Barab´ Barabási. Burstiness and memory in complex systems.

[45] K Grammer, M Honda, A Juette, and A Schmitt. Fuzziness of nonverbal courtship communication unblurred by motion energy detection. 77(3):487–508. ISBN: 0022-3514 (Print)\r0022-3514 (Linking).

[46] A. Grinsted, J. C. Moore, and S. Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. 11(5):561–566. Publisher: Copernicus GmbH.

[47] Ihshan Gumilar, Ekansh Sareen, Reed Bell, Augustus Stone, Ashkan Hayati, Jingwen Mao, Amit Barde, and Anubha Gupta. A comparative study on inter-brain synchrony in real and virtual environments using hyperscanning. Publisher: Elsevier Ltd.

[48] Joanna Hale, Jamie A Ward, Francesco Buccheri, Dominic Oliver, and Antonia F De C Hamilton. Are you on my wavelength ? interpersonal coordination in dyadic conversations. (123456789). ISBN: 0123456789 Publisher: Springer US.

[49] Edward Hall. *The Hidden Dimension: Man's Use of Space in Public and Private*. Anchor Books, Hamburg, Germany, 1969.

[50] Eugy Han, Mark R Miller, Cyan DeVeaux, Hanseul Jun, Kristine L Nowak, Jeffrey T Hancock, Nilam Ram, and Jeremy N Bailenson. People, places, and time: a large-scale, longitudinal study of transformed avatars and environmental context in group interaction in the metaverse. *Journal of Computer-Mediated Communication*, 28(2), 01 2023. zmac031.

[51] Eugy Han, Mark Roman Miller, Nilam Ram, Kristine L. Nowak, and Jeremy N. Bailenson. Understanding group behavior in virtual reality: A large-scale, longitudinal study in the metaverse.

[52] Ming Han, Xue Min Wang, and Shu Guang Kuai. Social rather than physical crowding reduces the required interpersonal distance in virtual environments. *PsyCh Journal*, (February):1–10, 2022.

[53] David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2):171–186, 2001.

[54] Leslie A Hayduk. Personal Space: Where We Now Stand. *Psychol. Bull.*, 94(2):293, 1983.

[55] Brittan Heller. Watching androids dream of electric sheep: Immersive technology, biometric psychography, and the law. *Vanderbilt Journal of Entertainment and Technology Law*, 23, 2020.

[56] Diane Hosfelt. Making ethical decisions for the immersive web, 2019.

[57] Markus Jakobsson, Elaine Shi, Philippe Golle, Richard Chow, et al. Implicit authentication for mobile devices. In *Proceedings of the 4th USENIX conference on Hot topics in security*, volume 1, pages 25–27. USENIX Association, 2009.

[58] Eunice Jun. Circadian rhythms and physiological synchrony : Evidence of the impact of diversity on small group creativity. 3.

[59] Kevin Kelly, Adam Heilbrun, and Barbara Stacks. Virtual reality: an interview with jaron lanier. 64:108–120.

[60] Negar Khojasteh and Andrea Stevenson Won. Working Together on Diverse Tasks: A Longitudinal Study on Individual Workload, Presence and Emotional Recognition in Collaborative Virtual Environments. *Frontiers in Virtual Reality*, 2(June):1–24, 2021.

[61] Daiki Kodama, Takato Mizuho, Yuji Hatada, Takuji Narumi, and Michitaka Hirose. Effects of collaborative training using virtual co-embodiment on motor skill learning. 29(5):2304–2314. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[62] Alex Kupin, Benjamin Moeller, Yijun Jiang, Natasha Kholgade Banerjee, and Sean Banerjee. Task-Driven Biometric Authentication of Users in Virtual Reality (VR) Environments. In *Internation Conference on Multimedia Modeling2*, pages 55–67, 2019.

[63] Marianne LaFrance. Nonverbal synchrony and rapport: Analysis by the cross-lag panel technique. 42(1):66–70. Publisher: [Sage Publications, Inc., American Sociological Association].

[64] Marc Erich Latoschik, Florian Kern, Jan Philipp Stauffert, Andrea Bartl, Mario Botsch, and Jean Luc Lugrin. Not alone here?! scalability and user experience of embodied ambient crowds in distributed social virtual reality. 25(5):2134–2144.

[65] K. M. Lee. Presence, explicated. 14(1):27–50.

[66] Sugang Li, Ashwin Ashok, Yanyong Zhang, Chenren Xu, Janne Lindqvist, and Macro Gruteser. Whose move is it anyway? Authenticating smart wearable devices using unique head movement patterns. *2016 IEEE International Conference on Pervasive Computing and Communications, PerCom 2016*, pages 1–9, 2016.

[67] Jonathan Liebers, Mark Abdelaziz, Lukas Mecke, Alia Saad, Jonas Auda, Uwe Grunefeld, Florian Alt, and Stefan Schneegass. Understanding user identification in virtual reality through behavioral biometrics and the efect of body normalization. *Conference on Human Factors in Computing Systems - Proceedings*, 2021. ISBN: 9781450380966.

[68] Brendan McSweeney. Hofstede's model of national cultural differences and their consequences: A triumph of faith - A failure of analysis. *Human Relations*, 55(1):89–118, 2002.

[69] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports*, 10(1):17404–17413, 2020.

[70] Mark Roman Miller, Hanseul Jun, and Jeremy N. Bailenson. Motion and Meaning: Sample-Level Nonlinear Analyses of Virtual Reality Tracking Data. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 147–152, Bari, Italy, 2021. IEEE.

[71] Mark Roman Miller, Neeraj Sonalkar, Ade Mabogunje, Larry Leifer, and Jeremy Bailenson. Synchrony within triads using virtual reality. 5. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[72] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. Within-system and cross-system behavior-based biometric authentication in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 311–316. IEEE, 2020.

[73] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. Using siamese neural networks to perform cross-system behavioral authentication in virtual reality. *Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2021*, pages 140–149, 2021.

[74] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. Combining Real-World Constraints on User Behavior with Deep Neural Networks for Virtual Reality (VR) Biometrics. *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2022*, pages 409–418, 2022.

[75] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. Temporal Effects in Motion Behavior for Virtual Reality (VR) Biometrics. *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2022*, pages 563–572, 2022.

[76] Alec G. Moore, Ryan P. McMahan, Hailiang Dong, and Nicholas Ruozzi. Personal identifiability and obfuscation of user tracking data from VR training sessions. *Proceedings - 2021 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2021*, pages 221–228, 2021.

[77] Fares Moustafa and Anthony Steed. A longitudinal study of small group interaction in social virtual reality. *Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST*, 2018.

[78] Joschka Mütterlein, Sebastian Jelsch, and Thomas Hess. Specifics of collaboration in virtual reality: How immersion drives the intention to collaborate. *Proceedings of the 22nd Pacific Asia Conference on Information Systems - Opportunities and Challenges for the Digitized Society: Are We Ready?, PACIS 2018*, 2018.

[79] Vivek Nair, Gonzalo Munilla Garrido, and Dawn Song. Exploring the Unprecedented Privacy Risks of the Metaverse. *ArXiv*, 2022.

[80] Vivek Nair, Gonzalo Munilla Garrido, and Dawn Song. Going Incognito in the Metaverse. *ArXiv*, 2022.

[81] Allen Newell. *Unified theories of cognition.* Unified theories of cognition. Harvard University Press. Pages: xvii, 549.

[82] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632. ACM Press, 2005.

[83] Eric Novotny and Gary Bente. Identifying signatures of perceived interpersonal synchrony. 46(4):485–517.

[84] Ilesanmi Olade, Charles Fleming, and Hai Ning Liang. Biomove: Biometric user identification from human kinesiological movements for virtual reality systems. *Sensors (Switzerland)*, 20(10):1–19, 2020.

[85] Aydin Ozdemir. Shopping malls: Measuring interpersonal distance under changing conditions and across cultures. *Field Methods*, 20(3):226–248, 2008.

[86] Ye Pan and Anthony Steed. Effects of 3d perspective on head gaze estimation with a multiview autostereoscopic display. 86:138–148.

[87] Ye Pan and Anthony Steed. A gaze-preserving situated multiview telepresence system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2173–2176. ACM.

[88] Ken Pfeuffer, Matthias J Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 110:1—-110:12, New York, NY, USA, 2019. ACM.

[89] Alexander Plopski, Teresa Hirzle, Nahal Norouzi, Long Qian, Gerd Bruder, and Tobias Langlotz. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *ACM Computing Surveys*, 55(3):1–39, 2023.

[90] Erik Prytz, Susanna Nilsson, and Arne Jönsson. The importance of eye-contact for collaboration in AR system. *9th IEEE International Symposium on Mixed and Augmented Reality 2010: Science and Technology, ISMAR 2010 - Proceedings*, pages 119–126, 2010.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[92] Nilam Ram and Manfred Diehl. Multiple-time-scale design and analysis: Pushing toward real-time modeling of complex developmental processes. In *Handbook of Intraindividual variability across the life span*, pages 308–323. Journal Abbreviation: Handbook of Intraindividual variability across the life span.

[93] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy: coordinated body movement reflects relationship quality and outcome. 79(3):284–295. ISBN: 1939-2117 (Electronic)\r0022-006X (Linking).

[94] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98*, pages 179–188. ACM Press.

[95] Byron Reeves, Leo Yeykelis, and James J. Cummings. The Use of Media in Media Psychology. *Media Psychology*, 19(1):49–71, 2016.

[96] Franziska Roesner, Taday Oshi Kohno, and David Molnar. Security and privacy for augmented reality systems. *Communications of the ACM*, 57(4):88–96, 2014.

[97] Daniel Roth, Constantin Klelnbeck, Tobias Feigl, Christopher Mutschler, and Marc Erich Latoschik. Beyond Replication: Augmenting Social Behaviors in Multi-User Virtual Realities. *25th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2018 - Proceedings*, pages 215–222, 2018.

[98] Daniel Roth, Jean Luc Lugrin, Dmitri Galakhov, Arvid Hofmann, Gary Bente, Marc Erich Latoschik, and Arnulph Fuhrmann. Avatar realism and social interaction quality in virtual reality. *Proceedings - IEEE Virtual Reality*, 2016-July:277–278, 2016.

[99] Robert J. Rydell and Allen R. McConnell. Perceptions of entitativity and attitude change. 31(1):99–110.

[100] Mohd Sabra, Nisha Vinayaga Sureshkanth, Ari Sharma, Anindya Maiti, and Murtuza Jadliwala. Exploiting out-of-band motion sensor data to de-anonymize virtual reality users, 2023.

[101] Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15. ACM.

[102] Christian Schell, Andreas Hotho, and Marc Erich Latoschik. Comparison of Data Encodings and Machine Learning Architectures for User Identification on Arbitrary Motion Sequences. In *2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 11–19, December 2022. ISSN: 2771-7453.

[103] Désirée Schoenherr, Jane Paulick, Susanne Worrack, Bernhard M. Strauss, Julian A. Rubel, Brian Schwartz, Anne Katharina Deisenhofer, Wolfgang Lutz, Ulrich Stangier, and Uwe Altmann. Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony. 51(1):361–383. Publisher: Behavior Research Methods.

[104] Yiran Shen, Hongkai Wen, Chengwen Luo, Weitao Xu, Tao Zhang, Wen Hu, and Daniela Rus. GaitLock: Protect Virtual and Augmented Reality Headsets Using Gait. *IEEE Transactions on Dependable and Secure Computing*, 5971(c):1–14, 2018.

[105] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. 4(11):1208–1214. Number: 11 Publisher: Nature Publishing Group.

[106] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, and Gordon Wetzstein. Saliency in VR: How do people explore virtual environments? _eprint: 1612.04335.

[107] Mel Slater, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, Dalila Szostak, Shivashankar Halan, Deborah Fox, and Jeremy Silver. The Ethics of Realism in Virtual and Augmented Reality. *Frontiers in Virtual Reality*, 1(March):1–13, 2020.

[108] Mel Slater, Amela Sadagic, Martin Usoh, and Ralph Schroeder. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, 9(1):37–51, 2000.

[109] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. 11(5):702–712.

[110] Yilu Sun, Omar Shaikh, and Andrea Stevenson Won. Nonverbal synchrony in virtual reality. pages 1–28. ISBN: 1111111111.

[111] Ivan E. Sutherland. A head-mounted three dimensional display. page 757. ISBN: 158113052X.

[112] Ivan E. Sutherland. Sketchpad: a man-machine graphical communication system. In *Proceedings of the May 21-23, 1963, spring joint computer conference on - AFIPS '63 (Spring)*, page 329. ACM Press.

[113] Ivan E. Sutherland. The ultimate display. pages 506–508. ISBN: 0897916670 _eprint: 1601.03459.

[114] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[115] Kota Takahashi, Yasuyuki Inoue, and Michiteru Kitazaki. Interpersonal Distance to a Speaking Avatar: Loudness Matters Irrespective of Contents. *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VRW 2022*, pages 774–775, 2022.

[116] Ron Tamborini, Eric Novotny, Sujay Prabhu, Matthias Hofer, Clare Grall, Brian Klebig, Lindsay S. Hahn, Janine Slaker, Rabindra A. Ratan, and Gary Bente. The effect of behavioral synchrony with black or white virtual agents on outgroup trust. 83:176–183.

[117] B. Tarr, M. Slater, and E. Cohen. Synchrony and social connection in immersive virtual reality. 8(1):1–8. Publisher: Springer US.

[118] R. Vertegaal, R. Slagter, G. Van Der Veer, and A. Nijholt. Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. *Conference on Human Factors in Computing Systems - Proceedings*, pages 301–308, 2001.

[119] Roel Vertegaal and Yaping Ding. Explaining effects of eye gaze on mediated group conversations: Amount or synchronization? *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 41–48, 2002.

[120] Janna N. Vrijsen, Wolf-Gero Lange, Ron Dotsch, Daniël H. J. Wigboldus, and Mike Rinck. How do socially anxious women evaluate mimicry? a virtual reality study. 24(5):840–847.

[121] Takumi Wakabayashi, Yukihiko Okada, and Keiichi Zempo. Effect of salesperson avatar automatically mimicking customer's nodding on the enjoyment of conversation in virtual environments. In *Augmented Humans Conference*, pages 334–337. ACM.

[122] Joseph B Walther. Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, 23(1):3–43, 1996.

[123] Changsheng Wan, Li Wang, and Vir V. Phoha. A Survey on Gait Recognition. *ACM Computing Surveys*, 51(5):1–35, September 2019.

[124] Xue Wang and Yang Zhang. Nod to Auth: Fluent AR/VR Authentication with User Head-Neck Modeling. *Conference on Human Factors in Computing Systems - Proceedings*, 2021.

[125] Yuntao Wang, Zhou Su, Ning Zhang, Rui Xing, Dongxiao Liu, Tom H. Luan, and Xuemin Shen. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Communications Surveys & Tutorials*, 25(1):319–352, 2023.

[126] Mark Weiser. The computer for the 21st century. Issue: 3 Pages: 94–104 Publication Title: Scientific American Volume: 265.

[127] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[128] Julie R. Williamson, Jie Li, and Vinoba Vinayagamoorthy. Proxemics and social interactions in an instrumented virtual realityworkshop. *Conference on Human Factors in Computing Systems - Proceedings*, 2021.

[129] Andrea Stevenson Won, Jeremy N. Bailenson, Suzanne C. Stathatos, and Wenqing Dai. Automatically detected nonverbal behavior predicts creativity in collaborating dyads. 38(3):389–408.

[130] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[131] Haley E. Yaremych, William D. Kistler, Niraj Trivedi, and Susan Persky. Path tortuosity in virtual reality: A novel approach for quantifying behavioral process in a food choice context. 22(7):486–493.

[132] Haley E. Yaremych and S. Persky. Tracing physical behavior in virtual reality: A narrative review of applications to social psychology. *Journal of Experimental Social Psychology*, 85(April):103845, 2019.

[133] Nick Yee, Jeremy N. Bailenson, Mark Urbanek, Francis Chang, and Dan Merget. The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *Cyberpsychology and Behavior*, 10(1):115–121, 2007.

[134] Pavel Zahorik and Rick L. Jenison. Presence as being-in-the-world. 7(1):78–89.

[135] Leshao Zhang and Patrick G.T. Healey. Human, chameleon or nodding dog? In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 428–436. ACM.