LONDON'S GLOBAL UNIVERSITY

Measuring, Analysing and Artificially Generating Head Nodding Signals in Dyadic Social Interaction

Patrick Lord Falk

Thesis submitted to UCL for the degree of Doctor of Philosophy

August 2022

I, Patrick Lord Falk confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Date:

Abstract

Social interaction involves rich and complex behaviours where verbal and non-verbal signals are exchanged in dynamic patterns. The aim of this thesis is to explore new ways of *measuring* and *analysing* interpersonal coordination as it naturally occurs in social interactions. Specifically, we want to understand what different types of head nods mean in different social contexts, how they are used during face-to-face dyadic conversation, and if they relate to memory and learning. Many current methods are limited by time-consuming and low-resolution data, which cannot capture the full richness of a dyadic social interaction. This thesis explores ways to demonstrate how high-resolution data in this area can give new insights into the study of social interaction. Furthermore, we also want to demonstrate the benefit of using virtual reality to *artificially generate* interpersonal coordination to test our hypotheses about the meaning of head nodding as a communicative signal.

The first study aims to capture two patterns of head nodding signals – fast nods and slow nods – and determine what they mean and how they are used across different conversational contexts. We find that fast nodding signals receiving new information and has a different meaning than slow nods. The second study aims to investigate a link between memory and head nodding behaviour. This exploratory study provided initial hints that there might be a relationship, though further analyses were less clear. In the third study, we aim to test if interactive head nodding in virtual agents can be used to measure how much we like the virtual agent, and whether we learn better from virtual agents that we like. We find no causal link between memory performance and interactivity. In the fourth study, we perform a cross-experimental analysis of how the level of interactivity in different contexts (i.e., real, virtual, and video), impacts on memory and find clear differences between them.

Impact Statement

The overall objective of this thesis is to understand how people coordinate and learn using head nods by establishing new ways to *measure*, *analyse* and *artificially generate* dyadic social interactions. We establish a data collection framework that is able to *measure* the precise multimodal nature of dyadic verbal and non-verbal behaviour using motion capture, and to *analyze* the coordination between people in a conversation using wavelet analysis. This could be very important for determining naturalistic parameters in social interaction, such as its timing or rhythmic properties. Understanding the timing and at which frequencies specific behaviours occur can help us answer how and why we use these as social communicative signals. Such knowledge could impact research on disorders of the social brain such as autism, and automatic detection or sensing of social signals.

The quantification of these social signals further demonstrates the experimental benefits of *artificially generating* social signals in virtual agents based on behavioural data from real interactions. This can then be used to further test and challenge our cognitive hypotheses about social behaviour to provide new insights and directions for research into human interpersonal coordination. Using motion capture to build better virtual models of interpersonal coordination with a grounding in psychology can subsequently contribute to improved realism of virtual agents that closely approximate real behaviour without the need to manually extract and code individual parameters. This will guide better research on virtual reality and provide fundamental new insights and directions for reasearch into interpersonal coordination.

The findings in this thesis also demonstrate that it is possible to identify specific head nodding behaviours and link these behaviours to conversational outcomes.

These head nodding behaviours can also be used as a way to quantify features of an interaction (e.g., affiliation, liking, interest), or as part of a clinical assessment.

Understanding how different conversational outcomes like memory and learning are affected by social interaction is important in education. In this new era of online social interaction following the recent coronavirus (COVID-19) pandemic, it is important to understand what makes an interaction work and how we can best implement the benefits of real social interactions in an online educational setting.

The convergence of research questions in both psychology and computing thus sets the scene for the studies presented in this thesis, which draws together these diverse research areas, combining cognitive and psychological hypothesis testing with new advances in *measuring*, *analysing*, and *artificially generating* interpersonal coordination to provide a new level of understanding of dyadic social interaction.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Antonia Hamilton for her invaluable support, guidance, and patience during my PhD. Antonia has been diligent in teaching me new ideas and techniques, and I am especially grateful to her for constantly getting me out of my comfort zone to grow as an academic.

I offer my sincere appreciation for the learning opportunities provided by the Leverhulme Trust for funding my PhD. It is next to impossible to alone command more than a portion of multi-disciplinary research, so I want to extend my special thanks to Jamie A. Ward whom I worked alongside during this time and who I owe greatly for his stout technical insights and for giving me much encouragement along the way. I also want to express a special thank you to Marco Gillies and Nadine Aburumman for their contributions and resourceful insights to this research.

It has been a great pleasure and privilege to work among so many brilliant academics from the Social Neuroscience group and the Institute of Cognitive Neuroscience over the years. I extend my heartfelt thanks to everyone for sharing their ideas and creating a supportive workplace. I especially thank Roser Cañigueral and Ye Tian for their help and friendship during the time shared in our office.

Thank you to my dear friend and freelance illustrator Mattias Fahlberg who helped illustrate some of the pictures in this thesis. Finally, I am endlessly thankful to my family, my wife Linnéa and our two dogs Tito and Cash for their love, understanding, and continued support to complete this research work.

Contributions

Patrick Falk carried out data collection for all experiments in this thesis, with help from masters student Jessica Kankram when collecting the video data for the cross-experimental study in Chapter 5. Patrick Falk designed the experiments and performed all analyses in MATLAB with help from Antonia Hamilton, Jamie A. Ward and Marco Gillies. The virtual task in Chapter 4 was created by Nadine Aburumman.

Table of Contents

Table of Figures

Table of Tables

Chapter 1. Introduction

Human social interaction involves rich and complex behaviours where verbal and non-verbal signals are exchanged in dynamic patterns (Paxton & Dale, 2013). In this thesis we are interested in learning more about head nodding during face-to-face dyadic conversations in naturalistic settings. Specifically, we want to understand what different types of head nods mean during different social contexts, how they are used, and if they relate to learning and memory. To understand social coordination, it is helpful to have detailed high-resolution data recordings and to analyse this with appropriate methods. However, many current methodological frameworks are limited by time-consuming and low-resolution data collection methods, which cannot capture the full richness of a dyadic social interaction. By using a multimodal data collection framework of dyadic face-to-face conversation, we also want to explore new ways of modelling coordinated behaviour as it naturally occurs in social interactions.

We begin this chapter by reviewing previous work on interpersonal coordination (Section 1.1). In this section we cover verbal and non-verbal social signals (1.1.1) and give an overview of two types of coordinated behaviours: Behavioural Mimicry (1.1.2) and Backchannel Signals (1.1.3). We will then focus specifically on head nodding behaviour (Section 1.2). In this section we look at head nods as mimicry (1.2.1), head nods as backchannels (1.2.2), and introduce approaches to the study of interpersonal coordination (1.2.3). During the next three sections, we review the methodological background to *Measuring* (Section 1.3), *Analysing* (Section 1.4), and *artificially generating* (Section 1.5) interpersonal coordination. These sections include our approach to recording dyads (1.3.1), how to analyse this data with time-frequency analysis (1.4.1–1.4.2), and to artificially generate social signals in interactive virtual agents (1.5.1–1.5.2). We conclude this chapter by presenting our overall aims, together with an overview of the experimental chapters (Section 1.6).

1.1 Interpersonal Coordination

Face-to-face conversations are undoubtedly our most important form of social interactions. Understanding the social interplay during conversations, how they work, and why they are important, remains a challenge for researchers studying their complex nature from different fields of interest, like psychology, cognitive neuroscience, linguistics, or computer science. The study of how people perform this behavioural and linguistic coordination during social interaction is essential to understanding the nature of social cognition and has attracted the attention of many social disciplines recently. Various terminology has been used to define interpersonal coordination, and a general definition is that interpersonal coordination is the temporal matching of body movements and/or linguistic utterances between people when they engage in social interaction (Bernieri & Rosenthal, 1991; Hoehl, Fairhurst, & Schirmer, 2020). In this thesis, we will use the umbrella term 'coordination' when we talk about ways in which people temporally organise their verbal and non-verbal behaviour in social interactions.

From early work (Chartrand & Bargh, 1999; Condon & Ogston, 1966; Kendon, 1970) to more recent studies (Abney, Paxton, Dale, & Kello, 2015; Ramseyer & Tschacher, 2011) it has been demonstrated that people coordinate their movements in different ways, including synchronizing rhythms (Richardson & Dale, 2005), mimicking (Chartrand & Bargh, 1999), structuring turn-taking behaviour (Duncan & Fiske, 1977), and assuming complementary roles (Garrod & Pickering, 2004).

The outcomes of interpersonal coordination have been widely studied in a variety of contexts. Many studies examine the positive outcomes, either at an individual level, or at the level of the dyad or group. Perhaps the most well-known finding related to positive outcomes of coordination is that it promotes liking and feelings of

closeness between interacting partners (Lakin, Jefferis, Cheng & Chartrand, 2003). Other interpersonal outcomes include findings that coordination increases rapport (Hove & Risen, 2009), and cooperation (Wilmermuth & Heath, 2009).

However, there are many challenges involved in understanding the functional benefits of interpersonal coordination as the research community is still uncertain about what mechanisms are involved and how they interact across multiple modalities and timescales (Dale, Fusaroli, Duran & Richardson, 2013). There is a growing recognition that studying isolated participants responding to stimuli or interacting with confederates does not translate well to real-world social interaction, which is often complex and dynamic (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Risko, Richardson, & Kingstone, 2016). Recent attempts have been made to fill this gap, calling for a 'second-person neuroscience' capturing the need for natural interactions, involving multiple modalities across different contexts and timescales (Heerey, 2015; Schilbach et al., 2013).

In this thesis, the aim is to explore new ways of 1) *Measuring* real-world interpersonal coordination using high resolution motion capture to identify fine grained behaviours based on hypotheses about the meaning of specific social signals; 2) *Analysing* interpersonal coordination as it naturally occurs in dyadic social interaction; 3) *Artificially Generating* interpersonal coordination in virtual characters that can interact with participants to test our hypotheses about the meaning of specific behaviours as communicative signals. These advances can allow us to investigate how to quantify simple interaction behaviour in higher resolution and test hypotheses with high experimental control.

In the next section we will review research on how two or more people coordinate exchanges of both verbal and non-verbal social signals.

1.1.1 Verbal and Non-Verbal Social Signals

Take a moment to consider an example of a complex social interaction – two poker players trying to outwit each other at the poker table. One of the players is confident he has the winning hand and is trying to hide his confidence to lure the other player in and bet the pot. As he does this, he is acting stoic not to give anything away that the other player can use to "read" him. In poker slang this is known as not giving away a "tell". However, his opponent knows this, and is trying everything he can to "talk up" the situation by verbally asking questions that will elicit a response that he can read and gather clues on how confident he is. The player with the winning hand is experienced enough to recognize that this is what the other player is doing, so he does not answer. At this point, the other player, realizing talking does not work, slams and rubs his hands together, leans back in his chair more relaxed, and laughs at the situation while saying "I don't know what to do?!". Now, the player with the winning hand does the same, he leans back laughing in his chair while soon realizing his own mistake. By responding in such a relaxed and confident manner, he just gave away multiple "tells" to the other player, which was exactly what the other player wanted, to coerce a non-verbal response. There is a lot to gather from this example on how people coordinate their verbal and non-verbal behaviour.

First, a social interaction is a shared behaviour between two or more individuals, resulting in a transfer of information through verbal and non-verbal messages. In this thesis, we will be discussing and referring to these messages as *social signals*. We define a social signal as a detectable and interpretable message produced, voluntary or involuntary, during social interactions that provide an exchange of information between people (Maynard-Smith & Harper, 2003). As such, a social signal is profitable to both the sender and receiver (Kleinsmith & Bianchi-Berthouze, 2013).

Secondly, interpersonal coordination is reciprocal, meaning that both players trade behaviours back and forth in a dynamic “dance” (Kendon, 1970). Traditionally, the role of the person that is receiving a signal has been viewed from two main perspectives in the language sciences. In what Clark and Krych (2004) called ‘unilateral’ views on conversation, speaking and listening are two separate processes. In this view, the receiver is a passive listener. In ‘bilateral’ views on conversation, speaking and listening is considered a joint activity where the receiver of a signal is not merely a passive observer but has to adapt and coordinate his behaviour with that of the speaker to maintain mutual understanding (Clark, 1996; Sacks, Schegloff, & Jefferson, 1974). A simple and clear-cut difference between speaker and listener roles is rare in real-world social interactions. In research, we typically use this ‘sender-receiver’ framework because it acts as a good model of explanation. For example, we distinguish between the person (sender) who sends, speaks, enact, or otherwise reveals a signal; and another person (receiver) who decodes and interprets the signal. This distinction is heuristic in the sense that it often works as a good-enough labelling in a world where any social interaction quickly becomes highly dynamical and complex.

Any description of conversation also requires consideration of how people can build mutual understanding between two or more people and the key concept here is ‘common ground’ (Clark, 1996). Common ground is the set of beliefs and knowledge held by both people, which they both know that the other also believes. During a conversation, common ground can be updated as each person hears new information from their partner or signals to their partner that they understand what they have received. This process is called ‘grounding’ (Clark & Brennan, 1991).

Face-to-face conversations is a type of social interaction that involve highly coordinated exchanges of verbal and non-verbal signals. Experimental studies of conversation, or discourse, have primarily focused on verbal communication (Clark, 1996). Since the 1960s, several disciplines, both inside and outside of linguistics, have been interested in how language and body movements are interconnected within human action. This introduced new challenges for disciplines traditionally focused almost exclusively on language, but also fueled new contributions in the social sciences on human action and communication. Within perspectives inspired by social interaction, such as Conversation Analysis (CA) (Duncan & Fiske, 1977; Sacks et al., 1974), emerged a field called interactional linguistics (Ochs, Schegloff, & Thompson, 1996) where focus is on the systematic organization and diversity of action within various social settings. In other words, action is understood as being organized not as isolated events by individuals but within social interaction. This includes gestures, gaze, facial expressions, body postures, body movements, prosodic speech (e.g., rhythm, pitch) and grammar. CA offered a new approach to study syntax and semantics, by encouraging a view of different modalities as interconnected, and language as integrated within this framework as one among many other modalities. According to Sacks and Schegloff (2002), CA has always been interested in bodily behaviour during social interaction and how it is sequentially organized. Even though the field of linguistics recognized early the relevance of gestures, this multimodality opened new avenues for analysis.

It is now widely recognized that non-verbal communication is important for successful interaction. Humans can produce non-verbal social signals in a variety of modalities, ranging from gaze (e.g., direction, blinks, pupil dilation) (Argyle & Cook, 1976; Kendon, 1967), body-movements including gestures (Kendon, 2004), posture

(Condon & Ogston, 1971), and head movements (Cerrato, 2005), to facial speech expressions (Ekman & Friesen, 1972). As we saw with the example of the two poker players, both verbal and non-verbal social signals are interconnected. For example, listening to someone speaking can influence you with what words they speak, the acoustics of the sound produced when speaking those words, the visual signals of the mouth moving, the posture and movement of their body and head. Social interaction is thus naturally multimodal, with a large variety of possible signals that each modality can produce. Importantly, by modalities, we mean not just the more specific and traditional notion of a sensory modality, but also measures derived from the human system that can lead to coordinated behaviour, including many non-verbal modalities, or 'modes' of interacting, like for example head nods, facial expressions, gestures, and eye-blinks (Burgoon et al., 2002). For this thesis, we will limit our investigations to the domain of non-verbal social signals.

In the next section, we are going to focus more closely on two specific types of interpersonal coordination, known as Behavioural Mimicry (1.1.2) and Backchannel Signals (1.1.3), which will drive our hypotheses used in this thesis.

1.1.2 Behavioural Mimicry

Humans copy each other in a variety of ways during conversations, including facial expressions (Bavelas & Chovil, 1997), moods (Neumann & Strack, 2000), speech (Giles & Powesland, 1975), postures and gestures (Chartrand & Bargh, 1999; Shockley, Santana, & Fowler, 2003). The copying of mannerisms, posture, gestures, and body movements is commonly called 'behavioural mimicry' (Lakin & Chartrand, 2003), or the 'chameleon effect' (Chartrand & Bargh, 1999). It typically arises

spontaneously during social interaction. For example, one person touches their hair and the other does the same shortly afterwards without thinking about it. Thus, mimicry differs from more goal-directed imitation (Hale & Hamilton, 2016a) where one person might copy to learn a skill or achieve a particular goal (Bekkering & Prinz, 2002; Bekkering, Wohlschlager, & Gattis, 2000). The spontaneous nature of behavioural mimicry makes it more difficult to fake, which in turn makes it seem more honest and revealing about socially relevant information (Pentland, 2010).

Early research observed that mimicry during clinical therapy sessions (Scheflen, 1963) and classroom interactions (Bernieri, 1988) was correlated with reported affiliation, empathy, and rapport. In their pioneering study Chartrand and Bargh (1999) experimentally manipulated behavioural mimicry and demonstrated a causal effect from being mimicked to positive outcomes. The authors had confederates copy the naturally occurring posture, movements, and mannerisms of the participants while taking turns to describe various photographs to each other. At the end of the experiment, participants who were mimicked rated the confederate as significantly more likeable. Following this study, the confederate paradigm became widely used for studying behavioural mimicry (Stel, Rispens, Leliveld, & Lokhorst, 2011; Van Baaren, Holland, Kawakami, & van Knippenberg, 2004), and the link between mimicry and affiliation led to the theory that mimicry functions as a 'social glue' to facilitate liking and affiliation to help people bond with members of our social groups (Lakin et al., 2003). This 'Social Glue Hypothesis' states that a person's initial liking of someone leads the person to mimic their posture, movements, and mannerisms in conversations, which in turn leads to greater mutual liking between the interacting partners. Support for this theory comes from evidence that people mimic others more when interacting with in-group members (Bourgeois & Hess,

2008), or have a goal to affiliate (Cheng & Chartrand, 2003). Behavioural mimicry has also been related to improved rapport (Chartrand & Van Baaren, 2009), altruistic behaviour (Van Baaren, Holland, Kawakami, & van Knippenberg, 2004), trust (Bailenson & Yee, 2005), and better collaborative performance (Fusaroli & Tylén, 2012; Marsh, Richardson & Schmidt, 2009). Moreover, during mimicry, the boundary between self and other is thought to become blurred (Georgieff & Jeannerod, 1998), and research has shown that behavioural mimicry appears to influence or affect the self-construal of the person being mimicked (Ashton-James, van Baaren, Chartrand, Decety, & Karremans, 2007), which led the researchers to propose that an increase in self-other overlap, where people feel closer or more like the other, leads to more prosocial behaviour, and not just towards the person mimicking, but to others in general. Some studies investigate the idea that being mimicked increases self-other overlap (Hale & Hamilton, 2016a), while others have suggested that a greater self-other overlap can induce more mimicry behaviour (Maister & Tsakiris, 2016). Other findings consistent with these proposals is that being mimicked generally induces cognitive changes in feelings of interdependence (Stel et al., 2011). It has also been suggested that some people show more spontaneous mimicry than others, leading to more liking (Salazar-Kämpf, Liebermann, Kerschreiter, & Krause, 2017).

However, the cognitive processes underlying this link or what exactly sustains this "social glue" are not yet clear and heavily debated. Moreover, even the basic link from mimicry to liking has not been replicated consistently, and it is not known if individual differences in mimicry are robust across different contexts. In Chapter 2, we will examine if there are reliable individual differences in social signalling behaviour. While some previous studies imply that some people show more mimicry than others (Salazar-Kämpf et al., 2017), there is little data to quantify this. Our

large-scale data collection will provide an opportunity to explore individual differences. In Chapter 3, we aim to determine if coordination between two people in a conversation might be related to measurable outcomes of the conversation, including how much the two people relate to each other in terms of self-other processing. Research has shown that mimicry appears to affect the self-construal of the person being mimicked (Ashton-James et al., 2007), and that being mimicked increases self-other overlap (Hale & Hamilton, 2016a), but so far only subjective reports and low-resolution methods have been used to examine this link. With higher resolution data and behavioural measures, we aim to explore the link between mimicry and self-other overlap further. In Chapter 4, we will examine mimicry in relation to liking, and how virtual reality can be used to test this relationship. Virtual mimicry has been shown to have significant effects on people's perception and attitude toward them (Bailenson & Yee, 2005), and could be used to test competing hypotheses about how people respond to mimicry. We aim to improve upon this method and artificially generate mimicry in virtual characters to test if virtual characters that mimic participants are more likable than those who do not mimic.

In the next section, we will introduce another type of interpersonally coordinated behaviour that is important for social interactions, known as Backchannel Signals.

1.1.3 Backchannel Signals

Research over the past decades has highlighted behavioural coordination that help establish and maintain common ground. Backchannelling is one such coordinated behaviour, believed to smooth interaction, and maintain grounding (Clark, 1996).

Descriptions of conversation behaviour often distinguish between the primary channel of information, from the speaker to the listener, and a secondary channel of information, or backchannel, from listener to speaker. Argyle (2013) notes that backchannel responses have a powerful effect on speakers. They afford an opportunity to communicate messages to the speaker without disrupting the flow of their speech. The speaker does not perceive them as interruptions (Duncan & Fiske, 1977). Whenever people listen to someone speaking, they do not just passively receive what is being said, but they actively participate in the interaction providing information about how they feel and think of the speaker's message. Thus, backchannels support the bilateral account of conversation. For example, when speakers send a signal, the listener provide feedback through a backchannel, which in turn affect the speaker (Clark & Krych, 2004). Verbal backchannels are represented as linguistic verbal answers like "Yes", or vocalizations such as 'uh-uh' and 'mm-hm' comments (Sacks et al., 1974). Non-verbal backchannels are usually associated with gestures, smiling, gaze, blinks, facial expressions, and head movements (Argyle, 2013; Goodwin 1981; Heylen, 2006; McCarthy, 2003; Hömke, Holler, & Levinson, 2017). It has been estimated that up to 20% of all facial expressions that occur during a conversation are backchannel signals (Bavelas & Chovil, 1997), and that backchannels occur more than a thousand times in an hour of conversation (Gardner, 1998).

Much of the research with non-verbal backchannel signals has been dedicated to identifying the social meaning behind them. For example, Hömke et al. (2017), investigated blinking as one potential additional type of backchannel, hypothesizing that it may serve a social-communicative function. To examine this, they studied dyadic face-to-face conversations by measuring and analysing short and long blink

duration and frequencies in high resolution using motion capture. The authors found that long blinks from a listener were related to changes in the conversation topic from the speaker and suggested that the long blink might signal ‘please move to a new topic’. This suggests that even subtle non-verbal signals such as eye blinks may be perceived as meaningful by others. In a follow up study, Hömke, Holler, and Levinson (2018), developed a novel experimental paradigm using virtual reality technology to selectively manipulate blink duration in a virtual listener. The participants were asked to have a conversation with three different virtual characters and to respond to open questions (e.g., “How was your weekend, what did you do?”). While participants were answering, the virtual character produced different types of visual feedback, or backchannels, which was triggered secretly by a confederate via video link whenever they felt appropriate to signal understanding. In some trials, those triggers produced a short blink (~200 ms), and in other trials they produced a long blink (~600 ms). The aim of their study was to experimentally test the claim that long blinks signal ‘please change topic’ by observing how participants would react when the virtual characters gave a long blink. The results showed that participants spoke less following a long blink compared to a short blink, consistent with their earlier hypothesis. This approach of using correlational data from close observation of naturalistic behaviour to build a good hypothesis, followed by testing with strong experimental control using a virtual character was crucial to determining the meaning behind long blinking as a backchannel behaviour. In the following sections, we will narrow our focus to head nods, and use a similar approach to explore the meaning of these signals.

1.2 Interpersonal Coordination of Head Nods

The way people move their head when they speak and listen to each other is an interesting form of communication to study because of the diversity of meanings that could be present. Researchers usually study head movements from two different perspectives: their role in speech production, and their communicative functions (McClave, 2000). From the perspective of speech production, Hadar, Steiner, Grant, and Rose (1983) found that when speaking the head moved a lot compared to when someone was listening, and that there was a correlation between head movement and verbalisations. The authors also showed that head movements play an important role in putting emphasis on important parts of the speech utterance, and on improving the auditory speech perception. Other functions of head movements during speech production include self-affirmation and managing speech planning and hesitation about what to say (Boholm & Allwood, 2010).

On the other hand, looking at head movements from the perspective of communicative function can tell us something about the way the movement might be, or intended to be perceived by others. Listeners head nods have been of particular interest to many researchers. Linguists, for example, have long recognized that listener head nods can be non-verbal backchannels that provide feedback to a speaker in an interactive way (McClave, 2000). Head nods can also function as stress markers, and head turns can have a deictic function (Birdwhistell, 1970). Boholm and Allwood (2010) considered the functions of repeated head movement, such as head nods, jerks, shakes, and tilts with co-occurring speech, and found that their main function, especially nods and shakes, was to provide communicative feedback. The power of head and facial movements to control interpersonal coordination even in the absence of speech was described by Kendon (1990).

Following the work from Kendon, research show that many other important non-verbal signals during conversation are centred around the head. Smiles for example are frequently reciprocated behaviour where people respond to their partners' smiles with smiles (Heerey & Kring, 2007; Hess & Bourgeois, 2010). Failing to reciprocate smiles can even affect what others think of you as they may find it to be aversive behaviour. It is not surprising then that listeners' attention is drawn to the speaker's head and face during conversation (Argyle & Cook, 1976).

Within the general context of head movements (Hadar, Steiner, & Rose, 1985; Kendon, 2002; McClave, 2000; Heylen, 2006; Cerrato, 2005), head nodding has been studied because of its central role in face-to-face conversations as an important source of social information (Argyle & Trower, 1979; Birdwhistell, 1970). The frequency of head nodding in face-to-face interactions has been shown to reveal personal characteristics or predict outcomes. For example, job applicants that head nod more in employment interviews are perceived as more employable than applicants who do not (Gifford & Wilkinson, 1985; McGovern, Jones, & Morris, 1979). In most cases, people producing the head nods are not even aware of the social signal they send as they are often the result of implicit processes.

It is important to note that the way social signals work can be different depending on a lot of factors, which may or may not be determined by a combination of functions. Thus, social signals are ambiguous, as they may correspond to more than one meaning. Head nodding, for example, is regarded as a distinct social signal that is particularly sensitive to conversational demands and can convey several different meanings, such as confirmation, agreement, and approval (Poggi, Errioco & Vincze, 2010), to signalling attention and understanding (Hadar et al., 1983; Kendon, 2002), and requests for information and passing or claiming turns (Duncan, 1972).

For this thesis, we will limit our investigations to the domain of non-verbal behaviour, and more specifically to one subset of non-verbal behaviour, namely head nods. A head nod typically consists of multiple up and down movements of the head. Poggi et al. (2010) define a nod as a vertical head movement in which the head, after a slight tilt up, bends downward and then goes back to its starting point. This is the theoretical definition of a head nod we are using in this thesis. It should be kept in mind that non-verbal signals and head nods is just one window through which we observe and measure social interaction, but that these signals are measured within the context and experimental framework of multimodal social interaction, including verbal communication. In other words, our experimental setup is built to capture and record a large variety of modalities, but we choose to analyse only head nodding behaviour in this thesis. This choice to delimit our work will naturally lead to consequences for interpretation, which we will cover at the end of the thesis. In the following sections, we will explain head nodding in conversations within the framework of behavioural mimicry and backchannel signals in respect to speaker-listener roles.

1.2.1 Head Nods as Mimicry

A long tradition of research into human interpersonal coordination describes patterns of synchrony or mimicry as integral non-verbal behaviours in dyadic social interaction (Bernieri, Steven, & Rosenthal, 1988; van Baaren, Janssen, Chartrand, & Dijksterhuis, 2009). As noted by Ramseyer and Tschacher (2006), the distinction between ‘synchrony’ and ‘mimicry’ is important as these concepts have recently been regarded as two distinct phenomena that have occupied separate research

fields. However, both perspectives apply the same reasoning and theory. For a simple interaction between two people, we can imagine different mechanisms at play that match with different timing properties. Mimicry occurs when one person performs an action and another performs the action afterwards, which implies a time lag between the two actions. Time lags are typically more than 300 ms (the limit of reaction times) but could be up to 10 seconds (Chartrand & Bargh, 1999). Synchronous behaviour (e.g., rhythmic walking, musical coordination), on the other hand, are those behaviours that are matched in time and occur simultaneously with no time lag (e.g., 0 ms) (Bernieri & Rosenthal, 1991). As such, synchrony and mimicry behaviour can be observed simultaneously during social interactions, and much research has focused on comparing their different causes and effects (Chartrand & Lakin, 2013; Hove & Risen, 2009). In this thesis we will use the term 'mimicry' to refer to behavioural mimicry with a time lag of head nodding.

In face-to-face social interactions (e.g., interviews or conversations), early research observed mimicry behaviour during clinical therapy sessions, and made a link to affiliation, empathy, and rapport (e.g., Scheflen, 1963). However, these types of studies were often based on single cases and anecdotal accounts. As we discussed earlier, after Chartrand and Bargh (1999) published their work on experimentally manipulating mimicry (i.e., the chameleon effect), many researchers started using confederate paradigms to study mimicry (Stel et al., 2011; van Baaren et al., 2004), which led to the hypothesis that behavioural mimicry facilitates liking and affiliation between people to help them bond (Lakin et al., 2003). Head nodding is one of the behaviours studied by Chartrand and Bargh (1999), but they also considered others, like general head movements, postures, gestures, and various other body movements and non-verbal mannerisms.

However, dyads in real-world interactions had seldom been assessed directly. In a study by Ramseyer and Tschacher (2011), based on sessions during psychotherapy, the researchers hypothesized that automated analyses of video might be able to detect mimicry in head and body movements. They used a motion energy analysis and found that coordination was more pronounced in the psychotherapy condition where the data come from the same interaction, compared to a control condition of pseudo-interactions, where the data come from different interactions generated by shuffling time structures. This automated method allowed a more detailed understanding of the patterns of mimicry in a real-world conversation. However, the resolution of video analysis will always be limited because one video recording provides only a flat image of the people (i.e., no 3D data).

Hale, Ward, Buccheri, Oliver, and Hamilton (2020) recently measured head nods using motion capture to achieve a greater level of detail, and to determine the precise parameters and timing of mimicry. 30 dyads completed a structured conversation where they took turns to describe a picture to their partner while head nodding was recorded. They calculated the wavelet coherence (i.e., coordination) of these head movements within dyads as a measure of their non-verbal coordination. They found that mimicry occurred as a low frequency (<1.5Hz) nodding behaviour in the listeners, following the speaker's head nodding with a time lag of only 600 ms. This is consistent with a reactive mechanism, as implied by simple reactive mimicry mechanisms (Heyes, 2011). The present thesis will build on the work of Hale et al. (2020) to explore mimicry across contexts and in a broader social context.

An alternative way to examine the role of head nodding mimicry in conversation is to manipulate mimicry in Virtual Reality (VR). VR is a popular tool for research on social interaction because people usually react to virtual characters similarly to how

they would with real people (Bailenson, Blascovich, Beall, & Loomis, 2003). Using this idea, Bailenson and Yee (2005) created a method for virtual mimicry where participants wore a head mounted display (HMD) that tracked their head movements while the virtual character either mimicked their movement or made head movements recorded from a previous participant. They report that participants preferred interacting with the virtual character who mimicked and suggest that a time lag of 4 seconds between the movement of the participant and that of the virtual character is optimal, though this was not systematically tested.

The VR method has the advantage of high experimental control over mimicry manipulation. Virtual humans can be artificially created to only perform the necessary behaviours, such as speaking, blinking, or nodding. In addition to overcoming many of the challenges associated with traditional paradigms, virtual mimicry could be used to test competing hypotheses about how people respond to mimicry or the impact of individual behaviours. For example, it could be used to manipulate the precise time lag of head nodding behaviour between the participant and the virtual character.

1.2.2 Head Nods as Backchannels

Head nodding is likely to be a major communication backchannel, that is a feedback signal from the listener in a conversation to indicate that one has understood or taking note of what the speaker is saying (Allwood & Cerrato, 2003; Duncan, 1972; Yngve, 1970). In addition to receiving or understanding, some researchers have also considered the head nod as a backchannel signal that can indicate acceptance or agreement with what someone is saying (Cerrato, 2005; Heylen, 2006; McClave,

2000). Sometimes, we nod to the speaker only because they are saying something that we had previously said or thought, so we are nodding not just to the speaker but to ourselves as a form of ‘back-agreement’, or we nod to say ‘yes’ to ourselves to confirm that we understood what exactly the speaker means as a form of ‘processing nod’ (Poggi et al., 2010). The speed and manner of the backchannel response may carry meaning as well. Backchannel nodding is usually considered to be fast and can communicate engaged attentiveness with the speaker (Clark, 1996).

In addition to the low frequency (<1.5Hz) nodding behaviour found in listeners following the speaker’s head nodding with a time lag of 600 ms, Hale et al. (2020) also found an unexpected pattern of high frequency (>1.5Hz) nodding behaviour from the listeners. They hypothesized that this could be a newly identified social signal with a different meaning to the low frequency mimicry behaviour they found. These results provided a step towards the quantification of real-world human behaviour in high resolution. The quantification of high frequency fast nodding is useful for testing different hypotheses about the meaning of head nodding in conversations. For example, the high frequency nodding is produced mainly when participants are listening, which suggests this is likely to be a backchannel behaviour. It is a very quick head nod that is visible but very subtle, and usually not something people notice during conversations, unless someone were to point them out. Behavioural mimicry, on the other hand, is thought to be a slower type of nodding in the low frequency range.

Some open questions that come from this distinction of two types of nods, include how these different nodding behaviours – fast and slow – change with context. In Chapter 2, we will examine this in more detail and try to determine what these different signals might mean and how they are used across different conversational

contexts. For such use of nodding to be meaningful it is important to have a robust understanding of when and why people engage in fast nodding and slow nodding. Context provides an important way to understand the meaning of social signals, because we would expect some signals to remain constant across contexts, while others might change. This may also allow us to determine if there are robust individual differences in nodding behaviour. Another interesting question to ask is if head nodding behaviour are related to conversational outcomes, like other cognitive factors. Based on results from Chapter 2, Chapter 3 will examine if coordination between two people in a conversation might be related to measurable outcomes of the conversations, including how much the two people remember new information. In other words, we will see if we can find a correlation between head nodding and memory, or the information learnt in a conversation. Based on results from Chapter 3, Chapter 4 will examine both the fast and slow nodding behaviours using VR to test competing hypotheses about how people respond to these different behaviours. We also have methodological motivations to study head nodding, and in the following sections of the thesis, we will cover our approach in detail.

1.2.3 Approaches to the Study of Interpersonal Coordination

Hömke et al. (2017, 2018) provide a good example of a modern approach to studying interpersonal coordination. In their studies, they first *measure* real-world blinking, and *analyse* blinks in relation to dyadic conversation, and then they *artificially generate* a virtual character who could blink to test how people respond to blinks. This thesis will follow the same approach. Both Hömke's studies and this

thesis rely on the idea that specific social behaviours can be identified and understood in terms of ‘behaviour rules’ (Hadley, Naylor, & Hamilton, 2022)

The idea behind this approach is that a series of simple behavioural rules that guide how behaviour are generated are sufficient to explain some complex social interactions. For example, the coordinated movement of schools of fish appears sophisticated and complex but has a very simple basis where the behaviour can be explained by combining the rules ‘avoid those too near’, ‘align with those at an intermediate distance’, and ‘move towards those further away’ (Huth & Wissel, 1994). Like the idea of a dynamic system, which has a cascading number of possible non-linear interactions, which implies that a small change in the initial conditions may have widespread effects throughout the system, so too can these simple rules at the individual level of a single fish result in apparently complex collective coordination at the group level of the school (Couzin, 2009).

Characterizing human social interaction in terms of behaviour rules then, might provide a good explanation for interpersonal coordination since the behaviour of one individual is linked to the behaviour of their partner. In this thesis, we are using this approach to identify and analyse head nodding behaviour. For example, people tend to mimic head movements with a delay of 600 ms (Hale et al. 2020). Thus, a simple rule of ‘copy his head with 600 ms delay’ might be enough to create naturalistic head mimicry behaviour (Hale et al., 2020). A behavioural rule like this is then relatively easy to implement in a virtual character (Bailenson & Yee, 2005), making it testable. In other words, it is possible to test how participants respond to these virtual characters with or without the behaviour rule.

In this thesis, we want to 1) *Measure* real-world head nodding using high resolution motion capture to identify a behaviour rule based on hypotheses about the

meaning behind behavioural mimicry and backchannel signals. We then want to 2) *Analyse* these head nodding rules in relation to dyadic conversation in speaker-listener roles. Lastly, we want to 3) *Artificially Generate* this behaviour rule in a virtual character that can then enact and engage in conversation with participants so that we can test how people respond in a conversation that includes this rule.

There are some subtle differences between this work and the studies by Hömke et al., particularly in the *Artificially Generate* stage. In their study with a virtual human (Hömke et al., 2018), the behaviour rule was implemented by an experimenter who controlled the virtual character. However, we aim to fully implement head nodding rules without needing human intervention. Implementing the behaviour rule in the virtual character that the participants interact with, will enable us to use the timing and frequency properties of head nodding to build highly responsive characters to make the coordination seem natural. This in turn can help demonstrate their sensitivity or how subtle social signals can be. Similarly, identifying where these behaviour rules break down can give us clues whether certain behaviours require more sophisticated rules or other cognitive models altogether (Hadley et al., 2022). For example, manipulating the social context could allow us to identify whether a behaviour rule breaks down or not under certain conditions.

This approach highlight how new technologies and experimental designs can be used to address fundamental theories and the formulation of hypotheses in social interaction. New methods of *measuring* and *analysing* social interaction are enabling a more detailed picture of behaviour and understanding this picture during real-world social behaviour can help in the development of the next generation of virtual characters to *artificially generate* interpersonal coordination, and effectively reverse-engineer social interactions to understand its underlying cognitive basis.

In the next three sections, we will cover all three aspects of this approach in more detail. This will require of us to first review some challenges associated with existing approaches to *Measuring* (Section 1.3), *Analysing* (Section 1.4), and *Artificially Generating* (Section 1.5) interpersonal coordination, before we introduce the research aims and overview of the experimental studies in the thesis (Section 1.6).

1.3 *Measuring* Interpersonal Coordination

As experiments in social neuroscience are striving toward more ecologically valid environments using naturalistic stimuli and paradigms, it becomes more and more important to capture the richness of social interaction in a manner which preserves the dynamic nature of multimodal non-verbal signals. Traditional cognitive approaches to understanding social interaction involve studying isolated participants responding to stimuli on a computer screen or interacting with confederates. Such designs tightly control the variables involved and isolate targeted behavioural or cognitive constructs. However, it is increasingly recognized that this is a poor model of real-world social interaction, as in many ways these studies do not resemble the complexity and dynamic nature of stimuli and behaviours in real life (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Risko, Richardson, & Kingstone, 2016). Real world conversations involve rich multimodal and dynamic coordination between individuals in which the behaviour of each person is strongly interdependent with those of their conversation partner. Natural mimicry, for example, occurs in a rich context of other coordinated actions, non-verbal signals, and speech. It has therefore been challenging to measure and model mimicry in these contexts. Psychologists and neuroscientists are now trying to understand dynamic social interaction, in which two people respond to each other in real time (Heerey, 2015; Schilbach et al., 2013). This could be very important for determining natural parameters in social interaction, such as its timing or rhythmic properties, since research has shown that these patterns cannot be accurately represented in non-naturalistic paradigms when one participant is replaced with an experimenter or confederate (Bevalas & Healing, 2013; Kuhlen & Brennan, 2013). In the following section we will review how methods for recording movement have rapidly developed.

1.3.1 Recording Interpersonal Coordination in Dyads

Manual Coding. Researchers have traditionally used several approaches to measure interpersonal coordination in dyads, but many use cumbersome manual audio-visual annotation methods, which yield low resolution data and is very time consuming. Early research involved recording video of natural conversations, which were then coded, frame-by-frame, to manually detect changes in speech, posture, or facial expressions (Bernieri, Gillis, Davis, & Grahe, 1996; Condon & Ogston, 1966; Kendon, 1970). Although these early experiments have been instrumental with detailed descriptive analyses revealing the fundamentals of social behaviour, they have not been able to capture the full richness of a social interaction in a precise and quantifiable manner. Moreover, annotating and interpreting video is an intensely manual process that generally must be performed in real-time as the video plays, requiring frequent pauses and playback. This was a major limitation and could result in analysis that is prone to human errors when many different behaviours are coded at the same time (Cappella, 1981; Grammer, Kruck, & Magnusson, 1998).

Automatic Recording. More recently, researchers have begun to use automated and quantitative measures to identify, and track coordinated non-verbal behaviours between people, which makes possible the discovery of very subtle and transitory patterns. This lets the researchers gather a lot more data to make better statistical analyses, like identifying behaviour rules. These methods are also faster and less labour intensive. An example of such a method is to use image processing techniques to calculate frame-to-frame differences in people's body movements from video to calculate their overall 'motion energy' during naturalistic conversations (Fujiwara & Daibo, 2016; Paxton & Dale, 2013; Ramseyer & Tschacher, 2010; Schmidt, Morr, Fitzpatrick & Richardson, 2012). Another example is to use computer

vision analysis (Dunbar, Jensen, Tower, & Burgoon, 2014). However, these methods lack the resolution to capture specific data from individual body parts or to capture movement in depth (relative to the camera). To achieve this level of measurement, another option is to use modern motion capture technologies to directly record the movements of each participant.

Motion capture (Bouaziz, Wang, & Pauly, 2013) provides more objective and detailed data about an interaction, and has been employed in a number of recent studies to study the specific modality of body movement in different scenarios (e.g., Feese, Arnrich, Tröser, Meyer, & Jonas, 2011; Hale et al., 2020; Poppe, Zee, Heylen, & Taylor, 2013), and with focus on specific body parts, like fingers (Oullier, Guzman, Jantzen, Lagarde, & Kelso, 2008), hands (Schmidt, O'Brien, & Sysko, 1999), and heads (Ashenfelter, Boker, Waddell, & Vitanov, 2009; Hale et al., 2020).

There are several different types of motion capture technologies. For the experiments in this thesis, we use an optical system which triangulates the position of retroreflective (passive) markers on the participant's body using a multi-camera setup (Optitrack, NaturalPoint Inc.). With proper calibration, the position of the markers can be translated to the position of the joints, and the position and rotation of skeleton segments calculated from the triangulation. With this system, there is the option to use light-emitting (active) markers. There are also magnetic systems, which use a weak magnetic field to detect markers (Feese et al., 2012), as well as cheaper options using 3D-camera systems like the Kinect (Won, Bailenson, Stathatos, & Dai, 2014). Optical systems generally have the disadvantage that markers can easily become obscured or occluded from the cameras' view (especially in the presence of more participants) and will require additional post-processing, whereas magnetic systems tend to be more precise and require less calibration.

The development of automated measures of recording behaviour is opening the way to new studies of social interaction, but also provides a new challenge in analysing it. In the following section we discuss how further advances in wavelet analysis (Issartel, Bardainne, Gaillot, & Marin, 2015; Schmidt, Nie, Franco & Richardson, 2014) can be used to understand social coordination in better detail. More specifically, we will demonstrate how it can help us quantify the relationship between the motion patterns of head nods to analyse the level of interpersonal coordination, or coherence, between people.

1.4 *Analysing* Interpersonal Coordination

The data we get when capturing body movement and its temporal progression is difficult to quantify and interpret. The head can move in different directions (i.e., yaw, pitch, roll), and with high or low frequency and amplitude. The interpretation of the head trajectory might also become more complex when we have a long time-course with two or more people interacting. To make sense of a biological trajectory or behaviour over time we must analyse time-series data. Recent analytical methods can deal with this non-linearity of movements and interpersonal coordination across multiple timescales. Time-series can be analysed according to the timing, frequency, or both timing and frequency of movement (Fujiwara & Daibo, 2016; Grinsted, Moore, & Jevrejeva, 2004; Issartel, Marin, Gaillot, Bardainne, & Cadopi, 2006). Next, we will cover a range of analytical methods as a guide to what each method are capable of and what its limitations are. We begin with temporal analysis methods, including Cross-Correlation and Cross-Recurrence Quantification Analysis. This is

followed by frequency analysis methods, including Fourier Transform, Windowed Fourier Transform, Wavelet Transform, and Cross-Wavelet Coherence Analysis.

Cross-correlations (Boker, Xu, Rotondo, & King, 2002; Rotondo & Boker, 2002) are the most common approach to measure the timing between two time-series of movement and observing at what time-lag the two time-series are most highly correlated. This peak correlation would indicate the timing at which the two participants match each other's movements (Fujiwara & Daibo, 2016). In other words, a correlation coefficient (*r*) is calculated for each relative time-lag between the two time-series. A lag of 0 would indicate mutual synchrony, whereas a lag of -1 would shift one time-series by one increment, decided by the length of the window that is applied (Figure 1-1). For example, while the movement of participant A is at time *t*, the movement of participant B is at *t* + 1, at which point the correlation is performed again. If two participants' movements are synchronized, *r* will peak at a lag of 0, representing that changes in their motion coincide (Ashenfelter et al., 2009).

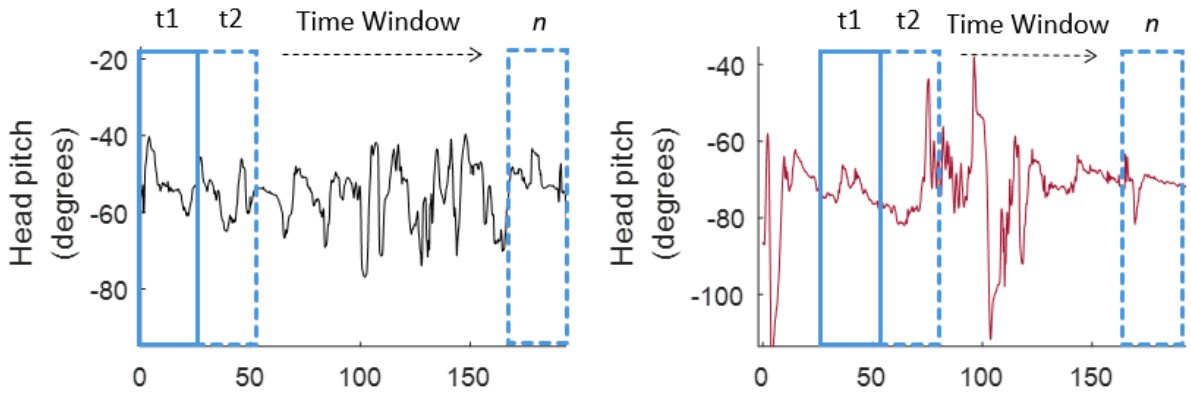


Figure 1-1. Windowed cross-correlation. Measures the timing between two time-series of movements (e.g., participants). A sliding time window observes at what time-lag the two have the highest peak correlation, or at what time the two participants match each other's movements.

Another method for analysing the temporal aspects, or the amount of coordination, in a dyadic interaction, is known as Cross-Recurrence Quantification

Analysis (CRQA) (Marwan, Romano, Thiel, & Kurths, 2007). This method originates as a non-linear analysis used to compare the temporal patterns of complex systems (Coco & Dale, 2014), and has shown to explain some properties of time-series data that linear methods, such as cross-correlations, cannot (Marwan, 2008). CRQA has two steps – first the researcher must identify 'system states' which occur repeatedly during an interaction. Within the field of interpersonal coordination these states represent the points in time that the social signals show similar patterns of change (Demos, Chaffin, & Kant, 2014). Secondly, CRQA can then test how often these states are visited and if there is coordination between the state of the two people in the interaction. A high rate of recurrence reflects a high temporal coordination of behaviours. CRQA can also measure patterns that are far apart in time to observe the similarity or influence those two signals can have on each other. This can work well for discrete measures of an interaction like word use or single gestures. However, for many continuous measures like head movements, the first step is challenging. There is no clear-cut way to identify states in head nodding for example, which makes it hard to use CRQA. In the following section, we will examine in more detail an alternative to temporal analysis methods known as frequency analysis methods, where we can measure the frequency, or the combination of time and frequency, of coordination.

1.4.1 Frequency Analysis

The frequency of movements, or the number of occurrences of a movement, can be observed using frequency analysis to mathematically transform the data in various ways from a time-amplitude signal into either a frequency-amplitude or time-

frequency representation of the signal. Taking our analysis of head nods as an example, the raw signals we get (i.e., head pitch rotation in degrees) are time-domain signals. When we plot these signals, we obtain a time-amplitude representation of the signal (e.g., degrees of head pitch on the y-axis, and the time-series on the x-axis). However, this representation is not always the best representation, and sometimes the most interesting information is hidden in the frequency content (i.e., spectral components) of the signal. The frequency is measured in cycles/second, or Hertz (Hz), and to measure or find the frequency content of a signal, we need to apply mathematical transformations to obtain information that is not readily available in the raw signal (Issartel et al., 2006).

Fourier Transform (FT). The most used transform is the Fourier Transform (FT). This frequency analysis provides a description of the frequency content of a time-series (Issartel et al., 2015). By applying the FT, we get a frequency-amplitude representation of that signal. This tells us how much of each frequency exists in the signal. However, it does not tell us *when* these frequency components exist. Thus, unless we are only interested in what frequency content exist in the signal, the FT is not a suitable method for non-stationary signals that change frequently, like in most biological signals (e.g., ECG, EEG). For example, the FT assumes that the time-series follows repetitive patterns and stable frequencies over time and can therefore be more suited for studying synchrony where participants make repetitive movements, such as rhythmic walking or musical coordination.

Windowed Fourier Transform (WFT). Methods have been created to account for both the time and frequency content of a time-series. By dividing the signal into segments, or windows, we could assume each window to be stationary (Figure 1-2). This method, known as Windowed Fourier Transform (WFT), though allowing for the

detection of sudden changes in the frequency content, suffers from a resolution problem which creates temporal ambiguity. What we can know are the *time-intervals* in which *frequency bands* exist. Thus, the problem with WFT has to do with the width of the window used. For example, a small window would not allow for the detection of an event larger than the window while maintaining a good localization in time. On the other hand, a large window will consider the long-term event (frequency domain) but with high imprecision in the temporal domain (Issartel et al., 2015).

Wavelet Transform (WT). When the time localization of the frequency component is needed, another transform giving us the time-frequency representation of the signal is the Continuous Wavelet Transform, or simply the Wavelet Transform (WT), which tells us what time a specific frequency component occurs (Morlet, 1983). We need the WT to analyse non-stationary signals and to overcome the resolution problems related to the WFT. Unlike the WFT which has a constant resolution at all times and frequencies (the width of the window is selected once for the entire analysis), the WT gives a variable resolution that changes the width of the window as the transformation is computed for every single frequency component (Figure 1-2).

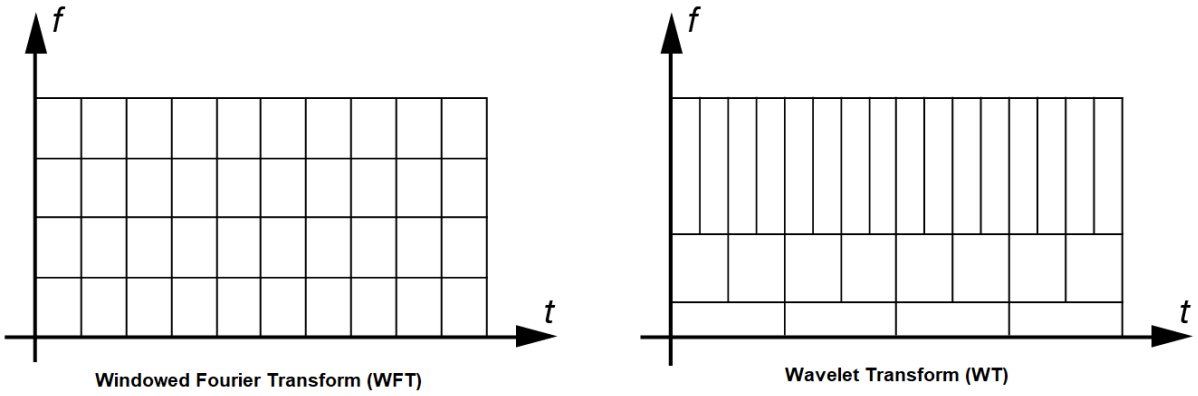


Figure 1-2. Time-Frequency analysis resolution. The Wavelet Transform (WT) gives a variable resolution that changes the width of the window as the transformation is computed for every single frequency component. This gives us good time (*t*) and poor frequency (*f*) resolution at high frequencies, and good frequency (less spectral ambiguity) and poor time resolution (more temporal ambiguity) at low frequencies.

Solving the resolution problem makes the WT efficient in the study of non-stationary (e.g., biological) signals that dynamically vary in frequency, and allows the detection of small changes and to analyse the temporal evolution of each frequency component. This in turn makes the WT more suitable for studying coordination in spontaneous or non-repetitive social interactions such as having a real-world conversation (Fujiwara & Daibo, 2016; Issartel et al., 2006). Take head nodding, for example, the time-amplitude representation of the signal (i.e., degrees of head pitch at different times) are transformed into the time-frequency domain (i.e., frequency of head nodding at different times).

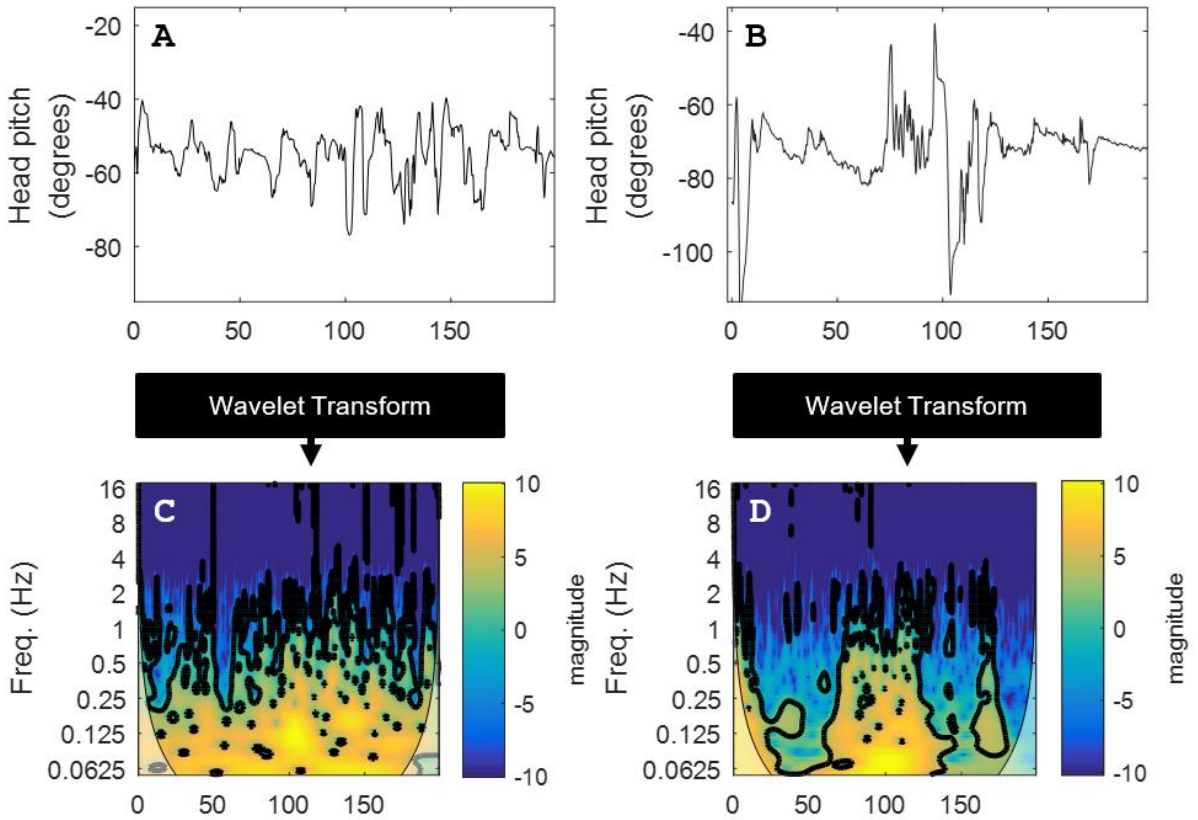


Figure 1-3. Continuous wavelet transform (WT). Measures the time-frequency component of the time-series (i.e., signal). The head pitch trajectories for both participants (A, B) are subject to a wavelet transform to get the time-frequency component (C, D). The magnitude of wavelet power is represented by color, where blue is low power, and yellow is high power. The time is represented on the x-axis (200s) and each frequency on the y-axis (0-16 Hz).

1.4.2 Cross-Wavelet Coherence Analysis

Cross-Wavelet Coherence (CWC) is a time-frequency based analysis following a wavelet transform, that allows us to consider the degree of similarity between two sets of data – or time-series, and the progression in time of this interaction (Grinsted et al., 2004). It can be useful to think of CWC as the correlation between two wavelet transforms. As such, this analysis evaluates the cross-frequency of two signals across time, and hence can uncover how the time-localized coherence at different frequency ranges (timescales) changes across the course of a trial. A cross-wavelet plot (Figure 1-4) displays the coherence (correlation between the speaker and listener in a dyad). In general, high coherence is interpreted as a high degree of coordination because it indicates that two people are moving with the same frequency in the same time window, but are not necessarily nodding synchronously (i.e., at exactly the same time).

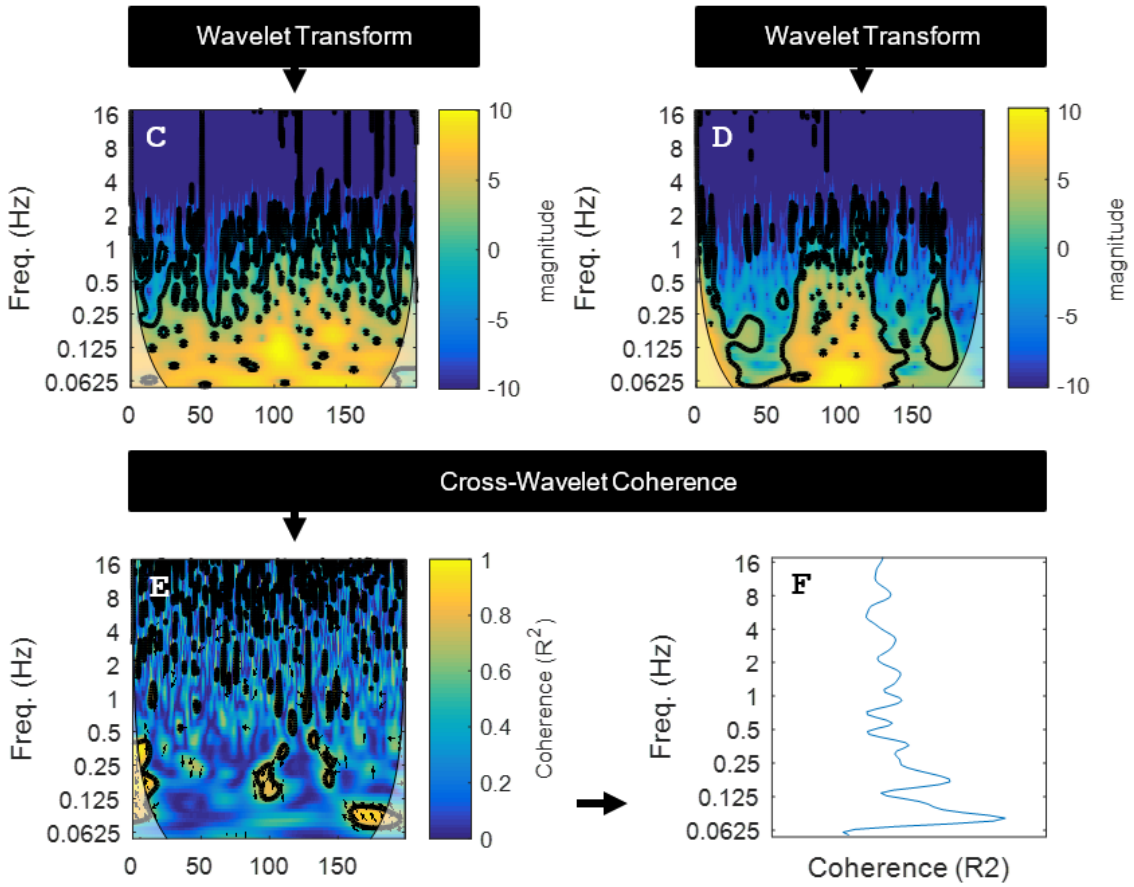


Figure 1-4. Cross-wavelet coherence (CWC). The CWC (E) is calculated between the two transformed time-frequency signals (C, D). The magnitude of wavelet power and wavelet coherence is represented by color, where blue is low power, and yellow is high power. The time is represented on the x-axis (200s) and each frequency on the y-axis (0-16 Hz). The coherence value (R^2) is then averaged over time and over all trials to obtain the overall frequency of coherence.

CWC gives us two measures of an interaction – a coherence measure (R^2) which tells us if two people move at the same frequency within the same time-window, and a phase measure which tells us the precise temporal relationship (i.e., time lag) between them. These measures can, for example, indicate whether one person led or followed the other at a specific frequency and for how long (Issartel et al., 2015).

Understanding the timing and at which frequencies all these behaviours occur can help us answer how and why we use these signals. While CWC analysis is a relatively new tool for studying social interactions (Issartel et al., 2015), several proof-of-concept studies using CWC have found that it can be applied to time-series data on body movements in a dyadic interaction. Varlet, Marin, Lagarde, and Bardy (2011) reported that dyads engaged in a visual tracking task influenced each other and produced spontaneous postural coordination. The researchers also used the phase measure to evaluate the occurrence of postural coordination. A similar coordination of postural sway was also reported (Sofianidis et al., 2012). In another study using CWC, Washburn et al. (2014) recorded movement data in dance settings and reported a higher level of coherence in trained dancers compared to non-dancers when they performed with their confederate dance partner. Walton, Richardson, Langland-Hassan, and Chemero (2015) also demonstrated that CWC could show the dynamics of movement coordination between improvising musicians.

Further studies in more social settings have also revealed that interpersonal coordination occurs at multiple timescales (Schmidt et al., 2014), and frequencies (Fujiwara & Daibo, 2016). In the study by Schmidt et al. (2014), they investigated

how bodily coordination is distributed across different nested timescales in a joke telling task during structured conversation. Dyads performed knock-knock jokes while the researchers recorded their movement using a Kinect camera. To assess the degree and pattern of participants' movement coordination at individual timescales, they used CWC and calculated both the coherence and phase measures for the whole trial, as well as at the minor timescales of the whole joke, the setup of the punch line, the turn-taking exchange, and the utterance. The coherence analysis revealed a greater than chance coordination (high coherence) of the joke teller's (speaker) and joke responder's (listener) movements at all timescales. In addition, the phase measure analysis revealed that the joke teller's movements led those of the joke responder at the longer timescales, which shows that their coordination are in-phase rather than anti-phase. These results demonstrate that complex interpersonal coordination is constructed from a set of rhythms associated with nested timescales within a structured social interaction.

Fujiwara & Daibo (2016) followed this up with a similar methodology using CWC but did not employ a specific task. Instead, they focused on unstructured conversations. In this type of scenario, the turn-taking between the participants in the dyad are not controlled. The researchers examined whether coordination of hand and head movements represented in the time-frequency plane would be observed even in such unstructured or free flowing conversations. To test this, they used a pseudo-pairing experimental paradigm first proposed by Bernieri and Rosenthal (1991). In this paradigm, video clips of dyads are isolated and re-combined in a random order. The coordination of these pseudo-pairs is then compared to the real pair. By using CWC, Fujiwara and Daibo (2016) hypothesized that the amount of coherence would be higher in the real pairs than in the pseudo-pairs. They

discovered a high degree of coherence at low frequencies (<0.25 Hz) and less coherence at high frequencies (>4Hz), which supports the validity and possible utility of using CWC to evaluate the structure of social interactions. However, Fujiwara and Daibo (2016), and researchers using similar pseudo-pairing paradigms before them, only tested whether their coherence pattern was present in the pseudo interactions *between* the pairs, and not *within* the same pairs. Moreover, none of the studies mentioned have recorded motion frequency in detail using motion capture to isolate the different movement coordinates of the head trajectory (i.e., pitch, yaw, roll).

Recently, Hale et al. (2020) improved on both these aspects – they used motion capture to isolate the different movement coordinates of the head trajectory, as well as provide pseudo-pairings *within* dyads to provide a stronger test where the pseudo pairs have the same general movement characteristics as the real pairs. They measured mimicry using single head motion sensors. A sample of 26 dyads (n=52) was engaged in a structured conversation where each participant took turns to describe a picture of a complex social scene to each other. They used CWC analysis to calculate levels of interpersonal coordination in real trials compared to a pseudo dataset created by matching data from different trials *within* the same pairs. What they found was different results for slow and fast head nods. The results showed a positive coherence at frequencies below 1.5 Hz, but also to their surprise, the real interactions showed lower than chance coordination of head movements at frequencies between 1.5 – 5 Hz. This unexpected finding suggests a systematic decoupling of head movements at higher frequencies, similar to findings of divergence in dialogue where people systematically diverge from one another in their use of syntactic constructions (Healey, Purver, & Howes, 2014).

These are interesting findings, and Hale et al. (2020) hypothesize that it could potentially be different signals carrying distinct social information. More specifically, head nodding behaviour in the high frequency range are quick and spontaneous, with a reported decoupling of head nodding that is driven by the listener. This 'fast nodding' could potentially contain different social information from the slower head nods in the low frequency range, since this frequency range is consistent with traditional observations of mimicry behaviour where modelling shows this behaviour is generated by a mechanism with a 600 ms time lag between speaker and listener.

Thus, Hale et al. (2020) interpret these results in the framework of two types of coordination, which they refer to as 'mimicry' and 'fast nodding'. They refer to mimicry as events where two people perform the same movement shortly after each other (e.g., one person nods and the other nods shortly after), which is a slower type of nodding behaviour. Fast nodding occurs when one person makes a small high frequency nod, and the other person typically does not move in this way. This thesis will explore the hypothesized distinction between these signals, as it is still not clear whether they are two distinct social signals.

The quantification of these two types of signals is also useful for allowing us to identify them as two separate behaviour rules based on different frequency ranges of head nodding. These behaviour rules can then be programmed in virtual characters to test our hypotheses about their meaning by having participants respond to these virtual characters with or without the behaviour rules. To *artificially generate* such interactive virtual characters, we have in this thesis taken steps to first *measure* and *analyse* social signals in a person's behaviour using motion capture technology and CWC (Section 1.3 and 1.4). In the next section (1.5), we will discuss how to *artificially generate* interpersonal coordination in more detail.

1.5 *Artificially Generating* Interpersonal Coordination

To *artificially generate* interpersonal coordination in virtual characters, we must first capture and understand natural interpersonal coordination. In previous sections, we have demonstrated how to do this and to identify behaviour rules. We now want to use these behaviour rules in combination with virtual reality technology to create highly responsive virtual characters that can interact and coordinate more naturally. The main challenge here is to generate animation that is highly responsive without unnatural jerkiness and that matches current models of coordination behaviour. Previous studies have also used various techniques to create such characters that can generate simple social behaviours such as body postures (Gillies, Kleinsmith, & Brenton, 2015), hand gestures (Kopp & Bergmann, 2013), or head nods (Huang, Morency, & Gratch, 2010). Such systems traditionally replay the measured movements with good timing but cannot yet create the detailed coordination found in real social interaction, nor have they been linked to psychological theory. Using motion capture technology to build better simulations of interpersonal coordination with a grounding in psychology has the potential to greatly improve the realism of social interaction with virtual characters. Hypothesis testing can then also benefit from seeing how people respond to these virtual characters, so that we can use virtual reality as a tool to further test and challenge our psychological theories.

In the next couple of sections, we cover how the virtual environment, including the visual appearance of the virtual characters, are created (1.5.1). Second, we describe how to generate realistic motion with high resolution motion capture data and mapping that on an interactive virtual agent (1.5.2). We will then present the research aims of the thesis and an overview of the experimental chapters (1.6).

1.5.1 Creating a Virtual Environment

To create a virtual experimental framework and display virtual environments to a participant, we need two components – a hardware and a software system. Researchers now have access to advanced computer graphics software that allow the creation of rich virtual environments. A graphics engine that is widely used in the research community today is Unity (Unity Technologies, 2022). Other graphics engines like Vizard (WorldViz, 2022) may have the same capacity to render virtual environments, but for the virtual reality experiment in this thesis, we are using Unity.

The hardware part is known as virtual reality headsets, or head-mounted displays (HMDs), which display stereoscopic images that is optically converged into a single image and updated in real-time. The recent popularity of HMDs has extended the reach of Unity to include the development of VR-compatible content that can use the HMDs ability to track the participants head movements to update the view. Thus, the rendering of the scene inside the HMD is determined by the head position and orientation of the participant. To add realism, HMDs often come with headphones to provide audio feedback and sometimes hand-tracking controllers that can provide haptic feedback. These can add significantly to the reality of what is being perceived and increases the likelihood that the participants would respond realistically.

However, how realistic must the visual appearance of the virtual characters we create as part of the virtual environment be to work in an experimental setting? The concept of the ‘uncanny valley’ was introduced by Masahiro Mori in 1970. He suggested that there is a non-linear relationship between how realistic a virtual character looks and how people perceive it and proposed that characters with near-realistic appearances are judged as ‘uncanny’ (Mori, MacDorman, and Kageko, 2012). A recent study suggest that this uncanniness occurs when there is an

incongruency between the characters appearance and behaviour (Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012). For example, a highly realistic character that behaves in a strange way, can still be perceived as more uncanny than a cartoon character behaving in a similar fashion. Other studies suggest choosing a middle ground between a stylized and realistic looking character (Zell et al., 2015). However, to generate socially realistic virtual characters, they need to be socially interactive, and in the next section we will describe how this can be achieved by generating interactive virtual agents, that can display non-verbal behaviour.

1.5.2 Generating an Interactive Virtual Agent

It is nowadays widely emphasized that social cognition is *embodied* in a living organism and *extended* into our technological and social environment (Clark, 2013). This has led to the prominence of “second-person neuroscience”, which studies the real-world interaction between two people (Shilbach et al., 2013). The most important form of face-to-face communication is the conversation, as it provides a forum or arena for all our social interactions. We have also seen that non-verbal signals are a crucial part of face-to-face conversations.

With virtual reality technology, we can manipulate the social interaction by changing the visual cues or signals to represent the behaviour of the virtual character that we want the participants to be able to observe. Hence, by programming behaviour rules into the behaviour of the virtual characters, they transform from being simply virtual ‘characters’ to being interactive virtual ‘agents’, allowing them to participate and enact in conversations in a more natural way or fulfil

certain social roles. Creating believable virtual characters and generating interactive behaviour is a great challenge for computer scientists (Pan, Gillies, Barker, Clark, & Slater, 2012; Rizzo & Talbot, 2016).

To generate interactive social signals, the virtual agent must first be able to detect the behaviour of the participant, for example by utilizing the positional head-trackers or microphones implemented with the HMD. The 'Unity' software can then be used to program specific responses (e.g., behaviour rules) conditional on the participants' behaviour. For example, knowing the location of the participant's head means that a virtual agent can be programmed to use behaviours such to copy the head movements (Bailenson & Yee, 2005; Hale & Hamilton, 2016b) or follow the gaze (Forbes, Pan, & Hamilton, 2016) of the participant.

Studies using virtual reality as a tool have shown that people keep appropriate social distance from virtual agents (i.e., proxemics) (Bailenson et al., 2003), and mimic their behaviours (Vrijsen, Lange, Becker, & Rinck, 2010). A range of studies have also successfully replicated psychological constructs with virtual reality, including rapport (Hale & Hamilton, 2016b; McCall & Singer, 2015), prosociality (Hale et al., 2020), and social anxiety (Pan et al., 2012) among others.

In the study by Hale and Hamilton (2016b), they used motion capture (Polhemus magnetic motion tracker) and virtual agents displayed on a projector screen to test the 'social glue hypothesis' that mimicry leads to increased rapport. The participants interacted with two virtual agents in an interactive picture description task commonly used in confederate studies (Chartrand & Bargh, 1999; van Baaren et al., 2009). In this task, one agent copied the head and torso movements of the participants after a specific delay (1 and 3 seconds), while the other agent showed pre-recorded natural head and torso movements without mimicry. The participants interacted with the

mimicking and non-mimicking agents one after the other, in a within-subjects design, and rated feelings of rapport toward the agents. The results showed no effects of mimicry on rapport, and the researchers conclude that being mimicked does not necessarily increase rapport. They argue that we should be careful in accepting the 'social glue hypothesis' and be cautious about the fragile effects of being mimicked. Some of the future directions that the researchers present with this study include more rigorous methods, as they only use a single magnetic marker to measure head and torso movements with the Polhemus motion tracker, as well as not utilizing fully immersive virtual HMDs to present their virtual agents.

With high resolution motion tracking technology (Section 1.3) we can capture the actual non-verbal behaviour displayed by real people and use this to create behaviour rules to drive the behaviour of virtual agents in a *controlled* manner. I like to emphasize the word 'controlled' here because experimental frameworks in virtual reality typically specify a certain behaviour or focus only on a small subset of the behaviours compared to real-world social interaction. This is due to the high computational demand of simulating a fully responsive multimodal and dynamic conversation. Nevertheless, even very basic forms of social interactions can be enough to be perceived as realistic, which makes the use of simple behaviour rules ideal for using in virtual reality experiments as they are both perceivable and testable. For example, both behavioural (Wilms et al., 2010) and neuroimaging (Schilbach et al., 2011) studies show that very basic but contingent eye gaze behaviour can be enough to elicit a sense of realism for participants interacting with virtual agents. Similar results have also been demonstrated with eye blinks (Bailenson & Yee, 2005), and head nods (Huang et al., 2010), showing that these simple behaviours might be enough to elicit naturalistic real-world behaviour.

Blascovich et al. (2002) emphasize that a general assumption that seems conspicuous in all psychological research is that experimental manipulations of perceived (i.e., real-world) and imagined (i.e., written scenarios) stimuli are essentially equivalent for understanding psychological processes. Experimental manipulations of imagined stimuli cost less, require less effort, and provide a greater degree of experimental control (i.e., precise manipulation of variables). However, a greater degree of experimental control often comes at the cost of ecological validity (i.e., the extent to which an experiment is like situations encountered in everyday life). Thus, a trade-off typically exists between experimental control and ecological validity. Technological advances in both motion capture and virtual reality systems have allowed researchers to lessen this trade-off by facilitating an increase in ecological validity without entirely sacrificing experimental control (Blascovich et al., 2002), or vice versa. We have already seen how virtual reality increases the ecological validity by creating more immersive and realistic scenarios, while maintaining a high degree of experimental control by being able to manipulate the virtual environment and generating interactive virtual agents.

In the next and final section of this chapter, we will summarise the research aims and give a brief overview of the experimental studies in this thesis.

1.6 Research Aims and Overview of Experiments

In this chapter we have reviewed previous work on interpersonal coordination and conversation behaviour. We have specifically focused on head nodding behaviour as a social signal during interpersonal coordination. Limitations in the current literature are partly attributed to methodological challenges, and in this chapter, we have reviewed several major challenges associated with existing approaches to *measuring*, *analysing*, and *artificially generating* interpersonal coordination. New methods and technologies are needed to test cognitive theories of interpersonal coordination and real-world social interactions. Recent advances in motion capture technology (Bouaziz et al., 2013) for *measuring* behavioural data and new methods for *analysing* the data (Fujiwara & Daibo, 2016; Issartel et al., 2015; Schmidt et al., 2014) in dyads and real-world interactions, together with *artificial generation* of virtual agents, make it possible to study interpersonal coordination and social interaction in much higher resolution. These advances allow researchers today to investigate how to quantify simple interaction behaviour and test hypotheses with high experimental control. However, to advance in the field it is important that these methods are guided by precise and well-specified psychological or cognitive theory to work (Pan & Hamilton, 2018). Together, method and theory will build on each other to create new testable hypotheses, which will lead to the development of new theories, which can then be challenged with new methods and models.

In this thesis, we aim to use this approach of *measuring*, *analysing* and *artificially generating* head nodding signals in dyadic social interactions, which will allow us to first “reverse-engineer” head nodding by picking the behaviour apart to determine which parameters are necessary for improving current theories, and then use these parameters to “engineer” head nodding in virtual reality to test our hypotheses about

the meaning of head nodding as a communicative signal. This will guide better research on virtual reality and provide fundamental new insight and directions for research into interpersonal coordination.

More specifically, in this thesis we will be working within the methodological framework to 1) *Measure* real-world head nodding in dyadic social interaction using high resolution motion capture to identify behaviour rules based on hypotheses about the meaning of two types of head nodding signals closely related to behavioural mimicry and backchannel signalling. We then want to 2) *Analyse* the time-frequency series of these head nodding rules across speaker-listening roles in the dyads. Lastly, we want to 3) *Artificially Generate* this head nodding behaviour in virtual agents that can interact with participants to test how they respond to the rules.

The convergence of research questions in both psychology and computing thus sets the scene for the studies presented in this thesis, which draws together these diverse research areas, combining cognitive and psychological hypothesis testing with new advances in motion capture, signal analysis, and virtual reality to provide a new level of understanding of dyadic social interaction. Next, we will give a brief overview of the experimental chapters in this thesis, and what drives the hypotheses and methods in each.

In Chapter 2, the aim is to capture two patterns of head nodding signals – fast nods and slow nods – and determine what they mean and how they are used across different conversational contexts. Context provides an important way to understand the meaning of social signals, because we would expect some signals to remain constant across contexts, while other might change. For example, the affiliation signal ‘I like you’ might be relatively unchanged across different types of conversation (e.g., about pictures versus movies), while other signals might change.

A secondary aim in Chapter 2 is to understand if there are reliable individual differences in social signalling behaviour. While some previous studies imply that some people show more mimicry than others (Salazar-Kämpf et al., 2017), there is little data to quantify this. Our large-scale data collection will provide an opportunity to explore individual differences. The results from this study suggest that fast nods are a signal of having received new information and that it has different meaning to that of slow nods. It also shows that nodding is consistently driven by context but is not a useful measure of individual differences in social skills.

In Chapter 3, the aim is to investigate the relationship between memory and learning and head nodding behaviour. We aimed to determine if the moment-by-moment coordination between two people in a conversation might be related to measurable outcomes of the conversation, including how much the two people remember new information and how they relate to each other in terms of self-other processing. This exploratory study provided initial hints that there might be a relationship between head nodding behaviour and performance on a later memory test, though further analyses were less clear.

In Chapter 4, we built on the preliminary result of Chapter 3 and aimed to test if head nodding behaviour in a virtual agent relates to memory performance. We created a virtual human who can show our head-nodding behaviour rules and test how much people remember from a conversation with the agent. In addition, in this study we also aim to examine if interactive head nodding can be used to measure how much we like the virtual agent, and whether we learn better from virtual agents that we like. The results from this study demonstrate no causal link between memory performance and interactivity and reports no significant results between measures of liking with no reliable correlations to head nodding or memory.

In Chapter 5, we analyse data from Chapters 3 and 4 together with new data from a video-based conversation task. We aim to summarise how the level of interactivity in different contexts (i.e., conversation with a real human, conversation with a virtual human, task with an unresponsive video) impacts on the memory performance of the participants. The results show that the level of interactivity supports memory and learning during conversations.

In Chapter 6, we will widen the scope and discuss both the methodological and theoretical implications and developments of this thesis, as well as its limitations.

Chapter 2. Why Do People Nod in Conversation? Head Nodding as a Social Signal Across Different Social Contexts

2.1 Abstract

Social interaction involves rich and complex behaviours where verbal and non-verbal signals are exchanged in dynamic patterns. Our overall aim is to explore new ways of modelling this coordinated behaviour as it naturally occurs in social interactions. This chapter explores the role of head nodding in different types of conversation, to define why people show this behaviour and what different types of nodding signals might mean. We present results on how people coordinate their head nodding behaviour across three different conversational contexts: a structured one-way information sharing task, an unstructured shared recall task, and an unstructured two-way joint planning task. We used high resolution motion capture to record head movements and wavelet coherence analysis to understand the dynamic patterns of head nods, testing if coordination is seen at different frequencies.

We find that dyads show coherence in slow nodding, but only in a structured one-way information sharing context, which implies that this behaviour is not a universal signal of affiliation but is context driven. We also find robust fast nodding behaviour in the two contexts where novel information is exchanged, but not when shared information is recalled. This suggests that fast nods are a signal of having received new information and that it has different meaning to that of slow nods.

Finally, we show that nodding is consistently driven by context but is not a useful measure of individual differences in social skills. These results will help us understand the role of nodding in human conversation and especially in the current era of social distancing, help us build virtual agents who engage in realistic nodding.

2.2 Introduction

Conversation is fundamental to social interaction, and people engaged in a conversation have access to many different modes of interaction, both verbal and non-verbal, which all help shape the complete experience of a social encounter. In this chapter we are interested in learning more about head nodding during face-to-face dyadic conversations in naturalistic settings. Previous work suggests that people may mimic head nods at some points in a conversation but may also show backchannel behaviour (Hale et al., 2020). To understand these social signals, it is helpful to have detailed high-resolution data recordings and to analyse this with appropriate methods. However, many current methodological frameworks are limited by time-consuming and low-resolution data collection methods, which cannot capture the full richness of a dyadic social interaction. Using multimodal data collection of dyadic face-to-face conversation, our aim with this study is to explore new ways of modelling coordinated behaviour as it naturally occurs in social interactions to understand in more detail the meaning of head nodding as a social signal.

Both verbal and non-verbal behaviours have been studied extensively by linguists and psychologists alike, trying to reveal the social signals that may be hidden behind them. A social signal is here defined as a communicative message that, either voluntarily or involuntarily, conveys information between people (Maynard-Smith & Harper, 2003). Experimental studies of conversation have primarily focused on verbal coordination, but it is now widely recognized that non-verbal coordination is important in face-to-face interaction. Despite the subtle nature of these signals, we can produce and read them in an efficient way, and to decode from them a surprising amount of social information. Non-verbal signals come in many varieties, ranging from gaze (e.g., direction, blinks, pupil dilation) (Argyle & Cook, 1976; Kendon,

1967), body-movements like gestures (Kendon, 2004), posture (Condon & Ogston, 1971), and head movements (Cerrato, 2005), to facial speech expressions (Ekman & Friesen, 1972). Social interaction is thus naturally multimodal in the sense that we rely on a variety of modes of interacting to send and receive social signals. Understanding how and why non-verbal signals are coordinated between people remains unclear and limits our ability to theorise about the underlying cognitive mechanisms. One valuable approach to this problem is the study of interpersonal coordination, which draws on a long tradition that aims to quantify and understand face-to-face conversations (Argyle & Trower, 1979). Given the large variety of signals that each modality can produce, in this study we limit our investigation to the meaning of head nodding and its use in face-to-face conversations.

2.2.1 Head Nodding as a Social Signal

A long tradition of research into human interpersonal coordination describes patterns of synchrony or mimicry as integral non-verbal behaviours in dyadic social interaction (Bernieri et al., 1988). From early work (Chartrand & Bargh, 1999; Kendon, 1970) to more recent studies (Ramseyer & Tschacher, 2011) it has been demonstrated that people coordinate their heads and bodies with one another during conversations in various ways, including synchronizing rhythms (Richardson & Dale, 2005), structuring their turn-taking behaviour (Duncan & Fiske, 1977), and assuming complementary roles (Garrod & Pickering, 2004).

Many non-verbal signals during a conversation are centred around the head (e.g., eye-gaze, blinks, facial expressions, and head movements). It is not surprising then that listeners' attention is drawn to the speaker's head and face during conversation

(Argyle & Cook, 1976). Head nodding has been studied extensively because of its central role in conversations as an important source of social information (Birdwhistell, 1970). Head nodding is also regarded as a distinct social signal that is particularly sensitive to conversational demands and can convey a lot of different meanings (Poggi et al., 2010), from signalling attention and understanding (Hadar et al., 1983), to requests for information and passing turns (Duncan, 1972).

Recent work from our lab has developed an automated method which can identify and quantify two distinct types of nods – fast nods and slow nods (Hale et al., 2020). This quantification is useful to allow us to build virtual characters who show these behaviours and may allow us to measure nodding in different people to quantify features of an interaction (e.g., affiliation, liking, interest) or as part of a clinical assessment (Pan & Hamilton, 2018). For such use of nodding to be meaningful, it is important to have a robust understanding of when and why people engage in fast nodding and slow nodding behaviours, that is, what social signals are being sent? It is also important to know if there are robust individual differences in nodding behaviour, that is, could nodding be useful as a clinical indicator?

To address these two questions, it can be useful to consider the impact that social context can have on nodding behaviour. If the amount of nodding someone engages in is to be used as a clinical measure, it should be robust across different conversational contexts and consistent within an individual. If nodding is a social signal, then the meaning of the signal can be inferred by how nodding changes across such contexts. Here we will consider three potential meanings of a head nod: (1) a head nod can act as a communication backchannel, or a feedback signal from the listener in a conversation, (2) it can be the result of joint attention or simple gaze

following, and (3) it can be that we use it to mimic each other, which in turn acts as a 'social glue' to facilitate bonding and affiliation (Lakin et al., 2003).

Backchannelling. Descriptions of conversation behaviour often distinguish between the primary channel of information, from the speaker to the listener, and a secondary channel or backchannel where information flows from listener to speaker. There are both verbal and non-verbal backchanneling signals. Verbal backchannels are often represented as linguistic vocalizations such as 'uh-uh' (Sacks et al., 1974), and in the visual modality they are usually associated with certain eye-contact, or a smile. A non-verbal example of a backchannel is a head nod. Throughout a conversation, the listener may nod their head to show that they are listening, or even to indicate that one is agreeing with what the speaker is saying (Allwood & Cerrato, 2003; Duncan, 1972). In previous research (Hale et al. 2020), results show that a high frequency fast nod is produced mainly when participants are listening, which suggests this is likely to be a backchannel. This is relevant because it can let us test different hypotheses about the meaning behind these different head nodding signals, for example "What does fast nodding mean?".

Joint Attention. A head nod could also be the consequence of joint attention or following someone's gaze (Richardson et al., 2007). For example, as one person looks down at an object, the other person either follows the downward direction of that person's gaze or attends toward the same object by looking down. More specifically, joint attention has been defined as the capacity to coordinate attention with a social partner and can either be regarded as attending toward the same direction, or toward the same object or event, that another person is attending (Emery, 2000). A head nod then, much like a head turn, might just be the product of coordinating gaze with one's conversation partner. It can therefore be important to

distinguish between a backchannel signal and the behaviour of joint attention because they could have different social meanings.

Mimicry. A head nod could also be a copy of someone else's head nod. Humans mimic each other in a variety of ways (e.g., facial expressions, gestures, moods). This imitation, termed behavioural mimicry (Lakin & Chartrand, 2003), arises spontaneously during a social interaction, for example one person touches her hair and the other does the same shortly after. Mimicry differs from more goal-directed imitation (Hale & Hamilton, 2016a) where one person might copy to learn a skill or achieve a particular goal. It is further believed that this behaviour can act as a 'social glue' to facilitate bonding and affiliation between people (Lakin et al., 2003).

How can we tell if the head nod of person A turned their attention in the downward direction of the gaze of person B, or if person A was mimicking the downward head movement of B? In other words, how can we differentiate between simple gaze following and genuine mimicry? In addition, how can we separate mimicry or joint attention from backchannel behaviour? These are the type of questions we are interested in answering in this study, and the way we differentiate between these behaviours is using social context. We created three tasks with different conversational context, which allowed us to manipulate the structure of the turn-taking behaviour between people to be either structured or unstructured, and how information is shared between people. To disentangle the different meanings behind head nodding behaviour, we used high resolution motion capture and advanced signal analysis to measure the interpersonal coordination of head nods between two people engaged in naturalistic conversation. In the following section, we review different methods that have been used leading up to this study.

2.2.2 Methods for Measuring Interpersonal Coordination

As experiments in social neuroscience are striving toward more ecologically valid environments using naturalistic stimuli and paradigms, it becomes more and more important to capture the richness of interaction in a manner which preserves the dynamic nature of multimodal non-verbal signals. Traditional cognitive approaches to understanding social interaction involve studying isolated participants responding to stimuli on a computer screen or interacting with confederates. Such designs tightly control the variables involved and isolate targeted behavioural or cognitive constructs. However, it is increasingly recognized that this is a poor model of real-world social interaction, as in many ways these studies do not resemble the complexity and dynamic nature of stimuli and behaviours in real life (Krakauer et al., 2017; Risko et al., 2016). Psychologists and neuroscientists are now trying to understand dynamic social interaction, in which two people respond to each other in real time (Heerey, 2015; Schilbach et al., 2013). Real world conversations involve rich multimodal and dynamic coordination between individuals in which the behaviour of each person is strongly interdependent.

Researchers have traditionally used several approaches to measure interpersonal coordination in dyads, but many are limited by time-consuming and low-resolution manual audio-visual annotation methods. Although these early experiments have been instrumental with detailed descriptive analyses revealing the fundamentals of social behaviour (Bernieri et al., 1988; Kendon, 1970), they have not been able to capture the full richness of a social interaction in a precise and quantifiable manner.

More recently, researchers have begun to use automated measures to calculate motion energy or other parameters from video (Fujiwara & Daibo, 2016; Paxton & Dale, 2013; Ramseyer & Tschacher, 2010; Schmidt et al., 2012). However, this

method lacks the resolution to capture 3D data as it is limited to a flat image of the people. Motion capture technology (Bouaziz et al., 2013) provides much higher resolution and has been employed in a few recent studies to study body movement in different scenarios (e.g., Feese, et al., 2011; Hale et al., 2019; Poppe et al., 2013). Automatic recording of behaviour like this can provide more objective and detailed data about an interaction. For example, with motion capture we can automatically extract facial expressions and detailed body movement between two people in real time. These developments are opening the way to new studies of social interaction, but also provides a new challenge in analysing it.

In the following section we discuss how further advances in wavelet analysis (Issartel et al., 2015; Schmidt et al., 2014) is used to understand social coordination in better detail – for example how it can help us to quantify the relationship between motion patterns to analyse interpersonal coordination.

2.2.3 Motion Capture Data Analysis

The data we get when capturing human movement and its temporal progression is often difficult to quantify and interpret. For example, the head can move in different directions (i.e., yaw, pitch, roll), and with high or low frequency and amplitude. In addition, our interpretation of the head's trajectory might become more complex when we have a long time-course with two people interacting. To make sense of a biological trajectory or behaviour over time we must analyse time-series data.

Wavelet Analysis is a spectrum analysis that transforms a time-series from the time-amplitude into the time-frequency space, which tells us what time a specific frequency component occurs (Morlet, 1983). This method is efficient in the study of

non-stationary (i.e., biological) signals that dynamically vary in frequency. It permits us to detect small changes and to analyse the temporal evolution of each frequency component in more detail. Cross-Wavelet Coherence (CWC) is a quantitative method in signal processing that allows us to consider the degree of similarity between two sets of data – or time-series, and the progression in time of this interaction (Grinsted et al., 2004). It can be useful to think of CWC as the correlation between two wavelet transforms. In general, high coherence is interpreted as a high degree of coordination because it indicates that two people are moving with the same frequency in the same time-window, but are not necessarily nodding synchronously (i.e., at exactly the same time).

CWC gives us two measures of an interaction – a coherence measure (R^2) which tells us if two people move at the same frequency within the same time-window, and a phase measure which tells us the precise temporal relationship (i.e., time lag) between them. These measures can, for example, indicate whether one person led or followed the other at a specific frequency and for how long (Issartel et al., 2015).

Proof-of-concept studies using cross-wavelet methods have found that it can be applied to time-series data on body movements in a dyadic interaction (Issartel et al., 2006; Varlet et al., 2011; Walton et al., 2015). Further studies have also revealed that interpersonal coordination occurs at multiple timescales (Schmidt et al., 2014), and frequencies (Fujiwara & Daibo, 2016). In the study by Fujiwara and Daibo (2016) they compared the amount of motion coordination, averaged over all frequency components, to the amount calculated from pseudo interactions where members of different dyads were randomly paired together. They split the data into two frequency ranges and found a high degree of coherence at low frequencies (<0.25 Hz) and less coherence at high frequencies (>4Hz). However, the authors did not test whether this

pattern was also present in data calculated from pseudo interactions *within* the same pairs. The studies mentioned have also not examined motion frequency in detail using motion capture to isolate the movement coordinates (i.e., pitch, yaw, roll).

In a recent study from our lab, Hale et al. (2020) revealed the importance of head nods as a mechanism for coordinating during conversation. Here they measured mimicry using single head motion sensors. A sample of 26 dyads (n=52) was engaged in a structured conversation where each participant took turns to describe a picture of a complex social scene to each other. They used CWC analysis to calculate levels of interpersonal coordination in real trials compared to a pseudo dataset created by matching data from different trials *within* the same pairs. What they found was different results for slow and fast head nods. The results showed a positive coherence at frequencies below 1.5 Hz, but also to their surprise, the real interactions showed lower than chance coordination of head movements at frequencies between 1.5 – 5 Hz. This unexpected finding suggests a systematic decoupling of head movements at higher frequencies. The authors hypothesize that these could potentially be different signals carrying distinct social information.

The authors further interpret these results in the framework of two types of coordination, which they refer to as ‘mimicry’ and ‘fast-nodding’. They refer to mimicry as events where two people perform the same movement shortly after each other (e.g., one person nods and the other nods shortly after). Fast nodding occurs when one person, typically the listener, makes a small high frequency nod, while the speaker typically does not. In the next section we present the current study. In this study we want to examine whether social context may cause changes in how the mimicry behaviours and fast nods are used. In turn this may tell us what these different social signals might mean for the people involved in the conversation.

2.3 The Present Study

The study described above (Hale et al., 2020) identified two specific social behaviours – mimicry of nods and fast-nodding – which were seen during one particular type of structured conversation. The aim of the present project is to identify what these social signals mean and how they are used across different conversational contexts. We chose to use three different conversational contexts which differ in how information is shared between people and how tightly the conversation is structured. First, we replicated the picture description task used by Hale et al. (2020). Second, we implemented a novel 'video discussion task' and an established 'meal planning task' (Chovil, 1991; Tschacher et al., 2014). These tasks allow us to manipulate the structure of the turn-taking behaviour between people to be either structured or unstructured, and how information is shared between people. In the following section, we will describe and compare these two variables across the three different contexts or tasks that were designed for this study.

2.3.1 Conversational Contexts

In the Picture Description Task (Figure 2-1A), participants were asked to take turns at describing a picture of a complex social scene to each other and later discuss its content. The task was divided into two parts. During the first part (monologue), the speaker describes the picture while the other participant just listens. During the second part (dialogue), both participants have a free conversation about the picture. For example, at this point the listener could start asking questions about the picture. Both parts of this task represent a structured conversation, which contains clearly defined turns of who is speaking and when. The dialogue phase is less structured

than the monologue, but even here the listener is most often asking questions about the content of the picture and waiting for the speaker to answer. Thus, this was a rather artificial and highly structured interaction, with one person speaking for most of the time. This slow alternation of turns is reflected in the sample turn-taking shown in the middle panel of Figure 2-1A. The Picture Description Task is also an example of a one-way information sharing context, where one participant had access to the picture and was sending information to the other participant.

Hale et al. (2020) have so far only examined coherence patterns during structured conversations of this type. In this context, participants naturally move their heads when alternating between looking down at the picture and at the other participant. This behaviour could reflect mimicry (i.e., copying) as a means to form a social connection (Lakin et al., 2003). However, it could also reflect joint attention behaviour because the speaker has an important gaze target (the picture in her hands), and the listener might follow her gaze towards the picture. Thus, for head nods in particular, the Picture Description Task is unable to discriminate between social mimicry and simple gaze following or joint attention towards the picture.

In preparation for the Video Discussion Task (Figure 2-1B), participants watch a children's cartoon (Roberts, 2011). The video was a 3-minute animation with no words, in which a drawn line creates obstacles for a character called DipDap. Our reasoning was to re-create a scene of remembering shared events with others. So later during the session after which the participants had been engaged in other tasks, both participants are instructed to recall the events of the video; they could freely discuss what happened and help each other remember as much detail as possible. This task allowed participants to have natural unstructured conversation with no exchange of novel information. Both participants watched the video together,

so they had access to the same information and had to work together to recall it in detail. Typically, this task involves a loosely structured conversation which may include long pauses, shown in the sample turn-taking behaviour in Figure 2-1B.

For the Meal Planning Task (Figure 2-1C), participants had five minutes to come up with a menu together, consisting of an appetiser, main course, and dessert. However, they could only use ingredients that they both dislike, which introduced a fun cooperative element. This task was a natural conversation which typically involves laughter, interruptions and overlapping speech (Figure 2-1C) in a loose structure. As both participants were sharing information about their own meal preferences, the task enabled a two-way exchange of information which must be dynamically regulated. Summarised in Figure 2-1, the tasks differ in the amount of structure imposed, the type of information exchange, and the use of objects.

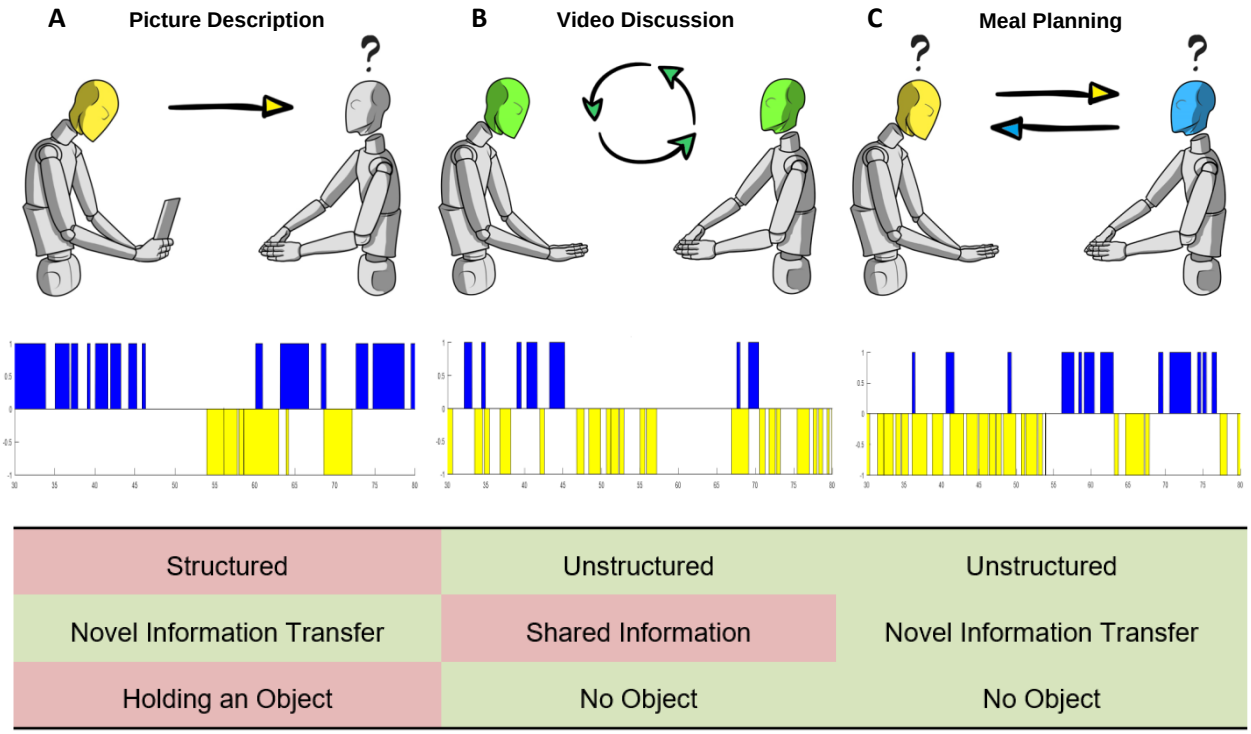


Figure 2-1. Conversational tasks. (A) Picture Description; structured one-way information transfer holding an object. (B) Video Discussion; unstructured shared recall. (C) Meal Planning; unstructured two-way information transfer. Graphs (middle) shows a sample of the turn-taking structure for each task in this experiment, highlighting the order and how often yellow and blue participants passed their turns.

2.3.2 Aims, Hypotheses and Predictions

In this study we are aiming to explore the social meaning behind head nods in naturalistic dyadic social interactions. We are looking to answer specific questions such as “What does fast nodding mean?”, and “What does slow nodding coherence mean?” across different social contexts. By varying the degree of (1) turn-taking, and (2) information sharing across three distinct conversational contexts, we can see if they differ in their patterns of coherence, which in turn will let us test different hypotheses about the meaning behind these head nodding signals. We will be testing four hypotheses related to the meaning of head nodding:

H_1: *Fast head nods are a signal of having received novel information.*

H_2: *Coherence of slow head nods is a signal of bonding and affiliation.*

H_3: *Coherence of slow head nods is a product of joint attention or gaze following.*

H_4: *Fast and slow head nods are stable features of a person and may be linked to personality traits.*

The first hypothesis (**H_1**) claims that fast nods are a signal of having received novel information and is based on the findings by Hale et al. (2020) that the coherence pattern of fast head nods could be a different social signal from that of the mimicry pattern associated with slow head nods. The authors based their hypothesis on results from motion analysis which showed more fast nods in the listeners compared to speakers, and more so when the speaker was verbalising. Testing this first hypothesis, we predicted (**P_1**) that fast nods should *change* across contexts. More specifically, If **H_1** is true, then we expect the coherence pattern associated with fast head nods to only be present in the Picture Description and Meal Planning

Tasks (Figure 2-1A, C) because these tasks both involve an element of novel information transfer, which promotes using a backchannel. In the Video Discussion Task (Figure 2-1B), participants have less reason to use a backchannel to signal anything because they already share the same information.

The second hypothesis (**H_2**) claims that coherent slow nodding signals bonding and affiliation and is based on the finding of coherent patterns of slow nods between two people, and the idea that these could be a form of social mimicry that acts as a 'social glue' to facilitate affiliation between people (Lakin et al., 2003). Testing the second hypothesis, we predicted (**P_2**) that the emotional attitudes of bonding and affiliation should be *similar* across the different tasks. More specifically, if **H_2** is true, then the positive coherence pattern associated with slow head nods should be present across all tasks, because the motivation to form a social bond should be equally present across all three conversational contexts.

In contrast, the third hypothesis (**H_3**) claims that coherent slow nodding is not related to behavioural mimicry, but rather the joint attention or simple gaze following behaviour between participants. Distinguishing between **H_2** and **H_3** will help us determine the meaning of the slow nodding coherence pattern. This pattern arises in the specific context of the Picture Description Task where one person is holding a picture on their lap. This context encourages the speaker to alternate gaze up and down between the picture and the face of the listener; the listener could then share the speaker's attention by also gazing down at the picture. If (**H_3**) is true, this leads to the prediction (**P_3**) that slow nodding coherence should change across contexts. More specifically, we expected the positive coherence pattern associated with slow head nods to be present only in the Picture Description Task since the Video Discussion and Meal Planning Tasks do not have a picture to draw attention.

The fourth hypothesis (**H_4**) claims that both fast and slow nodding are stable features of a person and may be linked to personality traits. This hypothesis is based on the idea that some people might be more prolific head nodders and show strong nodding across all contexts, while others might show little nodding. In other words, it tests the possibility that head nodding patterns could be used as a measure of individual differences in social behaviour or could be related to personality factors. If the propensity to engage in nodding is a fixed personality trait that is stable across contexts, we might also expect these individual differences in nodding to relate to measures of social skill. Testing the fourth hypothesis, we predicted (**P_4**) that fast and slow head nods should be similar across tasks within individuals and should correlate with the questionnaire measures of social behaviour. That is, we expected the coherence patterns of slow and fast head nods to be linked to fixed traits and not change depending on the social context. If we identify consistent nodding within participants and across tasks, we will have evidence that nodding can be used to quantify individual differences.

2.4 Methods

2.4.1 Participants

62 participants (M_{age}=24) were recruited from the UCL Psychology Subject Pool and the ICN Subject Database. Exclusion criteria included subjects that were not fluent in English. All participants were recruited and tested in pairs (31 dyads) and were randomly paired to arrive at the same time. On arrival, participants were asked to remove eye-makeup, bulky clothes, and jewelry as to not interfere with the recording equipment. The participants did not have any previous experience with the tasks and were unaware of the purpose of the experiment. Ethical approval for video, audio, and motion capture recordings was arranged via the UCL Research Ethics committee, and all participants gave their written informed consent. A monetary reimbursement was offered for participating in the study at a rate of £7.50/hour.

2.4.2 Equipment

In the present study we performed multimodal recordings from 31 pairs of participants (dyads) engaged in three different tasks with varied conversational structures. Each person in a dyad was randomly assigned to be either the 'Yellow' or 'Blue' participant and sat one meter apart on stools. Audio instructions, together with audio cues indicating the start and stop of a recording, were given to the participants via two speakers placed on the floor next to them. The video stimuli (Roberts, 2011) were shown on a projector screen next to them. Two LED lights were also stationed next to the participants to better illuminate their facial features. Curtains separated the participants from the experimenter, who remained in the room, but did not interact. Behind the curtains we had three computers that coordinated the whole experiment (Figure 2-2A, B, C). In this setup, one computer (A) acted as a client, which sent commands to the two server computers (B, C).

Motion Capture. Dyads were recorded with high resolution motion capture (Optitrack, NaturalPoint Inc., v.1.10), consisting of eight cameras (4 × Prime 13 and 4 × Prime13W) at a sampling frequency of 120 Hz. The system tracks a participant's body movements by detecting the position and rotation coordinates of retroreflective markers placed on their body. Each participant wore an upper-body suit with a set of 25 pre-determinately placed markers using the system specific software Motive.

Eye and Face Tracking. Mobile eye-tracking headsets (Pupil Labs Inc.) were used to track participants gaze and facial expressions. Calibrated for both members prior to recording, the eye-tracker outputs 2D gaze direction at a sampling frequency of 120 Hz. To identify and analyse facial expressions we used the free and open-source face recognition software OpenFace (Baltrusaitis, Robinson, & Morency, 2016). We do not report any eye or face tracking data in this thesis.

Audio and Video Recordings. Alongside the non-verbal elements, wearable microphones connected to an audio mixer were attached to each participant's chest and used to collect the verbal component of the interaction. Each participant's voice was recorded on two separate channels of a single audio file using the Audacity software. A Logitech webcam was also used to record video of the whole session. We do not report any audio or video data in this thesis.

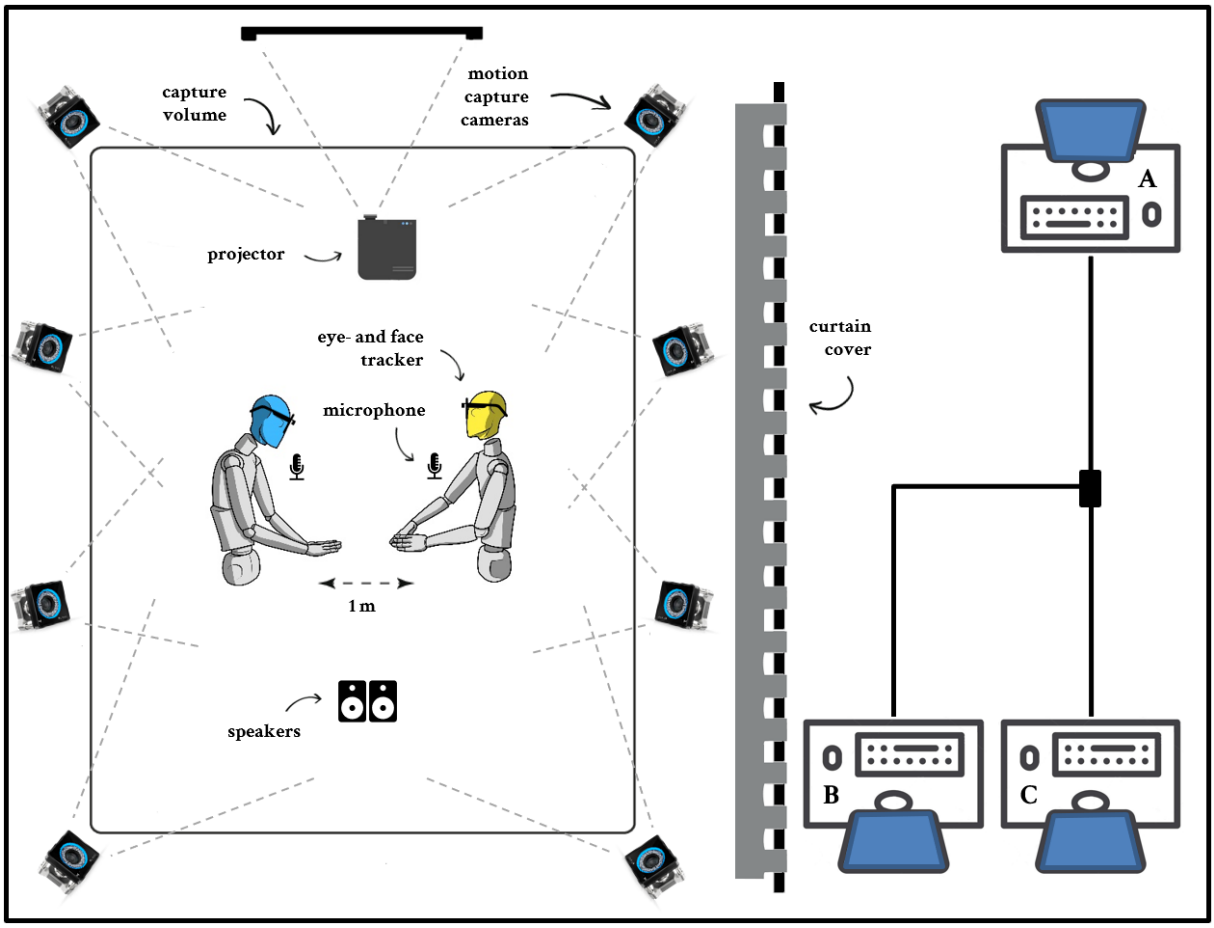


Figure 2-2. Lab setup. Equipment included motion tracking cameras (Optitrack Prime 13 & 13W), a projector, speakers, wearable microphones connected to an audio mixer, eye- and face trackers (Pupil Labs), and a curtain to separate the three computers running the experiment. Computer A acted as the client computer, that communicated with the two computers B and C acting as servers.

Recording Setup. The central requirement for multimodal data collection is the processing of multiple streams in a manner which preserves both the temporal and the spatial nature of the interaction. To capture all different data streams, we used

three computers (Figure 2-2A, B, C) running the recording software (Figure 2-3) with built-in socket compatibility across a local network. This was required to handle data being recorded from the various capture sources. This setup was designed to let us start and stop the different recording software within milliseconds of each other. To achieve this, we implemented a system-wide logging process that generated precise, machine-specific timestamps for each recording. We also instructed participants to perform a synchronised action (i.e., 3 hand claps) at the start of each phase of the task. The timing of these hand claps is visible in the audio, video, motion capture, and eye-tracker data, allowing us to check for synchronisation across all the different recording systems, and help localize certain events when analysing the data.

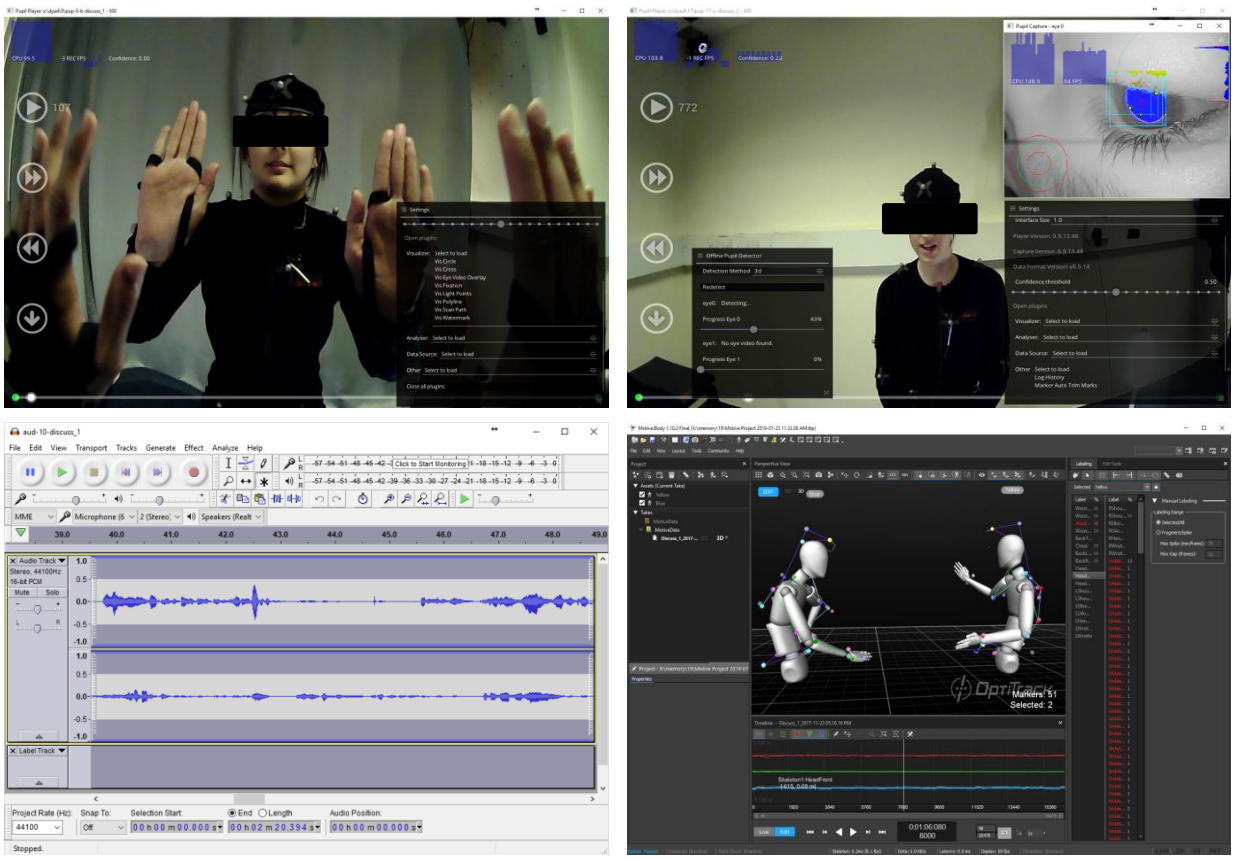


Figure 2-3. Recording software. *Top left and right*: Eye- and face tracking (Software: Pupil Player) from the perspective of the world camera on each participant; *Bottom left*: Voice recordings (Software: Audacity) from the wearable microphones; *Bottom right*: Motion capture (Software: Optitrack Motive) from the eight cameras tracking the retroreflective markers on the participants upper-body suits.

2.4.3 Procedure

Participants arrived at the lab and were shown all the equipment and informed of the procedures. They signed the informed consent, and then put on the motion capture suits, eye-trackers, and microphones. We then completed the calibration procedures for the motion capture and the eye-trackers. Each person in the dyad was randomly assigned to be either the 'Yellow' or 'Blue' participant, and sat one metre apart on small stools, before beginning the experimental tasks. As a pre-task in each recording block, the participants watched the DipDap video together. After watching the video, participants engaged in three tasks, each repeated once except for the third, for a total of five recordings per dyad (see Figure 2-4 for details).

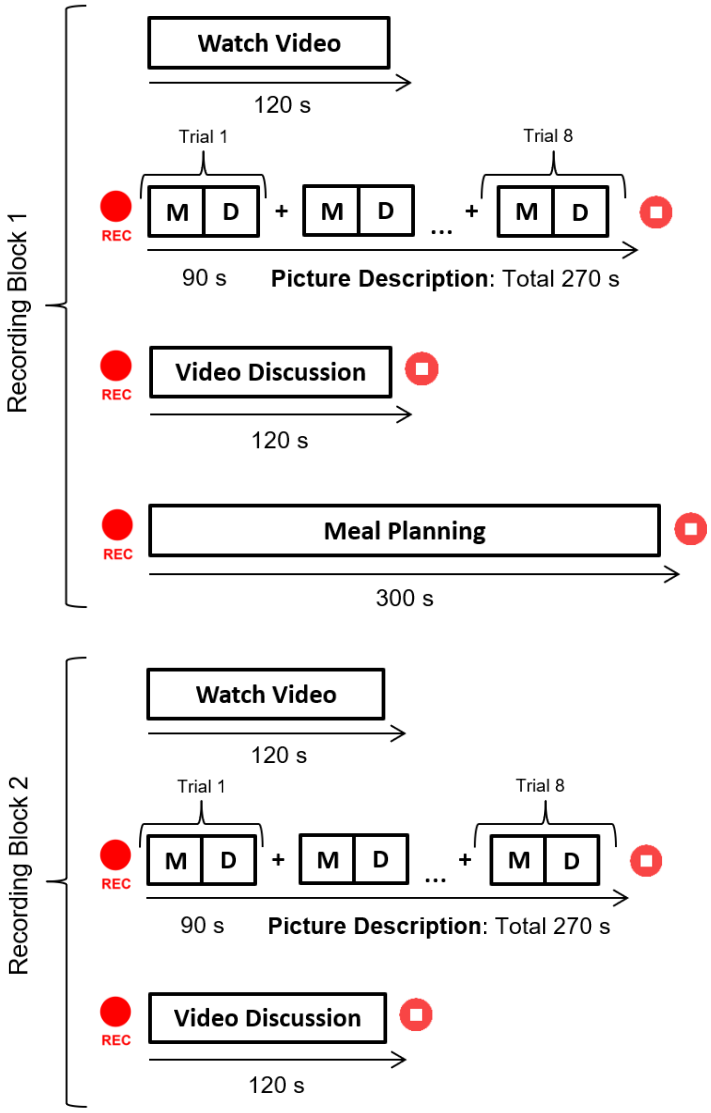


Figure 2-4. Task order and sequencing. The procedure consisted of a pre-task (watch video), followed by three tasks (Block 1). The Picture Description and Video Discussion tasks completed twice (Block 2), whereas the Meal Planning Task completed only once. This was due to its unrepeatable design in that participants by the end of the first block were familiar with each other's meal preferences. Repeating two of the tasks once provided us with the opportunity to collect more data and discover potential effects caused by familiarity and assess test-retest reliability. There was a total of five recordings per dyad.

Picture Description Task. Participants completed the Picture Description Task (Figure 2-1A) adapted from earlier behavioral studies (Chartrand & Bargh, 1999). This task is a form of one-way information sharing in which the participants were asked to take turns at describing a picture to each other and later discuss its content. Each trial was divided into two parts. During the first part (monologue), the speaker held a picture of a complex social scene and were instructed to describe it for 45 seconds while the other participant just listened. During the second part (dialogue), both participants were instructed to have a free conversation about the picture for another 45 seconds. At this point the listener could start asking questions about what had been described to them. Audible cues signaled the start and end of each trial, as well as the transition from monologue to dialogue. All dyads completed 8 trials in each of the two blocks for a total of 16 trials, taking turns in the role of speaker.

Video Discussion Task. Second, participants completed a Video Discussion Task (Figure 2-1B), in which they were instructed to recall a short three-minute video of a kid's cartoon (Roberts, 2011) that they had seen previously, and later discuss its content. It is a non-verbal short animation about a character called DipDap. This task could be described as unstructured shared recall where both participants had a free conversation for two minutes about the content of the video, and where they could help each other remember as much details as possible of what happened. This task does not require sharing new information between the participants. All dyads completed a single two-minute trial in each of the two blocks, for a total of 2 trials.

Meal Planning Task. Lastly, participants completed the Meal Planning Task (Figure 2-1C), based on Chovil (1991), and recently adapted by Tschacher et al. (2014). In this task the participants have five minutes to come up with a menu together consisting of an appetizer, main course, and dessert. However, they can only use ingredients that they both dislike, which introduces a fun cooperative element to the conversation. Like the Video Discussion Task, this is an unstructured conversation, but with two-way information sharing or joint planning. Participants completed a single five-minute trial in the second recording block.

Questionnaires. After the tasks, the participants completed four questionnaires. The Liebowitz Social Anxiety Scale (LSAS) (Liebowitz, 1987) assesses the range of social interactions that creates social anxiety. The scale features 24 questions relating to performance anxiety in social situations (e.g., “Eating in public places”). The Toronto Alexithymia Scale (TAS-20) (Taylor, Ryan, & Bagby, 1986), is a measure of deficiency in understanding, processing, or describing emotions. It has 20 statements (e.g., “I often don’t know why I’m angry”) rated on a five-point Likert scale. The Adult Autism Spectrum Quotient (AQ) (Woodbury-Smith, Robinson, Wheelwright, & Baron-Cohen, 2005) aims to assess whether adults of average intelligence have features of various autism spectrum conditions. It consists of 50 statements (e.g., “It does not upset me if my daily routine is disturbed”), each of which is a forced choice between “definitely agree or disagree”, “slightly agree or disagree”. The final one was the Experience of Gaze Questionnaire (unvalidated), which was created by two colleagues, Roser Cañigueral and Paul Forbes, measuring the participants subjective experience of eye contact. It consists of 20 statements (e.g., “I need to think about whether or not to make eye-contact”), with a forced choice on a five-point scale between “strongly agree” to “strongly disagree”.

2.5 Data Analysis

By selecting the pre-defined marker-set “25 Upper-Body” in the Motive software, the 25 retro-reflective markers auto-label and divide the skeleton model into different bone segments (head, torso, etc.). We selected the head bone-segment of each skeleton model (Figure 2-5), which gave us three data channels specifying its position in x, y, and z coordinates, as well as three channels specifying its rotation (yaw, pitch, roll) in degrees. The rotation signals roughly correspond to head turning, nodding, and tilting. Each channel was recorded at a sample frequency of 120 Hz.

In this study, we are interested in answering specific questions relating to the slow and fast head nodding behavior during dyadic social interaction, and we focus solely on exporting the head pitch data (i.e., degrees of rotation in the y-plane). Besides fitting our research questions, this rotation signal is interesting because it has been shown to be the most informative signal based on data from Hale et al. (2020).

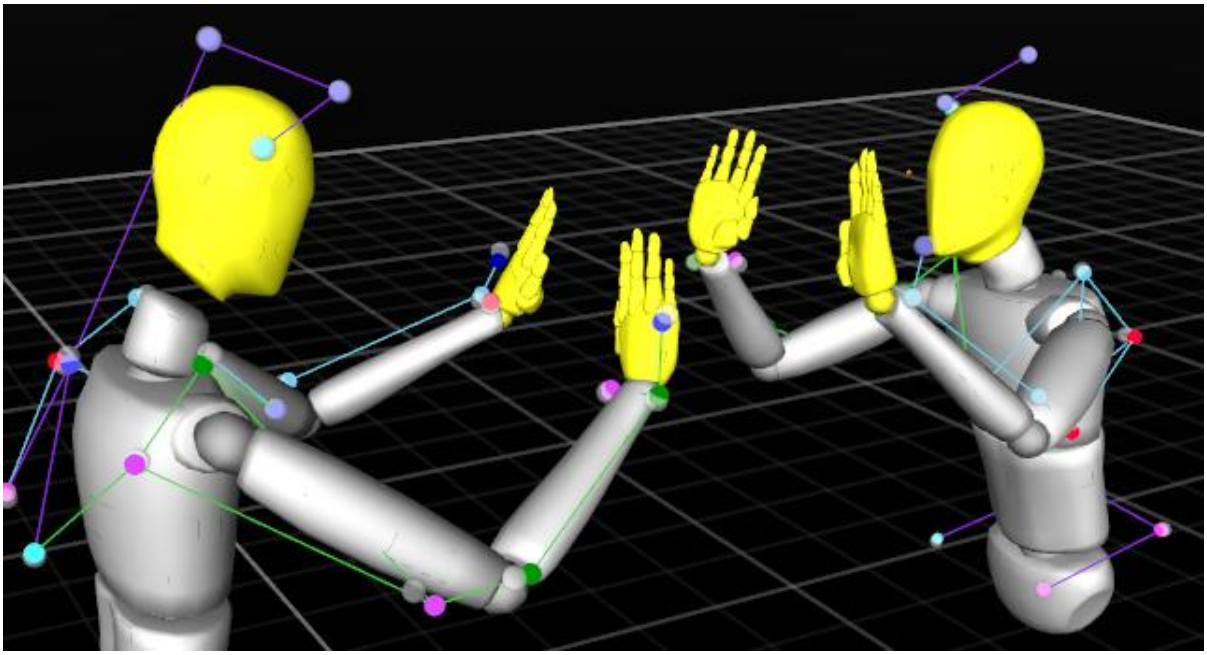


Figure 2-5. Motion capture data format. The head and hand bone-segments for each participant are highlighted in yellow. The raw head-pitch signal was exported from the center of each head bone-segment. Auto-labelled marker positions are color-coded according to left and right body parts.

2.5.1 Data Pre-Processing

A loss of marker tracking or incidents of misidentification where certain markers swap labels with each other will often result in jerky movements of the skeleton models. In such cases, we manually labelled these markers to correct this issue and ensure high quality data. We used the 'quick-label' mode in Motive with default settings. In specific cases of extensive marker swapping, we also used the edit tools to implement linear interpolations between the affected markers. In a minority of trials where the edit tools could not fix the loss of marker tracking, the wavelet toolbox in Matlab was unable to calculate the wavelet transform. Such trials were excluded from all analyses and reported as missing data (28/248 picture description trials; 5/62 video discussion trials; and 3/31 meal planning trials). The final head pitch signal ranged between -180 and 180 degrees, with no instances of a sudden switch from -180 to 180 found in the data. Thus, there was no need for circularity correction.

2.5.2 Cross-Wavelet Coherence Analysis

We carried out the following pipeline (Figure 2-6) using the Matlab Toolbox from Grinsted et al. (2004) to identify the wavelet power in the head pitch signals and to calculate cross-wavelet coherence. The input to this analysis is the raw head pitch trajectories for both participants with the head rotation in the y-plane in degrees (y-axis) as a function of time (x-axis) (A, B). We calculated the wavelet transform for each trial to get the time-frequency representation of each time-series (C, D). In total there were 133 wavelet scales using the Morlet wavelet (periods ranged from 0.1 – 19 Hz on the sampled data). Next, we calculated the cross-wavelet coherence between each of the two wavelet transforms (E), which gave us a measure of the time-frequency coordination between their movements.

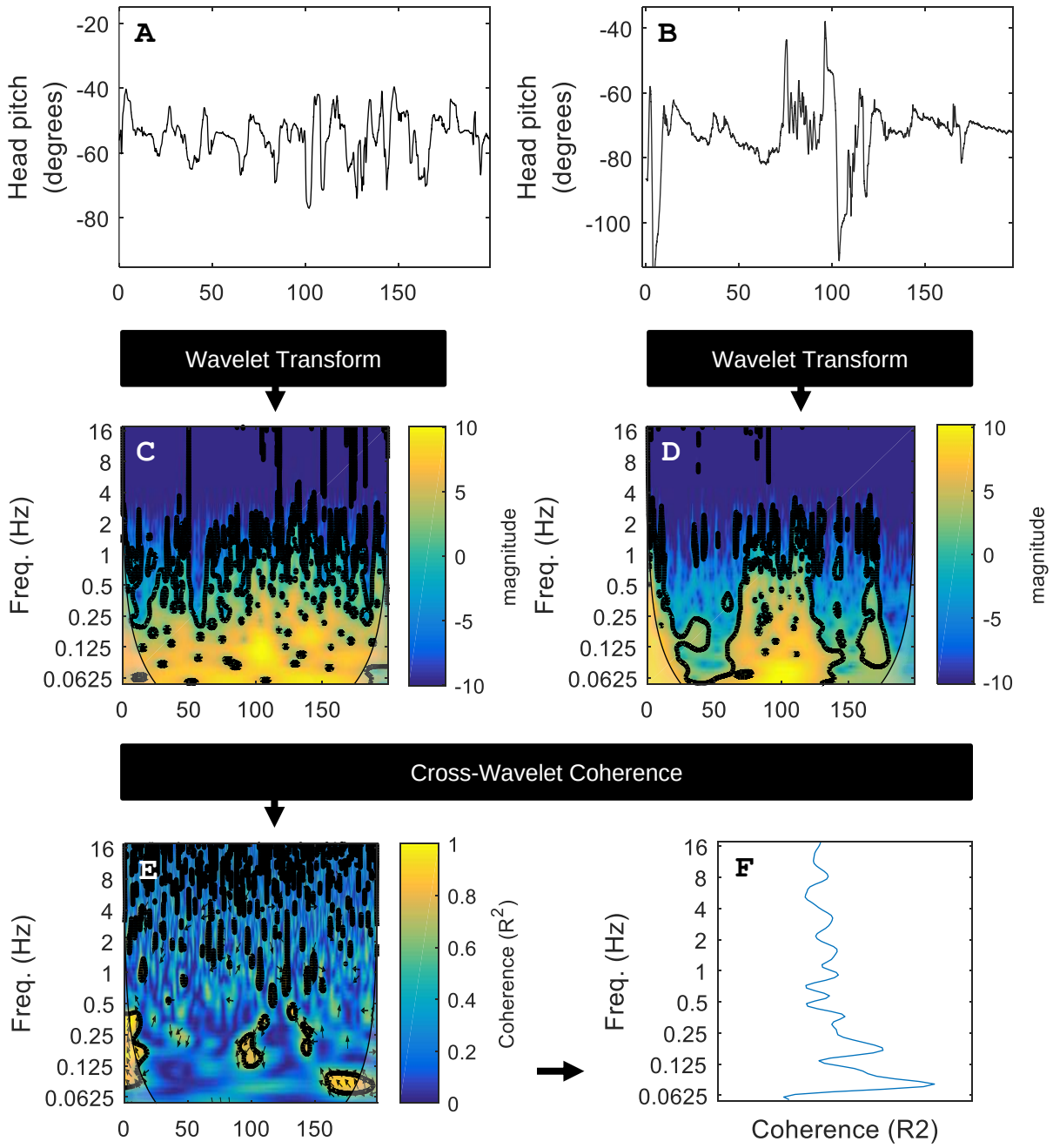


Figure 2-6. Cross-wavelet analysis pipeline. For each trial, the head pitch trajectories for both the Yellow and Blue participants (A, B) are subject to a wavelet transform (C, D). Then, the cross-wavelet coherence is calculated between the two participants (E). The magnitude of wavelet power and wavelet coherence is represented by color, where blue is low power, and yellow is high power. The time is represented on the x-axis (200s) and each frequency on the y-axis (0-16 Hz). The coherence value (R^2) is then averaged over time (F) and over all trials to obtain the overall frequency of coherence in head pitch between the two participants.

To ensure that the analysis was free from the influence of edge effects (influence on the wavelet from the discontinuities at the start and end of recordings), we

calculated the ‘cone of influence’ (COI) (opaque grey areas in the corners of Figure 2-6C, D, E) and zeroed any data outside it. We also applied cone-of-influencing zeroing around the monologue-to-dialogue transition in the Picture Description Task, this helped to minimize the influence of stimuli outside the dyad. We discarded data that was outside the 0.1 – 19 Hz range. In the final step, we averaged the cross-wavelet coherence (R^2) over the time-course of each trial to obtain a measure of the frequency of coherence without regard to the specific time at which it occurred (F).

2.5.3 Interpersonal Coherence in Real vs. Pseudo Interactions

To understand the patterns of head movement present in each task, it is helpful to compare the wavelet coherence values from the dataset to a baseline. A good measure or baseline test of interpersonal coordination is to compare coherence in real trials, where the two datasets come from the same interaction, with coherence in pseudo trials where the two datasets come from different interactions (Bernieri & Rosenthal, 1991; Fujiwara & Diabo, 2016). Earlier studies that have used this approach have created pseudo trials by matching datasets from different participants. A previous study from our lab has used a more rigorous approach by matching datasets from different trials *within* the same dyad (Hale et al., 2020).

We adopt the same approach here, where we match up the yellow participant’s signal from one trial with the blue participant’s signal from a different trial within the same dyad (Table 2-1). This gives us a strong test where the pseudo trials have the same general movement characteristics as our real trials, and any differences in the coherence levels between them must be due to a genuine live social interaction and will not be attributed to any individual differences between them.

Table 2-1

Example of generating pseudo trials for the picture description task.

Block-Trial	True Match		Pseudo Match 1		Pseudo Match 2		Pseudo Match 3	
1-1	1 S	1 L	1 S	3 L	1 S	5 L	1 S	7 L
1-2	2 L	2 S	2 L	4 S	2 L	6 S	2 L	8 S
1-3	3 S	3 L	3 S	1 L	3 S	5 L	3 S	7 L
1-4	4 L	4 S	4 L	2 S	4 L	6 S	4 L	8 S
2-5	5 S	5 L	5 S	1 L	5 S	3 L	5 S	7 L
2-6	6 L	6 S	6 L	2 S	6 L	4 S	6 L	8 S
2-7	7 S	7 L	7 S	1 L	7 S	3 L	7 S	5 L
2-8	8 L	8 S	8 L	2 S	8 L	4 S	8 L	6 S

Note. S = Speaker (Strong Colour); L = Listener (Light Colour).

In the Picture Description Task, each dyad completed 2 blocks of 4 trials, alternating between speaking and listening. For each real trial, we end up with 3 pseudo trial combinations, resulting in 24 pseudo trials per dyad (Table 2-1). In the Video Discussion Task, we had 2 blocks of 1 trial each, so we counterbalanced the two existing trials. In the Meal Planning Task, we had 1 block with 1 trial. To be consistent, we split the trial into two equal segments and counterbalanced the two.

We carried out wavelet analysis on the pseudo-data using the same pipeline as in the real trials. This gave us a set of coherence values for each real and pseudo trial of each dyad. Separately for the real dataset and the pseudo dataset, we averaged the coherence values across all trials for all dyads. We then calculated a coherence difference for each dyad, representing the average coherence in real interactions minus the average coherence in pseudo interactions, and performed t-tests on the coherence differences at each frequency (90 tests, one for each wavelet scale). To correct for multiple comparisons, we used a False Detection Rate (FDR) of 0.05 (Benjamin & Hochberg, 1995). By comparing real and pseudo trials in this way, we can see if interpersonal coordination that occurs in real conversations is different from the same people just speaking.

Following this, we aimed to define parameters to represent the fast and slow nodding behaviour in each dyad. Fast nodding is a behaviour shown by an individual, so we defined and quantified fast-nodding as the mean wavelet power for each individual participant in the frequency range of 2.6 – 6.5 Hz. However, to avoid circular analysis, or “double dipping”, we chose these frequency ranges from data presented by Hale and colleagues (2020). This is an individual measure equivalent to averaging the values in a horizontal band spanning the full time-range and the frequencies from 2.6 – 6.5 Hz on the two individually wavelet-transformed signals in Figure 2-6C (i.e., blue participant) and 2-6D (i.e., yellow participant). Thus, we obtain a single value for each participant which represents how much that person engages in fast-nodding behaviour. To emphasize, this is not a coherence measure, because fast nodding differs between speakers and listeners.

In contrast, coherence of slow nodding is a property of pairs of participants and emerges only in the analysis of the dyadic interaction. Thus, to characterise and quantify slow nodding, we calculated the mean wavelet coherence in the frequency range of 0.2 – 1.1 Hz. Again, these values were based on data from Hale and colleagues (2020) to avoid circular analysis. This is a dyadic measure equivalent to averaging all the values in a horizontal band spanning the full time-range and the frequencies form 0.2 – 1.1 Hz in Figure 2-6E. Thus, we obtain a single value for each dyad showing how much that dyad engages in coherent slow nods. Coherence (R^2) ranges on a scale of 0 to 1. We used these quantifications of fast and slow nodding for the next analyses.

2.5.4 Self-Report Questionnaires

First, we aimed to test if the tendency to engage in fast or slow nodding behaviour is a fixed personality trait that differs between people and is consistent across tasks. Thus, we correlated the fast-nodding score across tasks for each participant. Similarly, we aimed to test if the tendency to show coherence slow nods is a stable characteristic of dyads that is consistent across tasks, so we correlated the slow nod scores across tasks for each dyad. Finally, we calculated if either fast nodding scores or slow nodding scores were related to the subjective reports from four questionnaires at the end of the experiment.

2.5.5 Methods Summary

In the present study, the overall aim was to explore the temporal characteristics of head nodding in naturalistic dyadic interaction in different contexts to uncover the meaning behind the signal. Understanding the timing and at which frequencies head nodding behaviours occur can help us answer how and why we use these signals. So far it has been difficult to make firm predictions about the frequencies at which people naturally coordinate, both from a methodological and a design perspective. We set out to implement high resolution multimodal data capture along with sophisticated analysis methods to understand the fine-grained temporal patterns in relation to different hypotheses about the meaning of the head nodding signals.

Each dyad was engaged in different conversations across three tasks with varying structures of turn-taking and information sharing behaviour between the participants. These contexts were then analysed and compared to discover any differences in their pattern of coherence. From the pre-processed data, we calculated the wavelet

transforms (Figure 2-6C, D), where each capture of the participants head pitch (nod) is represented in terms of a wavelet in different frequencies and time points. Next, we calculated the cross-wavelet coherence (Figure 2-6E), which is a quantitative method in signal processing that enables us to compare specific characteristics from the interaction of two time-series, and the progression in time of this interaction (Grinsted et al., 2004). The coherence value (R^2) was averaged over time and over all trials. As a baseline test to the true interactions, we analysed the cross-wavelet coherence for pseudo interactions by shuffling the trials within each dyad. This test controls for any variables that may depend on individual differences and reveals features of movement that are specific to a genuine interaction.

At the end of the experiment, data from four questionnaires were collected. We analysed these by first testing if participants show a reliable pattern of fast or slow nodding by correlating these across tasks. We also tested if the tendency to nod is related to the personality traits measured in the questionnaires by performing correlations between these measures and relevant frequency bands of the wavelet data (i.e., high frequency fast nods and low frequency slow nods) for each task.

2.6 Results

2.6.1 Cross-Wavelet Coherence in Real vs. Pseudo Interactions

Results are shown separately for each task or conversational context (Figure 2-7). Graphs A, B and C show the mean and standard error of coherence (R^2) for real (red) and pseudo (blue) interactions. High coherence means a high degree of coordination, as it indicates that two people are moving with the same frequency. To

assess the difference in coherence between real and pseudo interactions, we performed t-tests (90 tests) at each frequency and calculated the effect size. Graphs D, E and F show the effect sizes (Cohen's d) calculated from the average coherence in real interactions minus the pseudo interactions. The dots indicate frequencies where there is a significant difference of coherence between real and pseudo interactions. Red dots represent points on the frequency range that pass a p<0.05 FDR significance threshold, while blue represent significant differences that did not.

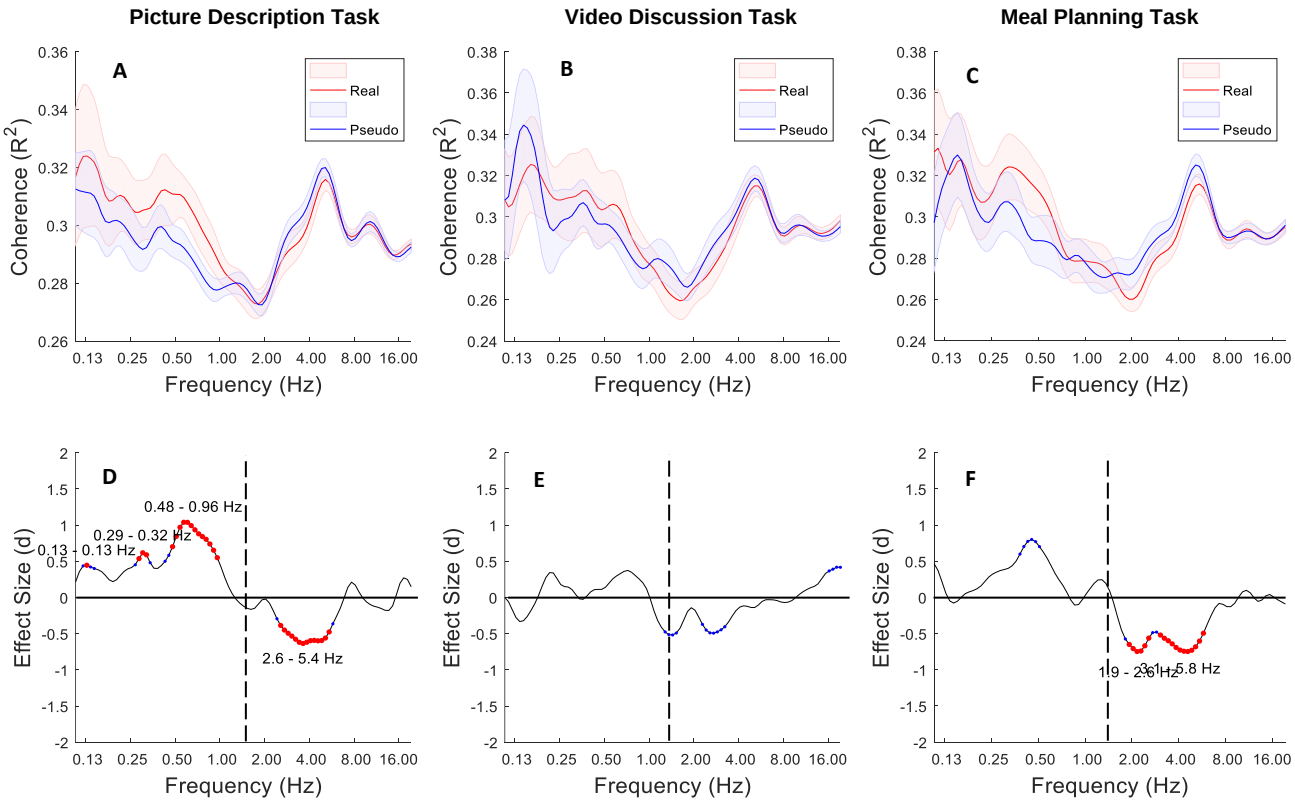


Figure 2-7. Cross-wavelet coherence. Graphs A, B, and C show the mean and standard error of coherence (R^2) for real and pseudo interactions across the three tasks. Graphs D, E, and F show the effect sizes (Cohen's d) for the difference between the real and the pseudo interactions. The dotted line indicates frequencies where there is a significant difference of coherence between real and pseudo interactions. Red dots represent points on the frequency range that pass a p<0.05 FDR threshold, while blue dots represent significant differences that did not pass this threshold.

On the above graphs we can observe two distinct patterns of coherence across the range of frequencies. These patterns are divided into two frequency ranges,

above and below 1.5 Hz, as indicated by the dashed vertical line (Figure 2-7D, E, F). In the low frequency range (<1.5 Hz) results show greater coherence in the real compared to the pseudo interactions for the Picture Description Task. However, this pattern was not observed in the Video Discussion and the Meal Planning Tasks. In the high frequency range (>1.5 Hz) results show less coherence in real compared to pseudo interactions in the Picture Description and Meal Planning Tasks but did not reach significant FDR corrected thresholds for the Video Discussion Task.

2.6.2 Individual Differences

We tested if the participants show a reliable pattern of fast or slow nodding by correlating these across tasks (Figure 2-8). The results show that there is no reliable positive relationship between fast nodding behaviour in any one task paired with any other task. There was also no reliable positive relationship between slow nodding coherence in any one task paired with any other task. There was a significant negative correlation, r=-0.43, p=.003, in slow nodding coherence in the Meal Planning and Picture Description Task, but it did not survive FDR correction.

In addition, we also tested if the tendency to nod is related to any of the personality traits measured in the questionnaires by performing correlations of the measures with relevant frequency bands of the wavelet data (high and low frequency nods) for each task separately (Figure 2-9). The questionnaires included the Liebowitz Social Anxiety Scale (Anx-Avoidance/Fear), the Toronto Alexithymia Scale (TAS), the Adult Autism Spectrum Quotient (AQ), and the Experience of Gaze Questionnaire (Gaze). The results show no correlations between the nodding measures and the questionnaire scores that survive a correction.

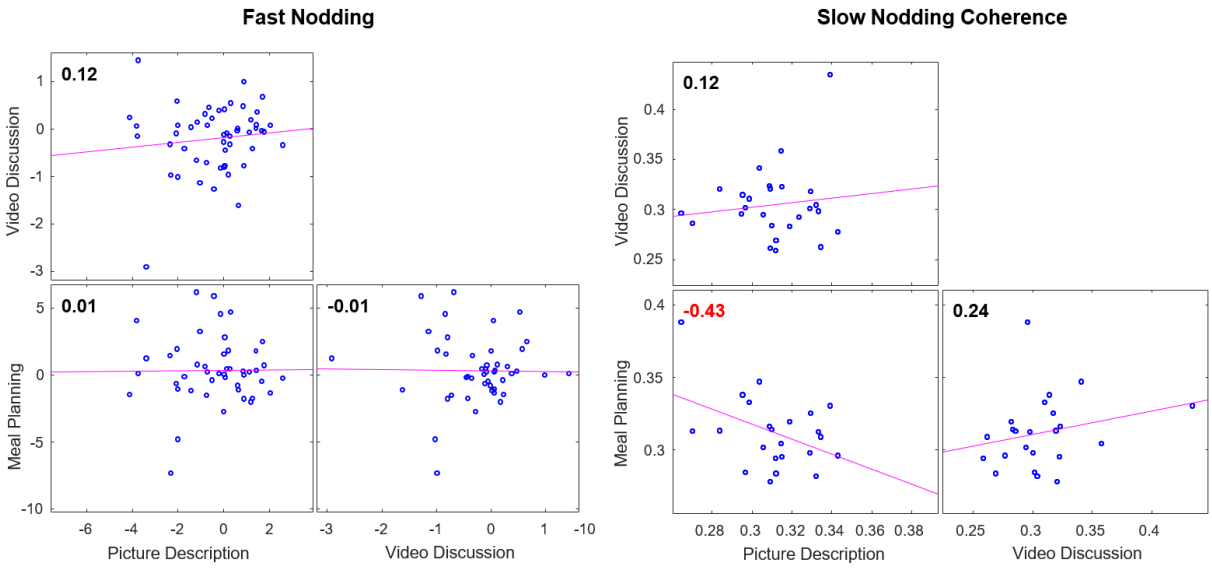


Figure 2-8. Across tasks correlation matrices. Red correlation scores indicate if the correlation is significantly (p<0.05) different from zero. The axis values for fast nods is the average power in the 2.6 – 6.5 Hz frequency band at the individual level (one data point for each participant). The axis values for slow nods is the degree of coherence (R^2) at the dyad level (one data point for each dyad).

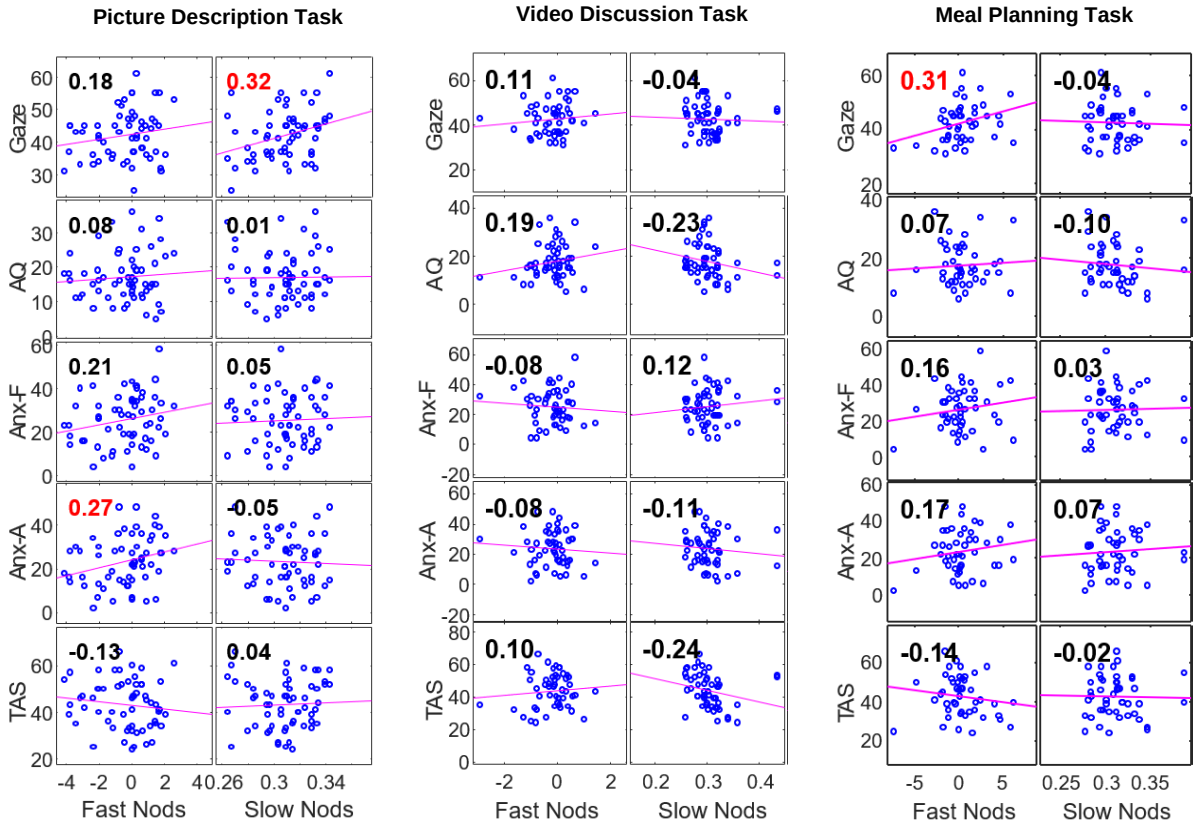


Figure 2-9. Within tasks correlation matrices. Red correlation scores indicate if the correlation is significantly (p<0.05) different from zero. The x-axis values for fast nods is the average power in the 2.6 – 6.5 Hz frequency band at the individual level (one data point for each participant). The x-axis values for slow nods is the degree of coherence (R^2) at the dyad level (one data point for each pair of participants). The y-axis values are the average questionnaire scores.

2.7 Discussion

Previous studies have identified patterns of fast nodding and slow nodding coherence in conversations but have not shown what this means. For example, recent work from our lab has shown that there are two distinct types of nodding behaviour – fast nodding and slow nodding (Hale et al., 2020). But for these measures to be meaningful and used to quantify features of an interaction, it is important that we understand why people engage in these behaviours. For instance, if nodding is a social signal, then the meaning of the signal can be inferred by how nodding changes across such contexts. Moreover, to use these nodding measures in clinical assessments, it is also important to know if there are individual differences between them. For example, if the amount of nodding someone engages in is to be used as a clinical measure, it should be robust across different conversational contexts as well as consistent within an individual.

This study tracked how these head nodding behaviours change across different conversational contexts, to understand the meaning of nodding as a social signal. More specifically, we considered three potential meanings behind a head nod: (1) a head nod can act as a communication back-channel, or a feedback signal from the listener in a conversation, (2) it can be the result of joint attention or simple gaze following, and (3) it can be that we use it to mimic each other, which in turn acts as a 'social glue' to facilitate bonding and affiliation (Lakin et al., 2003).

Consistent with previous work (Hale et al., 2020), we observe two distinct patterns of coherence across the range of frequencies. We define these as 'coherent slow nodding' (<1.5 Hz), and 'fast nodding' (>1.5 Hz). Examining the slow nodding behaviour, we find that dyads show greater coherence in the real interactions

compared to the pseudo interactions in the Picture Description Task. However, this pattern was not observed in the Video Discussion and Meal Planning Task. Examining the fast nodding behaviour, the results show less coherence in real compared to pseudo interactions in the Picture Description and Meal Planning Task but did not reach FDR corrected thresholds in the Video Discussion Task. We will proceed to explore the possible interpretations of these nodding patterns from the framework of the two frequency ranges, after which we will discuss the results relating to individual differences in nodding behaviour, followed by some methodological implications and limitations of the study.

2.7.1 Exploring the Coherent Slow Nodding Behaviour

We analysed slow nodding as the pattern of head pitch in the 0.2 – 1.1 Hz range. Our results reveal a positive above chance coherence of low frequency (i.e., slow) head nods between participants in the real vs. pseudo interactions during the Picture Description Task (Figure 2-7). This frequency range has traditionally been linked to behavioural mimicry (Chartrand & Bargh, 1999; Stel et al., 2009), which suggests that participants are mimicking each other's slow head nods during the Picture Description Task. However, the positive coherence does not mean that the participants are necessarily mimicking each other, but rather that one performs a head nod after which the other person nods within the time-window of the cross-wavelet analysis. This result is consistent with earlier studies that have used a similar paradigm (Fujiwara & Daibo, 2016; Schmidt et al., 2014), and effectively replicates the findings from Hale et al. (2020) using a higher resolution motion capture system. We did not observe any slow nodding coherence in the Video

Discussion and Meal Planning Tasks. This shows that slow nods changes across the different conversational contexts and cannot be generalized to all forms of social interaction. These results contradict the hypothesis (**H_2**) that slow head nods could be a form of social mimicry that facilitates bonding and affiliation between people (Lakin et al., 2003). Because if **H_2** were true, the coherence of slow nods should have been similar across the different conversational contexts due to the equal motivation to form social bonds during conversation.

Instead, these results rather seem to support our third hypothesis (**H_3**) that slow head nods are a product of joint attention or simple gaze following, which according to our prediction (**P_3**) arises in the specific context of the Picture Description Task where one person is holding a picture as a gaze target. This context encourages the speaker to alternate gaze up and down between the picture and the face of the listener; the listener could then share the speaker's attention by also gazing down at the picture. This is probably why we only see this behaviour in the Picture Description Task, and not in the Video Discussion and Meal Planning Tasks, because in these tasks there is no object to draw attention away. An exploratory analysis to further support this hypothesis is provided in the appendix to the thesis.

Based on this, we conclude that coherence of slow nodding is most likely not a general form of mimicry found across all contexts, but rather indicates that nodding is linked specifically to gaze targets. A prediction from this is that, if participants were in a context where a shared gaze target was located beside them, rather than in one person's hands, then we would instead see coherence of 'head shaking' as they turn their heads towards the target. It is possible to consider gaze following to be a subset of mimicry behaviour, in which the coherent head movements arising from following a person's gaze could be classified under a more general rubric of

‘interpersonal coordination’ or interaction. Indeed, some studies which score mimicry behaviour on observation of interpersonal coordination may not distinguish between gaze following and mimicry (Salazar-Kämpf et al., 2017). However, we suggest that it can be useful to make this distinction, because the two actions could have different social meanings. Gaze following is specific to the target of gaze (if an object is located on the left of A and on the right of B, then gaze following implies that A looks leftwards and B looks rightwards), whereas mimicry might be defined according to body-centred coordinates (I mimic a right-hand action with my right hand) (Liepelt, von Cramon, & Brass, 2008). This also shows the importance of understanding why people engage in specific head nodding behaviours, so we can figure out if the behaviour is used as a social signal, and if so what the exact meaning behind it is. Next, we will discuss the high frequency pattern of fast nodding.

2.7.2 Exploring the Fast Nodding Behaviour

In a secondary analysis by Hale et al. (2020), they demonstrated that the fast nodding behaviour is related to each participant’s speaking or listening behaviour by identifying a specific high frequency pattern between 2.6–6.5 Hz that is more prevalent when listening compared to speaking. This led them to the idea that fast nodding might be a backchannel signal related to listening. However, they were not able to properly test this idea. In the present study, we were able to test for the presence of fast nodding across different conversational contexts to determine if it really is being used as a backchannel to signal understanding. We analysed fast nodding as the pattern of head pitch in the 2.6–6.5 Hz range. Our results (Figure 2-7) reveal a less-than-chance coherence of high frequency fast head nods between

people in the real vs. pseudo interactions in the Picture Description Task and Meal Planning Task but did not reach significant FDR thresholds for the Video Discussion Task. These results positively replicate the results from Hale et al. (2020) in the task that they used (Picture Description Task), and we also found a similar fast nodding pattern in the Meal Planning Task, but not in the Video Discussion Task.

When we contrast the two hypotheses about the nature of the fast nodding behaviour, results from the real vs. pseudo comparisons (Figure 2-7) show that fast nods change across the different conversational contexts and cannot be generalized to all forms of social interaction. More specifically, these results support our hypothesis (**H_1**) that fast nodding are a signal of having received new information and that it is a different social signal to that of slow nodding coherence. In line with our prediction (**P_1**) fast nodding can only be observed in the Picture Description and Meal Planning Tasks, both of which involve a novel transfer of information.

From this we conclude that fast nodding most likely does not have the same meaning as slow nodding coherence and cannot be generalized to all forms of social interaction. Specifically, our results seem to indicate that fast nodding is a backchannel behaviour to signal that one has received new information. The Picture Description Task is a one-way information sharing context where the speaker is sharing new information to the listener about the picture. Similarly, the Meal Planning Task is a two-way information sharing context in which both participants are unaware of the other's meal preferences, while also having to share their own preferences. In both tasks, such exchange of new information promotes using a backchannel to signal to the other that you have received their message and are paying attention to what they are saying. On the other hand, the Video Discussion Task is about shared recall between members of a dyad and creates less motivation for them to use a

backchannel to signal that they have received any new information. This is probably because they both know they have acquired the same information regarding the video and believe that there is nothing meaningful to be learned from signalling when they have shared knowledge, or "common ground" (Clark, 1996). In this way, the lower-than-chance coherence pattern of fast nodding may only be present during conversations when there is a reason for people to signal that they have received new information and if they believe there is something meaningful to be learned.

Because both these tasks that promote using a backchannel involve an element of novel information transfer, we can begin to speculate as to whether changes in social signals can provide us with clues about our cognitive processes, such as predicting memory performance. At the very least, we could assess that fast nodding indeed seems to have a different meaning to that of coherent slow nods, and that these different social signals are used in different ways by the people involved.

2.7.3 Exploring Individual Differences in Nodding Behaviour

As an exploratory analysis, we used this dataset to test if the tendency to engage in fast or slow nodding behaviour is a fixed personality trait that differs between individuals. First, we tested if participants show a reliable pattern of fast or slow nodding by correlating these across tasks (Figure 2-8). That is, if a participant nods a lot in the Meal Planning Task, does that person also nod a lot in the Video Discussion and Picture Description Tasks? Second, we tested if the tendency to nod is related to any of the personality traits measured in questionnaires by correlating individual scores on fast nodding and coherent slow nodding with the questionnaire measures (Figure 2-9). If reliable individual differences in nodding behaviour could

be identified, this would motivate us to test in future studies if the tendency to nod reflects broader social skills. The four questionnaires included the Liebowitz Social Anxiety Scale (Anx-Avoidance/Fear), the Toronto Alexithymia Scale (TAS), the Adult Autism Spectrum Quotient (AQ), and the Experience of Gaze Questionnaire (Gaze).

In general, we did not find any evidence for reliable individual differences in nodding behaviour. Fast nodding behaviour in one task did not correlate with fast nodding in another task, nor did it correlate with any questionnaire measures. Slow nodding coherence in one task did not correlate with the same measure in a different task, and nor did it correlate with any questionnaire measures. Because we did not find any evidence of reliable individual differences, we did not further assess test-retest reliability for this analysis. This also means we can reject the hypothesis (**H_4**) that head nodding is linked to fixed personality traits, as it does not support our prediction (**P_4**) that fast and slow head nods should correlate with the subjective measures and show similar correlations between the tasks. Instead, our results rather seem to indicate that head nodding is dependent on the conversational context. The limitation here is that each person only appears in one dyad, so we are not able to quantify each person's behaviour independent of their interaction partner, as done by Salazar-Kämpf et al. (2017). However, at present there is no strong reason to use fast or slow nodding behaviour as a measure of an individual's social skills or as a clinical assessment. This is particularly relevant because studies are attempting to use automated analyses of interactive behaviour to identify and even diagnose disorders of social interaction such as autism (Georgescu et al. 2019). In the following section, we will discuss some limitations of the present study, and how future studies can extend the contribution.

2.8 Limitations and Future Directions

Limitations regarding data from the questionnaire measures presents a challenge for researchers who want to understand individual differences in social skills to identify the isolated contribution of everyone within a social interaction. This is difficult because data from naturalistic interactions are not independent but depend on the way people adapt their behaviours over time. Whereas social interactions have characteristics in addition to those of the individuals involved, the interaction might nonetheless be influenced by individual differences. In future studies, it is perhaps important to design questionnaire measures where such interactional parameters are sensitive to individual differences. Moreover, the data we have on this is based on a relatively small sample (i.e., 62 participants).

Secondly, we have no data on the temporal relationships between different movements, which makes us unable to make any inferences based on the timing between participants. As explained earlier on the data analysis for the motion capture data (Section 2.2.3), the Cross-Wavelet Coherence Analysis can give us two measures of an interaction – a coherence measure (R^2) which tells us if two people move at the same frequency within the same time-window, and a phase measure which tells us the precise temporal relationship (i.e., time lag) between them.

In addition to examining the coherence measure (R^2), Hale et al. (2020) also looked at the time lags using a simple cross-correlation measure between each participant's head pitch signal across a range of time lags to find a peak correlation of 0.6 seconds. This would indicate the timing at which the two participants matched each other's movements. This time lag data was backed up with further CWC-analysis of the phase relationships between the participants at every time-frequency

point, where they ended up with an optimal model revealing close to the same time lags as the cross-correlation measure (0.588 sec). This implies that the listener tends to match the head movements of the speaker with around 600 ms delay.

Initially we wanted to examine the timing of the frequency behaviour by performing cross-correlations between each participant's head pitch signal across all three conversational contexts to see if we got different results and to further strengthen our hypotheses. However, cross-correlation based lag analysis only makes sense if two people are, or are close to, moving in synchrony with one another – where one person's movement closely follows the other, like in a structured form of conversation with distinct turns between who is speaking and listening. Since the Video Discussion and Meal Planning Tasks are unstructured forms of conversation with a lot of overlap between who is speaking and listening, we considered using a method used by Tschacher et al. (2014) where they use the absolute values of each time-series. The use of absolute values means that both positive and negative cross-correlations contributes positively to the measure. This strategy yields values that are more representative of a dynamic unstructured dyadic interaction with more overlapping turn-taking behaviour. However, by using absolute values the researchers are essentially disregarding the sign of each value, giving different means and peak correlations. Moreover, the timing is also not as relevant for the fast nodding pattern as it is for the mimicry pattern because if one person is nodding quickly, the other is less likely to nod at all. For these reasons, we ultimately decided not to further investigate the temporal relationships using cross-correlations or phase analysis in this dataset. Such analysis is beyond the scope of this thesis.

We set out with the aim to improve upon the traditional approach of collecting unimodal data by implementing a multimodal setup that could capture dyadic social

interaction in higher resolution. This approach turned out to be a challenge with huge cognitive and computational demands. One of the difficulties in devising a multimodal data collection protocol is properly controlling for behavioural outcomes to ensure that the effects seen are not merely driven by effects created by the setup or the equipment itself. For example, care must be put into placing video stimuli and speakers so as not to influence where people look and move their heads. Having mobile eye-trackers equipped can also catch peoples gaze because it is not a natural thing to have on your face. Another disadvantage is that a setup like this is very obtrusive and requires specialized equipment and a dedicated recording space. Care when designing multimodal experiments are needed, but we believe that these more complex and contextually dependant setups will help to extend our knowledge of the interactive dynamics that regulate our social life.

This is not the first dataset to address multimodal dyadic social interaction. However, the inclusion of high-resolution motion capture, wide synchronization of socially relevant data and advanced analysis methods may provide us with many new opportunities to study social behaviour. In this study we created an integrated framework to account for the large variety of possible modalities that can interact to produce unique social signals in different contexts. This can be important for studies of the social brain and of disorders of social interaction by improving the automatic detection of social signals. It can also be important for creating computational models of realistic social behaviour, with the development of socially realistic virtual characters. In addition, we also implemented an analysis of comparing real interactions with pseudo interactions, renewed by Fujiwara and Daibo (2016), and improved upon by Hale et al. (2020). Here, we compare the real trials to a pseudo dataset created by matching randomized data from different trials *within* the same

pair of participants. By comparing real and pseudo trials in this way instead of *between* pairs, it enables us to observe if interpersonal coordination that occurs in real conversations between the same two participants is different from the same people just speaking without the context of the real interaction, they were involved in. This is a strong analysis that we believe should be used more often with data from dyadic interactions. For example, imagine that one dyad in the study is very energetic and the participants move about a lot, whereas in a second dyad both participants are quiet and mostly still. Comparing real trials within dyad to pseudo-trials across dyads might suggest a difference, but this might be driven only by the overall energy levels of the dyad and might not be specific to the social interaction. By using within-dyad pseudo-trials, we can control for the unique behaviour of each individual and identify only the coordination patterns which are specific to the live interaction of the participants.

A direction that we want to explore with the next study involves memory and learning during dyadic social interactions. In the present study, we have showed that fast nodding is represented more in conversational contexts in which there is a transfer of new information between participants. From these results, we can pose the question if fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received this information? And if so, could head nodding behaviour somehow be associated with memory? These are interesting questions since studying memory involves quantifiable outcomes from a conversation that we will try to test and measure using a new memory task. Memory and learning constitute a behavioural measure of acquired information and paired with our high resolution setup could provide useful insights into the natural parameters of head nodding in conversations, and its relationship to memory.

2.9 Conclusions

Previous studies have identified patterns of fast nodding and slow mimicry in conversation behaviour but have not shown what these could mean. With the use of high resolution motion capture and wavelet coherence analysis we tracked how these behaviours change across three different conversational contexts, in order to increase our understanding of the meaning of head nodding as a social signal.

First, we show supporting evidence that fast head nods are a signal of having received new information and that it is indeed a different signal to that of slow head nods in terms of facilitating the transfer of new information during conversations. Second, we show that slow head nods change across conversational contexts and find support for the hypothesis that slow nodding coherence are not a form of mimicry behaviour, but rather a consequence of having a shared gaze target, or simply following the gaze of another person. Third, we show that, in general, fast and slow head nods are not linked to stable personality traits that differ between people but are dependent on the conversational context. This implies that, at present, there is no strong reason to use fast or slow nodding behaviour as a measure of an individual's social skills or as a clinical assessment.

Our findings will be useful to explore if head nodding coordination might be related to measurable outcomes of a conversation, like memory and learning. It can also be used to measure nodding in different people to quantify features of an interaction (e.g., affiliation, liking, interest), and allow us to build virtual characters who show these behaviours. For such use of nodding to be meaningful, it is important to have a robust understanding of when and why people engage in fast nodding and slow nodding behaviours, which we have demonstrated with this study.

We try to demonstrate in this study that we need to continue to measure complex interactions between people in naturalistic settings and capture the coordination using new integrated frameworks. Still, there are problems with elaborate multimodal setups like this when it comes to controlling for all the behavioural outcomes. But with some extra care when designing multimodal experiments, we believe that these contextually dependant setups will help us capture social interaction in higher detail than before, and to extend our knowledge of the interactive dynamics that regulate our social life. Such knowledge could impact research on disorders of the social brain, automatic detection or sensing of social signals, and computational models of realistic or virtual social behaviour. It may even turn out to be more important in this new era of social distancing following the recent coronavirus (COVID-19) pandemic.

Chapter 3. Nodding Along as you Learn: Head Nodding in Conversation Predicts Memory?

3.1 Abstract

In our previous study (Chapter 2), we demonstrated that fast nodding behaviour is found more in conversational contexts in which there is a transfer of new information between participants. In this study we aim to investigate if encoding new information is associated with fast nodding. We collected a new dataset and tested if different types of head nodding frequencies during conversation could predict performance on an outcome measure in the form of a memory test. Using high-resolution motion capture, wavelet analysis, and multilevel modelling, we examine two hypotheses related to fast-nodding and slow-nodding behaviour respectively, namely if *encoding new information is associated with fast nodding behaviour* (**H_1**) and if *Self-other overlap is associated with slow nodding behaviour* (**H_2**).

Our results are ambiguous, depending on what statistical approach we implemented. We cannot conclusively claim that head nodding can be correlated with learning during conversations, but some analysis suggests that head nodding behaviours might be related to memory recall during unstructured conversations.

The findings of this study can provide useful insight into the natural parameters of head nodding behaviour and its relationship to memory. This can in turn allow us to build virtual agents who can simulate these natural backchannels to test how people respond to different types of interactions and drive the development of new and improved psychological theories of social interaction.

3.2 Introduction

In a recent study, Hale et al. (2020) developed an automated method which can identify and quantify two distinct types of nods – fast nodding and coherent slow nods. We used high resolution motion capture and wavelet analysis techniques to further explore the meaning behind these head nodding signals in different conversational contexts (Chapter 2). By shifting analysis away from the individual and widening the focus to the dyad, we began to explore the behavioural and cognitive dynamics that emerge from the contextual constraints of a dyadic interaction, such as the specific task or type of conversation.

Understanding the way in which people changed their behaviour in different social contexts provided us with important clues on how information is shared between people. We showed supporting evidence that fast head nods are a signal of having received novel information, as indicated by an increase in backchannel nodding behaviour in conversational contexts in which there is a transfer of new information between participants. We also found evidence that slow nodding coherence change across these different contexts. This challenged the 'Social Glue Hypothesis', which states that low frequency slow nodding, or mimicry, is closely related to social bonding and the desire to get on well with others (Lakin et al., 2003), and should therefore create the same motivation to form a social bond across all three conversational contexts. We interpreted this as favouring an interpretation of slow nods as a form of joint attention rather than a mimicry behaviour, and we concluded that fast nods are indeed a different signal to that of coherent slow nods, first hypothesized by Hale et al. (2020). However, it should be noted that the 'social glue hypothesis' is a very general hypothesis. It is not a narrow and strong hypothesis, and researchers are still uncertain about what mechanisms are involved to sustain it.

This chapter examines how non-verbal behaviour in a conversation relates to the outcomes of the conversation in terms of successfully remembering the information that was discussed. First, we review previous studies that link non-verbal behaviour to memory and learning outcomes.

3.2.1 Linking Conversational Behaviour to Outcomes

Pentland (2010) has described non-verbal signals as not just a complement to language, but as a separate communication network. If we understand this old channel of communication, he claims, we can predict the outcomes of many social situations. It is important to know if and how the social signals we have detected – fast and slow nodding – are related to conversational outcomes. A variety of studies have tried to link non-verbal behaviours in conversation to a range of outcomes, and this section summarises some studies in relation to memory and learning.

Previous studies that have used social signals to predict outcomes in social interactions include studies measuring attention and engagement in student classrooms to reveal the benefits for learning and e-learning (Chen, Wang, & Yu, 2015; Pinzon-Gonzalez & Barba-Guaman, 2021; Sümer et al., 2021). For example, Pinzon-Gonzalez and Barba-Guaman (2021) used computer vision techniques in classrooms and measured their head position as an index of social attention. They found that head position estimation can be used to detect levels of attention. A similar study also used a Kinect camera to measure levels of attention in classroom settings to predict memory in this context and to provide automated analytical tools of the learning process (Zaletelj & Košir, 2017). Specific focus has also been on head movements and how they are coordinated with the teacher's motion, which serves as a reliable predictor of the students' level of attention (Sümer et al., 2021).

Eye-movements have also provided insight into memory storage. During storage and retrieval of information, a link between interpersonal coordination and eye gaze patterns have been shown to influence the types of information that is stored (Richardson & Spivey, 2000). For example, people will make eye movements to empty regions of space when retrieving information from memory (Richardson & Kirkham, 2004). Because the success of a linguistic interaction is often dependent on a successful coordination of attention (Clark & Krych, 2004), Richardson and Dale (2005) predicted that the degree of eye movement coordination will reflect the degree to which the listener understood the speaker. Using cross-recurrence analysis, they showed that the more coordinated a listener's eye gaze were with the speaker in a conversation, the better the listener did on a comprehension test.

There is little work on specifically head nodding behaviour in relation to memory and learning outcome, and this study could demonstrate that different frequencies of head nods can be correlated with memory and be used to enhance teaching strategies. As we have seen, head nodding is particularly sensitive to conversational demands and can convey several different meanings (Poggi et al., 2010), such as signaling attention and understanding (Hadar et al., 1983; Kendon, 2002). Attention in human communication is needed for transmitting knowledge from one person to another and is closely related to learning performance (Chen & Huang, 2014; Chen, et al., 2015). According to Hedges et al. (2013) attention may be connected to the learning attentiveness of students to the teacher's instruction throughout a lesson. However, Smith, Colunga and Yoshida (2010) noted that effective learning depends on sustained attention, which plays a major role in acquiring knowledge. A related study also highlighted the importance of sustained attention in cognitive psychology, owing to its strong correlation with learning performance (Steinmayr, Ziegler, &

Träuble, 2010). In relation to this, researchers have also tried to use eye and head movement to predict student learning and selling software (Krithika & Priya, 2016). The authors could analyse if students felt bored or interested in a topic, thereby providing a continuous feedback mechanism for instructors to change their teaching styles. In all these studies, non-verbal behaviour is taken as an index of 'attention' or 'listening' (in contrast to boredom or mind wandering), and thus is expected to predict learning outcomes (i.e., memory). This links to our interpretation of fast nods in head movement as a potential backchannel signalling listening. Backchannels to indicate comprehension are a fundamental component of communication between people (Kendon, 2002), which help establish 'common ground' (Clark, 1996) between two people in a conversation to make the social interaction run smoothly.

An alternative approach to understanding non-verbal behaviour in social interaction can be found in the 'social glue hypothesis', which predicts that coordination should be related to liking and affiliation. Several studies have shown that being engaged in a conversation which includes mimicry can change one's liking for the partner and may even change the sense of self.

Social bonds are critical for our well-being, which forms the basis for much of social interaction. By creating a feeling of closeness to another person, interpersonal coordination can be seen as the glue that bonds social relationships. This has led to the theory that mimicry acts as a 'social glue' to facilitate liking and affiliation to help people bond with members of our social groups (Lakin et al., 2003). Research has also shown that behavioural mimicry appears to influence or affect the self-construal of the person being mimicked (Ashton-James, van Baaren, Chartrand, Decety, & Karremans, 2007). Participants who were mimicked also felt closer to others when completing an 'Inclusion of Other in the Self Scale' (IOS) (Aron, Aron, Tudor, &

Nelson, 1991). During behavioural mimicry, the boundary between self and other is argued to be blurred (Georgieff & Jeannerod, 1998), and Ashton-James et al. (2007) have proposed that behavioural mimicry induces a sense of self-other overlap, where people feel closer or more like the other, leading to more altruistic and prosocial tendencies. Thus, a self-other overlap between people generally assumes that the more they coordinate and mimic each other, the more they see the other as like them, which leads to them acting in a more prosocial manner towards the other person. Ashton-James et al. (2007) also demonstrated that being mimicked makes people behave prosocial towards others in general, and not just the person mimicking. Furthermore, being mimicked is shown to induce cognitive changes in feelings of interdependence (Stel et al., 2011). Taken together, this research suggests that behavioural mimicry also affects the way people think and behave. While some studies investigate the idea that being mimicked increases self-other overlap (Hale & Hamilton, 2016a), this relationship has also been demonstrated to be bidirectional. That is, greater self-other overlap can induce more mimicry behaviour (Maister & Tsakiris, 2016).

'Item memory' and 'self-other overlap' are interesting as potential outcomes of a conversation. In this study we are specifically interested to examine if and how fast and slow nodding behaviour relates to these two conversational outcomes. For example, does fast nodding predict how much one person remembers of the other person's speech? And can slow nodding predict how much self-other overlap there is between people during a conversation?

The first goal of this study is to examine if fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received or processed some new information. As such, we expect that participants

are more likely to engage in fast nodding when they recall more information from a conversation, and from this we could predict memory by correlating the fast nodding behaviour with an increased recall on a memory test.

The second and more exploratory goal of this study is to further explore the relationship between slow nodding coherence and a larger self-other overlap in terms of how close we feel to another person. In other words, such an interpersonal overlap should correlate with increased slow nodding coherence which is believed to signal liking and affiliation between people. We want to test this prediction by correlating the slow nodding coherence behaviour with how biased the participants are towards themselves compared to other on a memory test.

In the next sections, we will cover the two conversational outcomes in relation to memory research. The first section examines how memory relates to social learning (3.2.2), and the second examines how memory relates to self-other overlap (3.2.3).

3.2.2 Memory and Social Learning

Learning new information often occurs in social contexts. However, most research on learning examines isolated participants in front of computer screens, or through observational learning, where participants watch another person via video link. Learning as part of a real-world social interaction has been shown to be particularly valuable in studies with children (Kuhl, Tsao, & Liu, 2003), but has rarely been systematically studied in adults. In this study, we will investigate adult social learning – in particular the process of acquiring new information and factual knowledge – to gain a better understanding of how non-verbal social signals are linked to memory.

Social learning in humans refers to the acquisition of new information that is influenced by observation of, or interaction with, another individual (Heyes, 2012). Learning mechanisms are cognitive processes that encode information for long-term memory storage (Heyes, 2012). However, previous research has also suggested that when information received from the interaction with others is encoded for long-term storage (i.e., social learning), the encoding is achieved by the same cognitive processes that are responsible for the long-term storage of information received through other channels that are not distinctively social in nature (Sterelny, 2009).

After the information has been encoded and stored, it can be recalled from memory. Episodic memory is the ability to recollect previous experiences from long-term memory (Tulving, 1984). Episodic memory is distinguished from semantic memory, which is the ability to store general knowledge about the world without the involvement of personal experiences. In other words, episodic memory involves information about *where*, *when,* and *to whom* an event occurred (Tulving, 1984).

Long-term memory (i.e., both episodic and semantic memory), can be further categorized into explicit (i.e. declarative) and implicit (i.e., non-declarative) memory (Cohen & Squire, 1980). Declarative memory involves the recall of events and facts that can either be episodic (related to an associated event that you have personally experienced) or semantic (conceptual knowledge of an event). In contrast, non-declarative memory involves implicit skill learning and conditioning. Some areas in memory research would also consider the concept of 'learning' to be a more elaborate process where the goal is the acquisition and development of new abilities, skills, values, understanding, and preferences. These capacities in turn depend on socio-cognitive and environmental circumstances (Schwartz, 1992; Tulving, 1972). Together, episodic and semantic memory constitute explicit or declarative memory,

which is part of long-term memory. Episodic memory involves a person's recollection of temporally dated information of specific details of the event such as in what context (e.g., time and place), or to whom it happened.

In this study, we focus on different measures of declarative episodic memory which have been acquired through social learning. First, we are interested in the general recall rates of the participants and how much they have learned in terms of being able to remember if a fact they are presented with on a memory test is an old fact (i.e., one they could remember discussing with their partner) or a completely new fact (i.e., one they could not remember discussing with their partner). Secondly, we are interested in the self-preferential encoding of information associated to the self, and source memory confusion towards the self. In this study, recalling a fact counts as an episode and is not a factual semantic recollection. This is because (1) we have the participants interact in social learning in a dyadic real-world conversation, and (2) we have a post-hoc memory test that prompts the participants to recall their discussion with the other participant, which links the recollection to an event that they personally experienced (i.e., "Is this something I've heard from our discussion?"). Hence, in this study we are limiting our investigations to measuring declarative episodic memory. In the next section we will look at how memory relates to self-other overlap and how we measure it.

3.2.3 Memory and Self-Other Overlap

The 'social glue hypothesis' proposes an explanation for *why* behavioural mimicry may be an effective means of feeling closer to someone, conceptually presented as a larger self-other overlap. Less research has focused on explanations of *how* such

consequences of mimicry might occur. The term ‘self-construal’ is often used to define the interactive relationship of the self with others (Brewer & Gardner, 1996), and is the extent to which people define themselves, or construe their identity with reference to their social roles, groups, status, or relationships. For example, someone with an *independent* self-construal might identify themselves by their individual skills and attributes (e.g., “I am tall”), whereas someone who has an *interdependent* self-construal would be more likely to define themselves by their relationships with others (e.g., “I am a sister”). As such, people with an interdependent self-construal generally feel closer to others, both emotionally, psychologically, and physically (Aron, Aron, & Smollan, 1992; Holland, Roeder, van Baaren, Brandt, & Hannover, 2004). Thus, an interdependent self-construal is associated with a more other-focused, and hence prosocial orientation. Research on self-construal has inspired the view that mimicry helps increase the interdependence of one’s self-construal, which lead to a larger self-other overlap between people and more positive social consequences (Ashton-James et al., 2007). Recent findings challenge this view. For example, Cross, Turgeon, and Atherton (2019) suggest that interpersonal coordination in general is a result of categorization processes, where we self-identify as a group member, rather than feelings of closeness. This highlight the fact that the mechanisms linking mimicry and self-other overlap are still debated.

As we mentioned previously, the concept of ‘self-other overlap’ was first measured using the Inclusion of Other in the Self Scale (IOS) (Aron et al., 1991). The scale measures overlap between self and other in terms of how close two people feel but cannot distinguish between whether the self is being included in the other, or vice versa. However, in their ‘self-expansion’ model, Aron et al. (1991) argue that close relationships are characterized by including the other in one’s own mental self-

representation. One of the difficulties in much of the early work about the self was that it depended on subjective reports. Hogeveen, Chartrand and Obhi (2014) also found that mimicry did not lead to increased self-other overlap using the IOS scale. Moreover, most research using self-report measures to explore self-other overlap and prosocial behaviour has focused on the individuals' observations of their recently mimicked conversation partners and their subsequent relationships (Hove & Risen, 2009; Wiltermuth & Heath, 2009). Relatively few studies have explored how we perceive ourselves during a conversation with someone – that is, how mimicking or being mimicked can affect our self-construal, or our sense of identity relative to the other person. Looking at aspects of an individual's self-construal might prove to be a useful paradigm to explore the relationship between mimicry and self-other overlap. In this study, we aim to use more behavioural measures of memory recall in relation to two well known memory effects associated with the idea of self-construal.

The Self-Reference Effect (SRE). The SRE in memory research refers to the preferential encoding of information associated to the self compared to others (Klein, 2012; Symons & Johnson, 1997). For example, if you go out to have lunch together with a friend, you are more likely to remember what you had yourself rather than your friend. The SRE has been demonstrated with a wide range of different types of stimuli (e.g., faces, traits, geometric shapes) and tasks (e.g., perception, memory, decision making) and are linked to distinct patterns of neural activity (Klein, 2012; Macrae, Moran, Heatheron, Banfield, & Kelley, 2004; Powell, Macrae, Cloutier, Metcalfe, & Mitchell, 2010). This illustrates the broad range of self-reference effects on information processing. The term self-*reference* can be rather misleading, and what researchers mean when using the term is information being *associated to* or *linked to* the self, rather than just information referring to the self.

In their seminal study, Rogers, Kuiper, and Kirker (1977) presented trait adjectives to participants who self-rated the adjectives on how well they described themselves. The participants were later asked to recognize the adjectives from a list of words as either “old” (a word on the list) or “new” (not on the list). They counted the number of times that the participants incorrectly selected “old” (false alarm rate) and found a correlation between the number of false alarms and self-reference from the ratings, which meant that they had good memory of adjectives that they had judged in relation to themselves rather than other people.

The SRE has been replicated in subsequent studies. Maki and McCaul (1985) compared the memory recall of trait adjectives to that of nouns and showed patterns of the SRE based on the stimuli materials used: when traits adjectives were used recall was highest when the words were describing themselves, but when nouns were used recall was higher when the nouns were describing someone else. The reason was thought to be that traits are part of the self-schema (Markus & Kitayama, 1991), but nouns are not, hence only traits are encoded in a self-reference task.

Cunningham, Turk, Macdonald, and Macrae (2008) investigated this further in dyadic interactions with children by presenting them with pictures of their own or another child’s face along with an object. On a subsequent memory test, the children not only demonstrated better recall on objects judged in relation to themselves, but they also remember better which face the remembered object was presented with. These results beg the question of whether the SRE not only reflects an underlying memory process but can also bind the elements within an episode, like binding the memory to its source (i.e., who), and when it was processed in relation to the self (i.e., when). Whereas the benefits of the SRE are supported, less is known about this ‘who’ question, which leads us into an area related to ‘source memory’.

Self-Bias Effect (SBE). In a classic study by Russell and Jarrold (1999), instead of being tested on their ability to recall self-related information, a group of children with autism were tested on their ability to distinguish who (i.e., “self” or “other”) made an action to place a specific set of cards down on a table. A significant bias was observed, in which the participants were more likely to claim “I placed it” even if the other person had. As such, the Self-Bias Effect (SBE) is about the misidentification of the origin of a memory towards the self or another person (Schacter, 2001). The SRE, in comparison, can increase our attention to self-related events, engaging processes that increase recognition, without binding the event to the self – a process which would be necessary to enhance source memory and elicit a self-bias. Again, imagine going out to have lunch with a friend, and instead of being more likely to remember what you ordered rather than your friend, you ask yourself: “Did I mention the weekend trip to the sea, or did my friend say that?”. Here you are making a source memory judgment. Recent research shows that observing another person’s action can lead people to mistakenly recall that they have performed the action themselves (Schain, Lindner, Beck, & Echterhoff, 2012). Similarly, an example of a SBE is when you make the judgement if it was you or your friend who mentioned the weekend trip yesterday, you can be biased in claiming that it was your idea or statement, when in fact it may not have been.

Research on the SRE has established that inputs from the environment that are perceived to be related to the self are difficult to ignore (Bargh, 1982). In other words, the input from the environment specifies what people will and will not remember. With the SBE, rather than recalling self-related *input*, in a way it is more about self-related *output* in the form of recalling *who* performed a specific action and being biased, for example by claiming another person’s action as your own (saying ‘I

did it' even if you didn't). This idea is closely related to the concept of 'destination memory', which is about remembering to whom one has told what or remembering from whom one has received information (Gopie & MacLleod, 2009).

In the next section we will present the current study. In this study we have two goals, each aiming to investigate if and how the social signals we have observed are related to conversational outcomes. The first goal is to investigate if encoding new information is associated with fast nodding, and the second goal is to explore the potential relationship between slow nodding coherence and self-other overlap. We collected a new dataset and tested if these different types of head nodding frequencies – fast and slow – during conversation could predict performance on an outcome measure in the form of a memory test. This will hopefully give us clues as to whether we recall more information when interacting with others using head nodding in conversations. We also used both the SRE and SBE as two behavioural measures in an exploratory fashion to see if real-world behavioural mimicry is associated to any self-other effects.

3.3 The Present Study

In our previous study (Chapter 2), we demonstrated that fast nodding behaviour is found more in conversational contexts in which there is a transfer of new information between participants. From these results, we posited that fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received this new information. In this study, we suggest that measuring changes in social signals like head nodding during conversation can provide researchers with clues about our cognitive processes, such as predicting memory performance. Thus, the main goal of this study is to investigate if fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received or processed some new information. If this is the case, we predict that if fast nodding is seen during a conversation, participants are more likely to recall information from that conversation in a later memory test. That is, can we find a relationship between the fast nodding behaviour and an increased recall on a memory test.

In our previous study (Chapter 2) we also found evidence to support the idea that the slow nodding behaviour changes across different conversational contexts, in favour of being a form of joint attention rather than a mimicry behaviour. Seeing as both joint attention and mimicry can be interpreted as a form of social glue that make people feel closer to each other, we are interested to explore the potential relationship between a larger self-other overlap in terms of how close we feel to the other person, and the slow nodding coherence pattern. We will test this prediction by correlating the slow nodding coherence with how biased the participants are towards themselves compared to other in a memory test, using the SRE and SBE measures described above.

3.3.1 Conversation Task and Memory Test

We created a new conversation task, which is similar in structure to the Picture Description Task used in the previous study (Chapter 2). We wanted to keep the turn-taking structure from the Picture Description Task, but have explicit, yet uncommon, facts to remember so that we could create a sensible memory test and measure recall of each fact both speakers and listeners in a dyad. In this task, the participants are referred to as either *speakers* or *listeners*. A 'speaker' is a participant that during the task is reading the facts and conveying the information to their partner by speaking, and a 'listener' is the one receiving the information. It is difficult to define a good memory test using pictures as they can be interpreted in different ways depending on what person is receiving the visual stimuli and are often more specific. As such, they are often used more as a prompt to conversation rather than as an explicit measure. During this task, the facts that the speakers read were designed to be uncommon to increase the likelihood that the participants would not recognize them and see them as new facts. For example, the most well-known facts were omitted in favour of less known ones.

This study aimed to assess how well participants remember different facts relating to American states. The conversation task that they had to engage in was a form of one-way information sharing in which they were asked to take turns at reading a set of cards to each other, each of which contained three quirky facts associated with an American state to provoke comments. Each trial was divided into two parts. During the first part (monologue), the speakers held the card in front of them while they read the facts aloud for 45 seconds, while the other participant listened. During the second part (dialogue), both participants had a free conversation about the facts for 30 seconds. For example, at this point the listener could start asking questions about

the facts. A total set of 16 cards were used, split up into two piles for each recording session, and each card, or trial, referred to a different American state.

Following each recording session, the participants were asked to complete a surprise memory test on two separate computers in order to assess: (1) How many facts they remember from the conversation (i.e., general recall); (2) if they remember more facts when speaking compared to listening (i.e., SRE); and (3) if they are biased in claiming that it was them reading a fact even if it was not (i.e., SBE).

At this point, the facts about the American states that the participants would have spoken or listened to during the task will no longer be new to them during the memory test at the end of the experiment. At the point of the memory test, both speakers and listeners were asked to make a judgement on a prompt to recall if a fact was Old (i.e., a fact they could remember) or New (i.e., a fact they did not remember) in relation to the task. This is similar to what Rogers et al. (1977) did when they presented trait adjectives to participants who later were asked to recognize them from a list of words as either “old” (a word on the list) or “new” (not on the list). This judgement was relevant for assessing the participants general recall rates and the self-reference effect (SRE). If, and only if, the participants recalled and answered that it was an old fact on the first question, a second question followed, prompting them to make a source memory judgement to decide whether it was them (i.e., “Self”), or the other participant (i.e., “Other”) who read that fact. This follow-up question was relevant for assessing the participants self-bias (SBE). For more details on this procedure, see Methods, Section 3.4.3.

3.3.2 Aims, Hypotheses, and Predictions

We use our data to test if different types of head nodding frequencies (i.e., fast, and slow) could predict performance on an outcome measure in the form of a memory test. Specifically, we will test two different hypotheses, each based on one of the two head nodding frequencies:

H_1: *Encoding new information is associated with fast nodding behaviour.*

H_2: *Self-other overlap is associated with slow nodding coherence behaviour.*

Our first hypothesis (**H_1**) claim that encoding new information is linked to fast nodding, and is based on our findings from Chapter 2, that fast nodding behaviour is found more in conversational contexts in which there is a transfer of new information between participants. From this we hypothesize that encoding new information is associated with more fast nods because they are used as backchannels to signal to the other person that we have received and encoded the information. More specifically, if **H_1** is true, then we expect that if fast nodding is seen during a conversation, participants are more likely to recall information from that conversation. Testing this hypothesis, we predicted (**P_1**) that fast nodding should correlate with increased general recall on the memory test (Figure 3-1, P1), measured by how many facts they remember.

Our second hypothesis (**H_2**) claim that self-other overlap is linked to coherent slow nodding, and is based on the idea that the coherence of slow nodding is related to social bonding and the desire to get on well with others, either through mimicry or joint attention (i.e., 'social glue hypothesis'). From this we hypothesize that a larger self-other overlap between people should be associated with more slow nodding

coherence because such head nodding frequencies are used to signal liking and affiliation. More specifically, if **H_2** is true, then we expect participants to be more likely to engage in coherent slow nodding behaviour when they feel a greater self-other overlap (i.e., increased feeling of closeness), and this could in turn be correlated with self-related effects and biases. For example, the larger the self-other overlap is between people, the more confused we should become at distinguishing between *who* said what during the conversation (i.e., source memory), which in turn makes us more prone to mistakenly claim an idea as our own (i.e., SBE). Testing this hypothesis, we predicted (**P_2**) that slow nodding coherence should correlate with a SBE on the memory test (Figure 3-1, P2), measured by how biased they were in claiming that it was them reading a fact even if it was not.

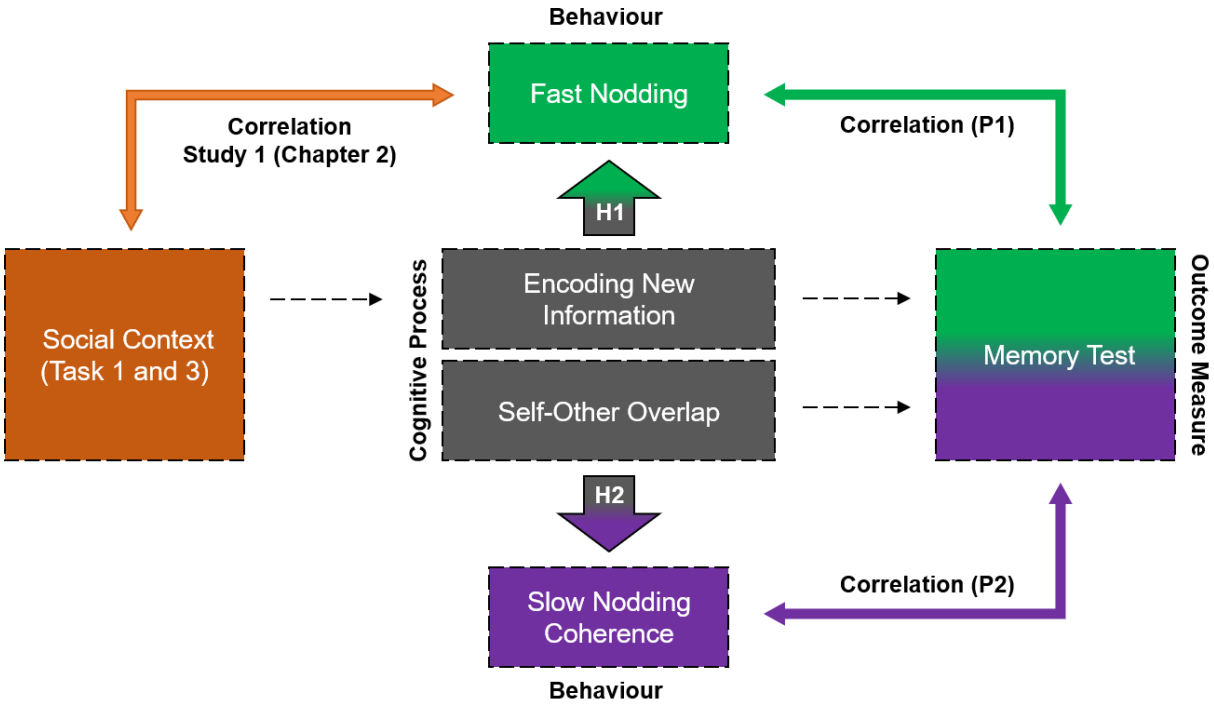


Figure 3-1. Box diagram of hypotheses and predictions. From the conversational contexts used in the first study (Chapter 2) (orange box) we found a correlation (orange arrow) between tasks that involved a novel information transfer and fast nodding behaviour (green box). In this study, we hypothesize (H1) that encoding new information is associated with fast nodding behaviour by predicting (P1) a correlation between fast nodding and increased general recall on the memory test (green arrow). We also hypothesized (H2) that self-other overlap is associated with slow nodding coherence behaviour, by predicting (P2) a correlation between slow nodding coherence and greater memory performance for the SBE measure on the memory test (purple arrow).

3.4 Methods

3.4.1 Participants

60 participants (M_{age}=25) were recruited from the UCL Psychology Subject Pool and the ICN Subject Database. Exclusion criteria included subjects that were not fluent in English. All participants were recruited and tested in pairs (30 dyads) and were randomly paired to arrive at the same time. On arrival, participants were asked to remove eye-makeup, bulky clothes, and jewelry as to not interfere with the recording equipment. The participants did not have any previous experience with the tasks and were unaware of the purpose of the experiment. Ethical approval for video, audio, and motion capture recordings was arranged via the UCL Research Ethics committee, and all participants gave their written informed consent. A monetary reimbursement was offered for participating in the study at a rate of £7.50/hour.

3.4.2 Equipment

In the present study, we took multimodal recordings from 30 pairs of participants (dyads) engaged in a conversation task, followed by a post-test measure. Audio instructions, together with audio cues indicating the start and stop of a recording, were given to the participants via two speakers placed on the floor next to them. Two LED lights were stationed next to the participants to better illuminate their facial features. Curtains separated the participants from the experimenter, who remained in the room, but did not interact. Behind the curtains we had three computers that coordinated the whole experiment (Figure 3-2A, B, C). For further details on the components used in this study, see Chapter 2, Section 2.4.2.

3.4.3 Procedure

Participants arrived at the lab and were shown all the equipment and were informed of the procedures. They signed the informed consent, and then put on the motion capture suits, eye-trackers, and microphones. We then completed the calibration procedures for the motion capture and the eye-trackers. Each person in the dyad was then randomly assigned to be either the 'Yellow' or 'Blue' participant, and sat one meter apart on small stools, before beginning recording the experimental task.

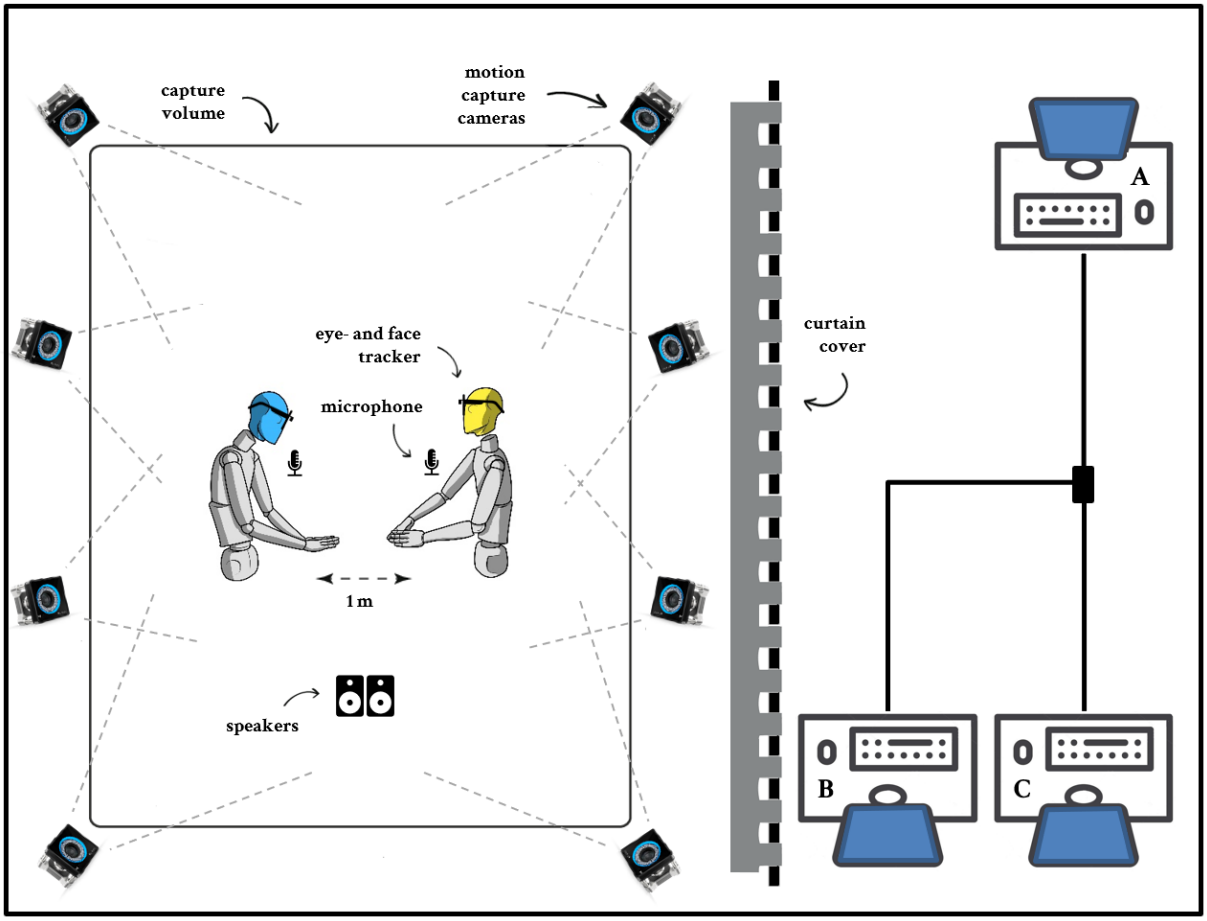


Figure 3-2. Lab setup. Equipment included motion tracking cameras (4x Optitrack Prime 13 & 4x Prime 13W), audio speakers, wearable microphones connected to an audio mixer, eye- and face trackers (Pupil Labs), LED lights, and a curtain to separate the three computers running the experiment. Computer A acted as the client that communicated with the two computers B and C acting as servers. For further details on each of the components, see Chapter 2, Section 2.4.2.

Conversation Task. This task was a form of one-way information sharing task in which the participants were asked to take turns at reading a set of cards to each other, each of which contained three facts associated with an American state (Figure 3-3B). Each trial was divided into two parts. During the first part (monologue), the speakers held the card in front of them while reading the facts out loud for 45 seconds, while the other participant just listened. During the second part (dialogue), both participants had a free conversation about the facts for 30 seconds. For example, at this point the listener could start asking questions about the facts. Audible cues signaled the start and end of each trial, and the transition from monologue to dialogue (Figure 3-3A).

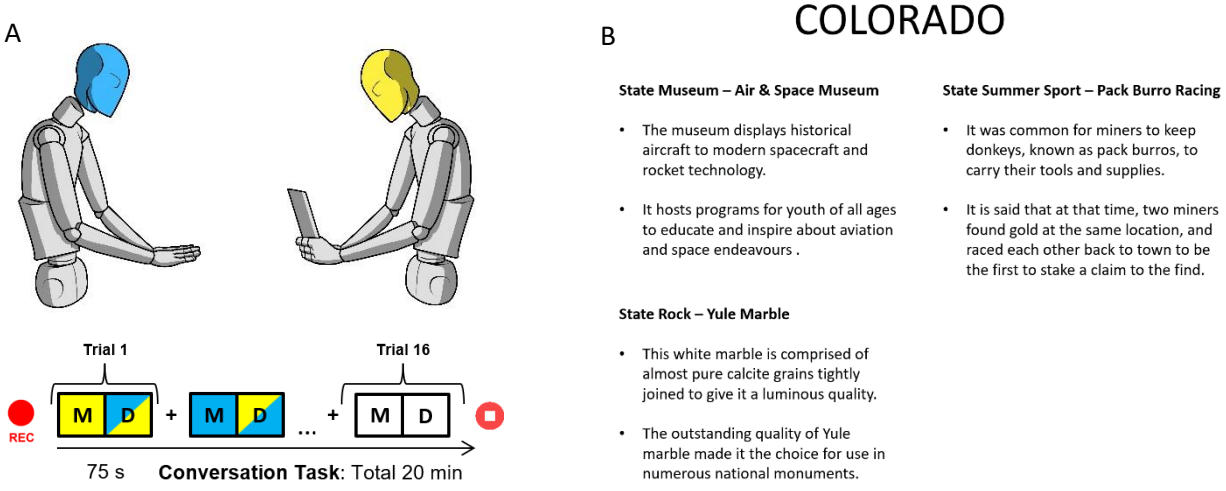


Figure 3-3. Trial timeline. (A) Conversation task with turn-taking order of monologue [M] and dialogue [D] over 16 trials between the yellow and blue participant. (B) Example card of facts from one state.

A total of 16 cards were used for each recording session, and each card, or trial, referred to a different American state. The selected states and facts on each card were chosen with the intention of minimizing prior knowledge of American states. For example, the most well-known states and facts were omitted in favour of less known ones. The 16 states were equally divided into two sets, which represent the two piles

of 8 cards that each participant in the dyad received in a random order. The sets given to the blue participant was alternated from dyad to dyad (i.e., Set 1 was given to the blue participant on odd dyads, and Set 2 on even dyads). The participants had a short break in the middle of the session, after they had completed 8 trials. See Table 3-1 for a complete list of the 16 states divided by set.

Set 1	*Set 2*
Massachusetts	Maine
Illinois	Ohio
Georgia	Florida
Pennsylvania	New Hampshire
Arkansas	Maryland
Idaho	Wyoming
Oregon	Indiana
Delaware	Mississippi

Table 3-1. States divided by set. The 16 American states divided by set of cards (Set 1 and Set 2), which represent the two piles of 8 cards that each participant in the dyad received in a random order. The sets given to the blue participant was alternated from dyad to dyad (i.e., Set 1 was given to the blue participant on odd dyads, and Set 2 was given to the blue participant on even dyads).

Memory Test. Following the recording session, the participants were assisted in removing the equipment and moved to separate tables where they were instructed to individually complete a memory test on a computer. The test was created with the graphics toolbox Cogent for Matlab 2018b to assess their memory on the state facts (Figure 3-4). The test included a total of 96 facts, 48 of which were the real facts presented to them during the task, and the other 48 which were new ones. All facts were presented in a random order and the participants were instructed to decide if it was an "old" fact – one they could remember discussing, or if it was a completely "new" fact – one they did not remember discussing during the task.

After deciding whether the fact had been discussed during encoding, and only if they answered that it was an old fact, a subsequent question followed, prompting the participant to make a source judgment about the encoding task to decide whether it was them ("Self"), or the other participant ("Other") who read the fact. This follow-up question was relevant for assessing the participants self-bias (i.e., SBE).

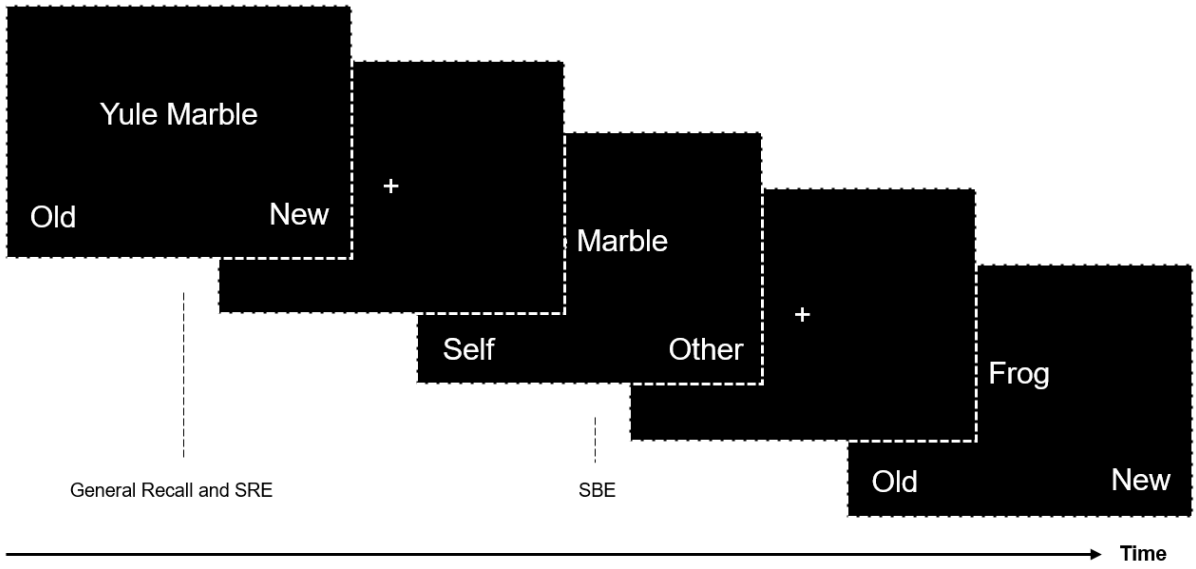


Figure 3-4. Memory test. A timeline of the screens presented to the participants. The nouns "Yule Marble" and "Frog" refer to state facts. The "Old" or "New" labels refer to the decision that the participant had to make regarding those facts of whether it was something they remembered discussing during the task or not. The "Self" and "Other" labels refer to the follow-up judgement the participant had to make if they previously answered "Old", of whether it was them or the other participant who read the fact. The First judgment assessed the general recall and the SRE, whereas the second assessed the SBE.

3.5 Data Analysis

Our aim was to record and measure the interpersonal coordination of head nodding patterns found in both the high and low frequency bands, and link these to memory performance. For the analysis of the head nodding data, we employed the same wavelet analysis approach used in Chapter 2 (See Section 2.2.3 for an introduction to wavelet analysis, and sections 2.5.1–2.5.2 for details on extracting the signals and analysing the head nodding patterns). In the following sections I will present how we analysed the memory performance measures (3.5.1), and how we matched the head nodding data to the memory data using a mixed-effects model analysis (3.5.2), followed by a short summary of the methods (3.5.3).

3.5.1 Analysis of Memory Performance

This study aimed to assess participants' (1) general recollection, (2) the SRE, and (3) the SBE, in relation to different facts about American states. Memory decisions were categorized into one of four response categories based on the "old/new" or "self/other" status of the fact and the given response: hit, miss, false alarm, and correct rejection. Among these, hits and correct rejections are considered correct responses, whereas misses and false alarms are considered incorrect.

General Memory Recall. If information was recalled, participants gave an "Old" response to indicate that they remember the fact from the conversation. If no information was recalled, a "New" response was given as participants were not able to recognize the fact. The participants' ability to discriminate Old from New facts on the memory test was calculated by considering the number of hits and correct rejections (i.e., the sum of Old/New responses correctly identified as Old/New), presented as a percentage of correct recall (hits + correct rejections divided by 2).

We further calculated the distance between the Old and New response distributions using D-prime (d') as a sensitivity measure on how well the participants were able to discriminate between Old and New responses. This was attained by transforming the hits and false alarm rates (i.e., the sum of New responses identified as Old, as a proportion of the total number of responses) into z-scores, and subtracting the z(false alarm) from z(hit). A larger d' represents a greater discrimination ability between the two chosen distributions – or better memory recall.

Self-Reference Effect (SRE). Another way to quantify the relative contributions of self- and other-related information when judging general recall responses is to analyse each participant's general memory recall based on whether they were speaking or listening. For example, if the participants remember more facts when

speaking compared to listening, it will give us the indication that people are more likely to remember (i.e., better encode) information that relates to themselves, regardless of how accurately they are able to identify the source of who did the encoding (SBE). The SRE was calculated by taking the number of correctly recalled facts between speaking and listening trials, based on which set of cards the participants started with, and comparing the means using a paired samples t-test.

Self-Bias Effect (SBE). For each fact that the participants responded "Old", they got a follow-up question prompting them to identify whether it was them ("Self") or their partner ("Other") who read the fact during the conversation. The participants' ability to discriminate Self from Other on the memory test was calculated by the number of hits (i.e., the sum of Self responses correctly identified as Self), presented together with correct rejections, misses, and false alarms. However, because the participants could only give a response if they already decided it was an Old fact, a separate index was calculated and defined as the proportion of facts correctly identified as Old that were also correctly identified as Self or Other.

Same as with the general recall, we calculated the distance between the Self and Other response distributions using D-prime (d') as a sensitivity measure on how well the participants were able to discriminate between Self and Other responses. Finally, we also analysed the participants response bias on their willingness to claim that it was them reading a specific fact. We describe our results in terms of the bias estimate 'criterion' (*c*) (Macmillan, 1993), a widely used measure given as the opposite of half the sum of the *z*-converted hit and false alarm rates. A negative value of *c* indicates a criterion to the self- and other-item distributions and a bias to respond 'self'. A positive value means that they are biased to respond 'other' and require stronger evidence before claiming to have read the fact themselves.

3.5.2 Mixed-Effects Model Analysis

After extracting the signals from the head nodding behaviours and analysing the memory performance data, we aimed to correlate the two types of head nodding behaviours – fast nodding and slow nodding coherence – with the memory performance measures. We calculated the number of correct answers that the participants got for each state (i.e., 0–3 correct facts for each of the 16 states) and divided it into speaking and listening trials. However, we were not able to do a full trial-by-trial analysis on the Self-Bias Effect (SBE) because of uneven trials. For instance, when the participants made their first judgment on whether they remember discussing the fact during the conversation (i.e., Old/New), all responses were balanced, with 4 possible responses for each state (i.e., 0-3). On the other hand, the source memory responses, whether they remember being the one reading the fact or not (i.e., Self/Other) depended on the participants recalling the fact to begin with (i.e., responding “Old”). This resulted in the source memory data ending up with some trials only having 1 or 2 responses. Future studies or further analysis would benefit from equating the number of responses for each state.

Figure 3-5 shows the pipeline of the wavelet analysis of the head nodding data used in Chapter 2 (see Section 2.5.2 for further details). The raw head pitch trajectories for both participants (A, B) were transformed using wavelet transforms to get the time-frequency representation of each time-series (C, D). Next, we calculated the cross-wavelet coherence between each of the two transformed signals (E), and as the final step averaged the coherence (R^2) over the time-course of each trial (F).

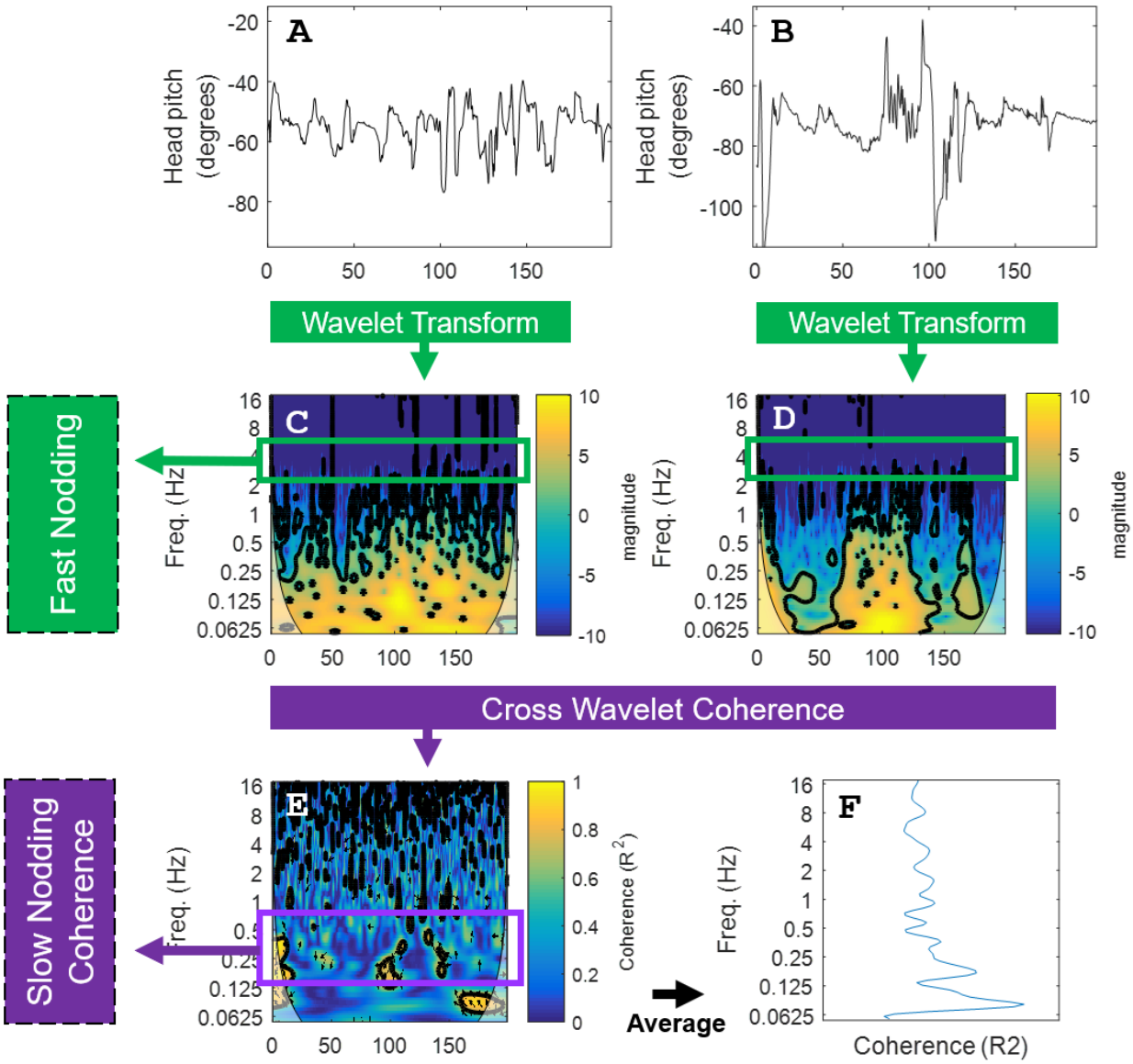


Figure 3-5. Selection of head nodding frequency bands. Fast nodding behaviour was selected from the high frequency band of the individual wavelets (green). Slow nodding coherence was selected from the low frequency band of the cross-wavelet coherence (purple).

Data on the head nodding behaviour were selected from two different stages of the wavelet analysis (Figure 3-5). Because fast nodding has shown to be a high frequency backchannel behaviour from the listener to the speaker in a conversation, it can be analysed on the participant level. Hence, for the fast nodding behaviour we selected data from the high frequency band of the individual wavelets (Figure 3-5C, D) before performing the cross-wavelet coherence. On the other hand, slow nodding

is more spontaneous and overlapping in terms of turn-taking and has been shown to be related to mimicry and joint attention behaviour. What this means is that slow nodding must be analysed as a coordinated behaviour on the dyadic level. Hence, for the slow nodding behaviour we selected data from the low frequency band of the cross-wavelet coherence (E). To avoid circular analysis, or "double dipping", we used the frequency ranges presented by Hale et al. (2020). This included coherence (R^2) at 0.2–1.1 Hz for slow nodding and the average power at 2.6–6.5 Hz for the fast nodding. However, a higher average power in the high frequency range means that the participants are not necessarily making more fast-nods but are putting more energy into their fast nodding. This measure can be interesting, but for our analysis we want to measure fast nod count of individual nods.

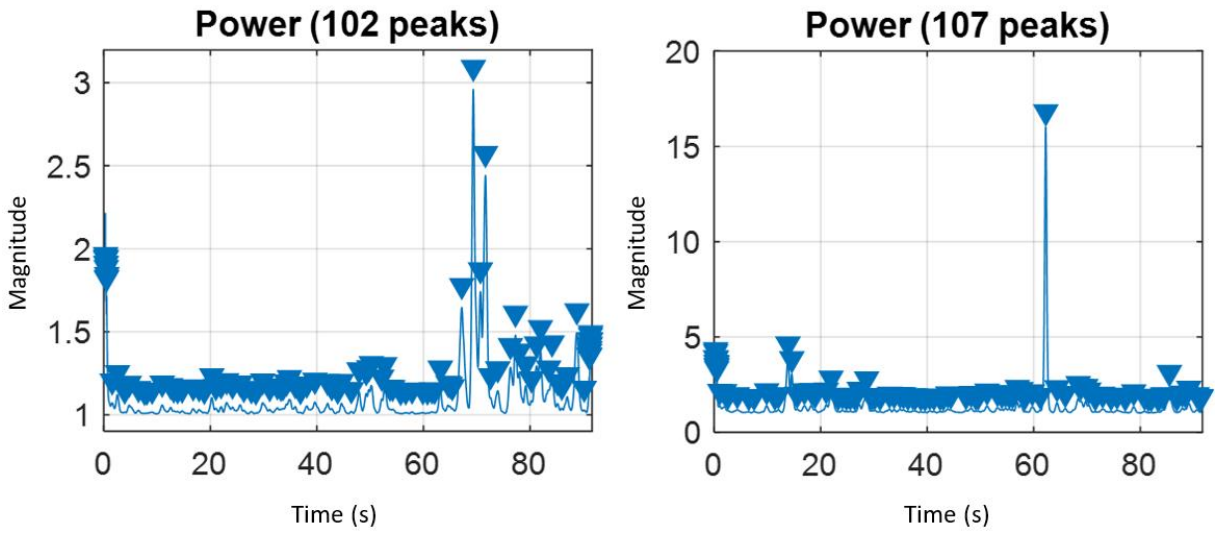


Figure 3-6. Fast nod detector. The graphs show the magnitude of the power from the selected high frequency band (2.6-6.5 Hz) averaged over each time point (i.e., period). The blue triangles on the peaks of the signal represents a fast nod, which was identified using a peak detection algorithm.

To calculate the average fast nod count, we sum up the average power for each time point (i.e., period) in the selected frequency band (i.e., 2.6–6.5 Hz) of the individual wavelets. We then used a peak detection algorithm called **findpeaks** in Matlab 2019b to identify the number of peaks (nods) across the signal (Figure 3-6).

In addition to the steps taken when pre-processing the head motion data in the previous study (Chapter 2, Section 2.5.1), we excluded 7 dyads in this dataset during post-processing due to extensive marker swapping, which created too many unnatural spikes or slopes in the signals. In cases of minor dropped signals, we used linear interpolation between the signals to remove NaNs so that the wavelet toolbox in Matlab was able to calculate the wavelet transforms. In total, this affected approximately 500 / 100.000 (0.5%) datapoints in each of the affected samples.

The final sample was used to create five models (M_1–M_5) that consisted of 46 participants (23 dyads) sorted into speaking and listening roles; with 16 trials sorted by Set (Table 3-1), for a total of 96 facts per dyad. The data included in the mixed-effects model were the memory recall responses (i.e., 0-3 recalled facts), the fast nod count calculated from the average power in the 2.6–6.5 Hz frequency band, and slow nods as the degree of coherence (R^2) in the 0.2–1.1 Hz frequency band:

M_1: *Memory ~ (1 | Participant)*

M_2: *Memory ~ Fast Nodding + (1 | Participant)*

M_3: *Memory ~ Slow Nodding Coherence + (1 | Participant)*

M_4: *Memory ~ Fast Nodding + Slow Nodding Coherence + (1 | Participant)*

M_5: *Memory ~ Fast Nodding * Slow Nodding Coherence + (1 | Participant)*

We estimated the models of the sample using multilevel statistical modelling, and a model comparison approach (Judd, McClelland, & Ryan, 2008). Prior to model comparison, we performed a linear multilevel regression for both speakers and listeners for all models, except for M_1 which was the null model (i.e., compact model). We used two-level models with the head nodding frequencies as predictors (level 1)

nested within participants (level 2). We had no interest in analysing the grouping variable of participant as a random effect but needed to factor this out for individual variation in the model parameters. The dependent variable was the memory performance (the number of correct answers for each state). The models M_2 and M_3 were created to test whether fast nodding and slow nodding coherence behaviour would individually predict memory performance. Model M_1 was the null model (compact model) that was used to compare the goodness of fit and parameter estimates of M_2 and M_3. We also combined the two factors into a single saturated model M_4, using both fast and slow nodding to predict memory performance. The final model, M_5, was created to test for an interaction between fast and slow nodding on the dependent variable. Furthermore, speaking and listening were treated in separate models. This decision was based on evidence from Hale et al. (2020) showing that nodding looks different during speaking and listening. Using separate models means that each model is simpler and easier to interpret.

In recent decades, statisticians and psychologists have developed methods of model comparison that go beyond traditional significance testing (i.e., NHST). We use one such approach, called the model comparison approach (Judd et al., 2008). From the model comparison approach, we compare the explanatory power of the five models that we created, with the goal of both identifying the model that best explains variance in our dependent variable (i.e., goodness of fit), as well as to estimate the parameters of interest in each model that best support or contradicts our proposed hypotheses. The essence of the model comparison approach to statistical testing is that it conceives of statistical tests of experimental effects as a comparison between two alternative models of the data that differ in the assumptions that they make. The nature of these assumptions of the two compared models

determines which research question is targeted (Prins & Kingdom, 2018). For example, at the bottom of the hierarchy is a null model M_1. Above the null model are two models with a single covariate: M_2 with the main effect of fast nodding and no other effects; M_3 with the main effect of slow nodding and no other effects. At the top of the hierarchy is model M_5 with the fast and slow nodding interaction as well as the main effects.

We used the function **fitlme** in Matlab 2019b to perform all the analyses and model comparisons. This function allows us to fit our linear mixed effects models which gives us more accurate results for grouped data, as the p-values for fixed effects coefficients do not generalize across levels of the random factors, which in our case are the participants because each has a different ability to recall information from memory. We then used the function **compare** in Matlab 2019b to test the goodness of fit with the Akaike's Information Criterion (AIC), and to estimate the parameters of interest between the different linear mixed effects models with Theoretical Likelihood Ratio Tests. These tests are the standard approach to evaluating mixed (i.e., multilevel) models.

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible parameters, which means that low AIC scores are better because it requires less information to predict with almost the same precision. So, if two models explain the same amount of variation, the one with less parameters will have a lower AIC score and will be the better-fit model. When comparing models using AIC-scores, if a model is more than 2 units lower than the other, then it is considered significantly better. We also looked at the maximum likelihood estimate to represent the likelihood that a model could have produced the observed memory performance. From the perspective of the model comparison

approach, the best-fit model is immediately interpretable: The evidence for the effect of a factor is in the comparison to the null model. As AIC values are not “tests” but measure the goodness of fit, Theoretical Likelihood Ratio Tests instead focus explicitly on decision thresholds and can be interpreted with the model comparison approach as the likelihood or probability of seeing the data you collected given your model. The benefit of this approach is that we can tailor the model precisely to match our hypotheses, providing more ways to compare the data than is possible with standard ANOVA procedures (Rouder, Engelhardt, McCabe, & Morey, 2016). However, interpreting p-values for the likelihood ratio test (i.e., LTStat) for mixed models are not as straightforward as they are for the linear model. There are multiple approaches with differing opinions about which approach is the best and if there’s even a correct way (Hox, Moerbeek, & van de Schoot, 2010; Winter, 2013). Here, we present two major analyses of our data (i.e., the multi-level modelling and the model comparison approach) for fairness and completeness.

3.5.3 Methods Summary

In the present study, the aim was to investigate if different types of head nodding frequencies (i.e., fast and slow) could predict performance on an outcome measure in the form of a memory test. Specifically, we will test two different hypotheses, each based on one of the two head nodding frequencies.

Dyads were recorded with motion capture and engaged in a new conversation task, followed by a post-test measure to assess participants' memory performance relating to facts about American states. Various analyses of the memory data were performed to assess (1) general recollection, (2) the SRE, and (3) the SBE.

We used wavelet analysis to measure the two types of head nodding behaviours. The frequency bands for fast nodding and slow nodding coherence were selected from different stages of the wavelet pipeline. Fast nodding was measured by running a fast nod detector to calculate the fast nod count in the 2.6–6.5 Hz band, and slow nodding was measured as the degree of coherence (R^2) in the 0.2–1.1 Hz band.

To get a more detailed picture of the link between memory and head nodding, we correlated the measures of the two head nodding behaviours with the memory performance. We estimated five mixed-effect models and performed a linear multilevel regression for speakers and listeners on all parameters. We also compared the models using AIC for goodness-of-fit and Likelihood Ratio Tests to observe the probability of observing the collected data.

3.6 Results

3.6.1 Mean Memory Performance

Figure 3-7A shows the % of general memory recall across speaking and listening conditions (See Section 3.6.2 for the full analysis). A paired-samples t-test shows that there is a significant increase in general memory recall when speaking (M = 0.93, SD = 0.09) compared to listening (M = 0.72, SD = 0.19), $t(90) = 6.9$, $p < .001$. This result shows that people are more likely to remember information they produced themselves, giving support to the existence of the SRE in memory.

Figure 3-7B shows the percentage of the Self-Other response rates across the four response categories. We can observe a high response rate of hits (92%) and correct rejections (98%), with low rates of misses (8%) and false alarms (2%). This indicates that people are good at remembering who was speaking during the encoding process. However, a result of 93% Old facts correctly identified as Old for example, does not necessarily mean that the participants have successfully discriminated between Old and New facts on the memory test – they could have just responded randomly when they weren't sure of the answer. In other words, the test is not sensitive to *why* any particular response was given. To support these results, we calculated the participants response sensitivity, or their ability to discriminate Old from New facts on the memory test using d-prime as a sensitivity index (Figure 3-7C). A higher d-prime represents the participants' ability to discriminate between their responses, and the results indicate that the participants show good sensitivity to the Old-New responses (d'=3.24) and what they can remember from the conversation. We also calculated d-primes for the Self-Other responses and the results demonstrate a similarly good sensitivity (d' 5.08) to express how good they are at remembering the source from the conversation.

Finally, to examine the SBE, we analysed the participants response bias on their willingness to claim that it was them reading a specific fact. Negative values of *c* indicate a bias to respond 'self', but the results indicate that they are biased in answering 'other' (*c*=0.65). However, because the participants could only give a response if they already decided it was an Old fact, a separate index was calculated, which resulted in uneven trials between the two decisions, with some trials having fewer responses. Thus, because the two distributions have unequal variance, it limits our interpretations of a continued trial-by-trial analysis of these results.

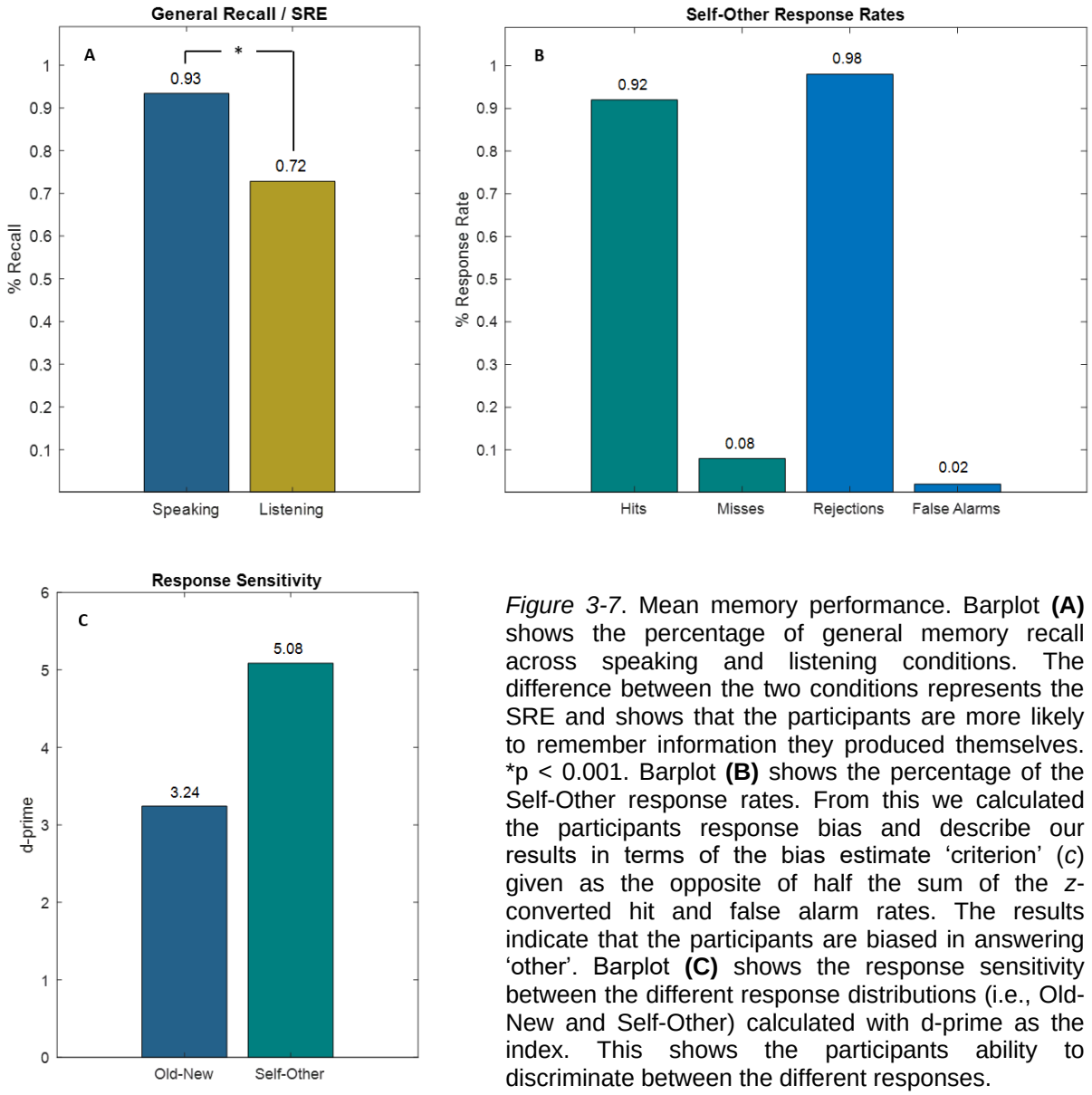


Figure 3-7. Mean memory performance. Barplot **(A)** shows the percentage of general memory recall across speaking and listening conditions. The difference between the two conditions represents the SRE and shows that the participants are more likely to remember information they produced themselves. *p < 0.001. Barplot **(B)** shows the percentage of the Self-Other response rates. From this we calculated the participants response bias and describe our results in terms of the bias estimate 'criterion' (*c*) given as the opposite of half the sum of the *z*-converted hit and false alarm rates. The results indicate that the participants are biased in answering 'other'. Barplot **(C)** shows the response sensitivity between the different response distributions (i.e., Old-New and Self-Other) calculated with d-prime as the index. This shows the participants ability to discriminate between the different responses.

3.6.2 The Effect of Memory Performance on Head Nodding

Figure 3-8 show scatter plots of the degree of head nodding as a function of memory recall for both speaking and listening members of the dyad. The y-axis values for the fast nods represent the fast nod count of individual nods in the 2.6–6.5 Hz frequency band; and the values for slow nods represent the degree of coherence (R^2) in the 0.2–1.1 Hz frequency band. The x-axis values represent the number of correctly recalled facts for each trial, with the data points jittered in the y-axis. The red and blue dots represent the sample median, with the red line showing the key trend from 2 to 3 correctly recalled facts. The dashed blue line shows the trend from 0 to 2 correctly recalled facts but should be interpreted with caution because of the limited number of samples. The unfilled blue circles show outliers.

The trend of the medians for both fast nodding and slow nodding coherence seem to increase with the number of correctly recalled facts – except for the listening condition during fast nods – which would indicate that certain head nodding behaviours correlate with memory performance. We used medians instead of means here because otherwise the between subject variance would mask the within subject variance, which makes it difficult to observe the correct trends. In addition, we removed the one trial in each of the speaking conditions that resulted in 0 recalled facts as there were too few trials for analysis here.

The data on the memory performance was further analysed using multilevel statistical modelling. Prior to model comparisons, we performed a linear multilevel regression on all models for both speakers and listeners. We used two-level models with head nodding frequency as predictors (level 1) nested within participants (level 2). The results from the analysis of the full mixed-effect model are presented in Tables 3-2, 3-3, and 3-4 for both speaking and listening trials.

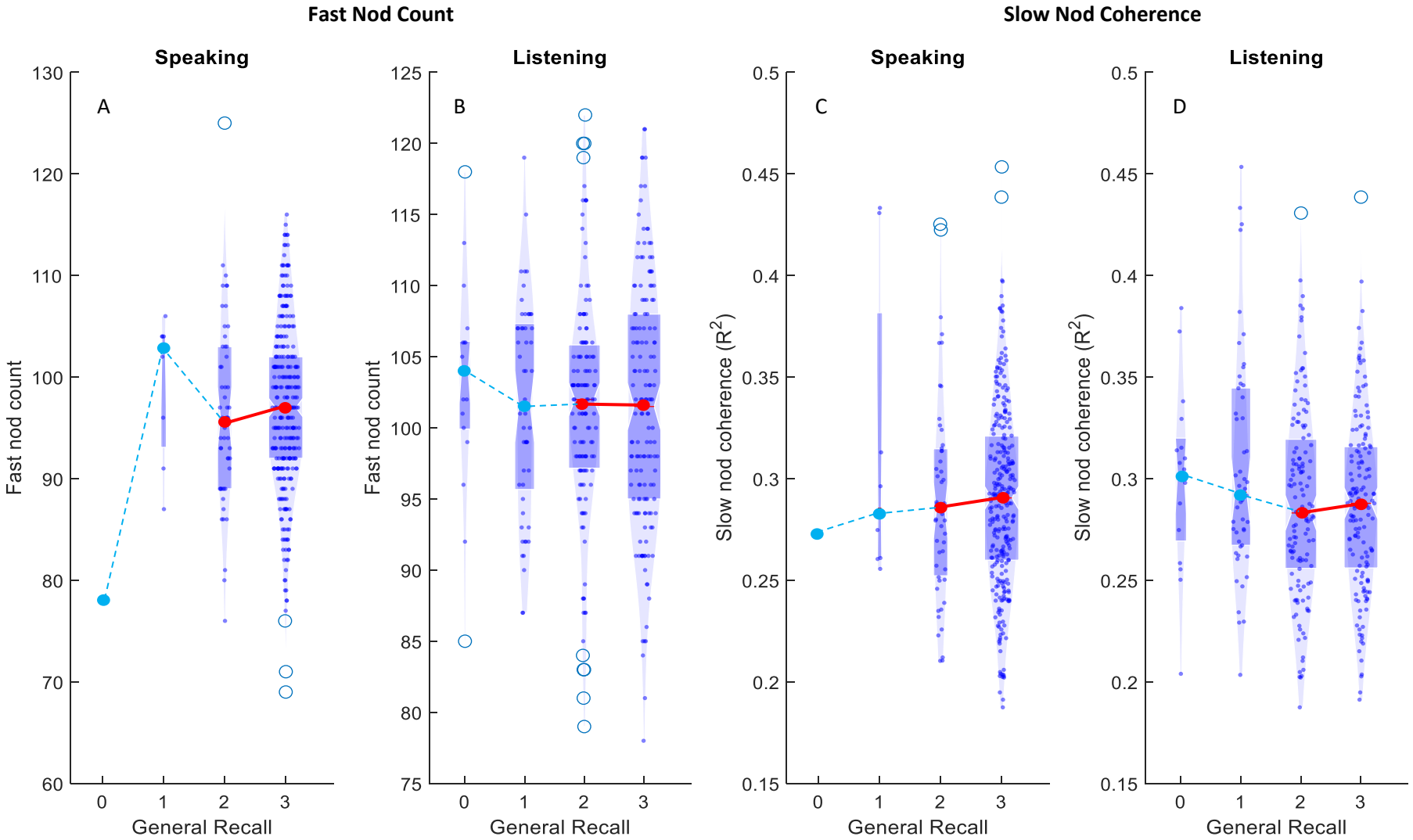
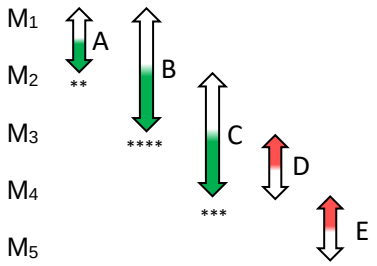
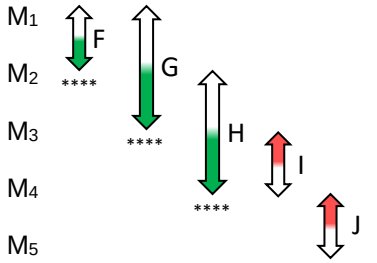


Figure 3-8. Violin scatter plots for general memory recall. The plots show the degree of head nodding as a function of general recall across speaking and listening conditions. The y-axis values for fast nods represent the fast nod count of individual nods in the 2.6–6.5 Hz frequency band; the values for slow nods represent the degree of coherence (R^2) in the 0.2–1.1 Hz frequency band. The x-axis values represent the number of correctly recalled facts for each trial.

Table 3-2

Mixed-effects model comparisons for general recall.

Speaking	Mixed-Effects Model Comparisons
	Recall ~ Participant Recall ~ Fast Nodding + Participant Recall ~ Slow Nodding Coherence + Participant Recall ~ Fast Nodding + Slow Nodding Co. + Participant Recall ~ Fast Nodding * Slow Nodding Co. + Participant
Listening	**Mixed-Effects Model Comparisons**
	Recall ~ Participant Recall ~ Fast Nodding + Participant Recall ~ Slow Nodding Coherence + Participant Recall ~ Fast Nodding + Slow Nodding Co. + Participant Recall ~ Fast Nodding * Slow Nodding Co. + Participant

Coloured arrows represent the best fit model for each comparison measured by differences in AIC-scores, and if the alternative model was accepted (Green) or rejected (Red) in favour of the compact model measured by the likelihood ratio, *$p < 0.05$; **p < 0.01; ***p < 0.001; ****p < 0.0001. *Note:* M_5 includes the fixed effects of fast and slow nodding, as Matlab automatically calculates those when running the interaction model.

Table 3-2 shows the mixed-effects model comparisons for general memory recall, divided by speaking and listening. Results show that, when speaking the inclusion of fast nodding (Arrow A, diff-AIC=5.85, p=.005) and slow nodding coherence (Arrow B, diff-AIC=17.64, p<.0001) significantly improved model fit compared to the null model (M_1). Similarly, when listening the inclusion of fast nodding (Arrow F, diff-AIC=14.28, p<.0001) and slow nodding coherence (Arrow G, diff-AIC=47.12, p<.0001) significantly improved model fit compared to M_1.

For a more robust comparison, we compared both fast nodding (M_2) and slow nodding coherence (M_3) with a saturated model (M_4), which had the combined fixed effects to provide a more natural baseline. These comparisons showed that, when

speaking the inclusion of slow nodding coherence (Arrow C, diff-AIC=11.39, *p*=.0002) significantly improved model fit, but the inclusion of fast nodding did not (Arrow D, diff-AIC=0.4, *p*=.2). Similarly, when listening, the inclusion of slow nodding coherence (Arrow H, diff-AIC=31.37, *p*<.0001) significantly improved model fit, but the inclusion of fast nodding did not (Arrow I, diff-AIC=1.47, *p*=.47).

The last comparisons we made was between the saturated model (M_4) and a fifth model (M_5) which added the interaction effect between fast nods and slow nods in addition to the main effects of fast and slow nods. These comparisons showed that, when speaking there is no interaction effect between fast and slow nodding (Arrow E, diff-AIC=1.05, *p*=.32). Similarly, when listening, there is no interaction effect between fast and slow nodding (Arrow J, diff-AIC=0.57, *p*=.23). The fixed effects estimate for all five models are listed in Tables 3-3 and 3-4.

Table 3-3

Results of the generalized linear mixed models for the predicted fixed effects during speaking

Models	Fixed Effects	β	SE	*t*	*AIC*	*p*
M_1: Recall ~ 1 + (1 | ppt)	Intercept	2.8	0.04	74.5	460	8.4e-224
M_2: Recall ~ 1 + fastnod + (1 | ppt)	Intercept	2.4	0.34	8.02	454	1.5e-14
	Fast Nods	.004	.003	1.31		.188
M_3: Recall ~ 1 + mim + (1 | ppt)	Intercept	2.8	0.16	17.8	443	1.15e-50
	Slow Nods	.037	0.52	0.07		.943
M_4: Recall ~ 1 + fastnod + mim + (1 | ppt)	Intercept	2.4	0.34	7.1	443	4.4e-12
	Fast Nods	.004	.003	1.27		.204
	Slow Nods	-.005	0.52	-.009		.992
M_5: Recall ~ 1 + fastnod * mim + (1 | ppt)	Intercept	3.94	1.6	2.47	444	.0138
	Fast Nods	-.012	0.02	-0.72		.472
	Slow Nods	-5.34	5.4	-0.99		.324
	Fast * Slow	.054	0.05	0.99		.322

Table 3-4

Results of the generalized linear mixed models for the predicted fixed effects during listening

Models	Fixed Effects	β	SE	*t*	*AIC*	*p*
M_1: Recall ~ 1 + (1 | ppt)	Intercept	2.19	0.08	26.3	856	2.1e-86
M_2: Recall ~ 1 + fastnod + (1 | ppt)	Intercept	2.45	0.55	4.48	842	1e-05
	Fast Nods	-.002	.005	-0.5		.613
M_3: Recall ~ 1 + mim + (1 | ppt)	Intercept	2.61	0.26	9.8	809	3.5e-20
	Slow Nods	-1.45	0.87	-1.66		.096
M_4: Recall ~ 1 + fastnod + mim + (1 | ppt)	Intercept	3.0	0.6	5.01	811	8.3e-07
	Fast Nods	-.003	.005	-0.73		.467
	Slow Nods	-1.44	0.87	-1.65		.099
M_5: Recall ~ 1 + fastnod * mim + (1 | ppt)	Intercept	-.584	3.05	-0.19	811	.848
	Fast Nods	.003	.03	1.05		.294
	Slow Nods	11.07	10.4	1.05		.291
	Fast * Slow	-.123	0.1	-1.19		.231

Looking at the fixed effects when participants were speaking, we find that neither fast (M_2, β = .004, t = 1.31, p = .188) nor slow (M_3, β = .037, t = 0.07, p = .943) nodding correlates with memory recall. Looking at the fixed effects when participants were listening, we observe similar results for both fast (M_2, β = -.002, t = -0.5, p = .613) and slow (M_3, β = -1.45, t = -1.66, p = .096) nodding. From the fixed effects of the saturated model when participants were speaking, we find that neither fast (M_4, β = .004, t = 1.27, p = .204) nor slow (M_4, β = -.005, t = -.009, p = .992) nodding correlates with memory recall. Looking at the fixed effects when participants were listening, we observe similar results for both fast (M_4, β = -.003, t = -0.73, p = .467) and slow (M_4, β = -1.44, t = -1.65, p = .099) nodding. We did not find an interaction effect between fast and slow nodding when the participants were speaking (M_5, β = .054, t = 0.99, p = .322) or listening (M_5, β = -.123, t = -1.19, p = .231).

3.7 Discussion

In chapter 2, we identified head nodding patterns that were repeatable across participants and meaningfully related to task performance. We also showed that fast nodding behaviour is found more in conversational contexts in which there is a transfer of new information between participants in a dyad. We hypothesized that fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received new information. We also found evidence to support the idea that the slow nodding behaviour changes across different conversational contexts, and that slow nods were a different signal to fast nods.

In this chapter, we have presented a study that investigates if memory for facts is associated with different types of head nodding behaviours during conversations. We collected a new dataset with a new task to test if different frequency patterns of head nodding behaviour could predict performance on an outcome measure in the form of a memory test. We present results based on the two main hypotheses relating to the fast and slow nodding behaviours. We find ambiguous results depending on what approach we take when it comes to interpreting the relatively novel statistical analysis. We cannot conclusively claim that head nodding correlate with memory during conversations, but some analysis seems to reflect on its importance. In addition, we also explored the potential relationship between slow nodding coherence and self-other overlap during conversations.

We will consider the possible interpretations for our two hypotheses from the framework of the two types of head nodding patterns. We will then proceed to discuss the theoretical implications and limitations of the study, what this could mean for social learning, and how future directions can extend the contribution.

3.7.1 Can Fast Nodding Predict Memory?

In our first hypothesis (**H_1**) related to fast nodding, we predicted that if this behaviour is a listening and backchannel signal for the transfer of new information, it should correlate with increased general recall on the memory test. The results from the model comparison approach using the AIC-scores and Likelihood Ratio Tests for general memory recall (Table 3-2) reveal that the inclusion of fast nodding significantly improved model fit and likelihood ratio compared to the null model, both when speaking and listening. However, for the saturated model M_4, which included every parameter, this comparison showed that the inclusion of fast nodding when speaking and listening did not significantly improve compared to the less complex model M_3, which only included slow nodding coherence as a parameter. From the perspective of the model comparison approach, both fast and slow nodding behaviour seem to matter for memory, because the model which includes these behaviours provides a substantially or significantly better fit than the null model (**M_1**). Such results would partly support our first hypothesis (**H_1**) that fast nodding as a listening behaviour and backchannel signal for the transfer of new information, correlates with learning during conversation as measured on a memory test. In addition to the fast nodding behaviour, the inclusion of slow nodding coherence when speaking and listening significantly improved model fit and likelihood ratio compared to the null model. From the model comparison approach, this also implies that slow nods seem to matter. The model which includes slow nodding provides a substantially or significantly better fit than the null model.

Closer inspection of the model comparisons to the saturated model supports this claim by showing that the inclusion of slow nodding coherence in model M_4 significantly improved model fit and likelihood ratio compared to both model M_2 and

M_3, whereas the inclusion of fast nodding did not. This tells us that adding fast nodding to the model did not increase the maximum likelihood enough to counter the effects of the increase in the number of estimated variables, whereas adding slow nodding coherence did, which hints at slow nods introducing less trivial information than fast nods to M_4.

Lastly, we also investigated whether there is an interaction between fast and slow nodding behaviour on memory recall using M_5 which included both the fixed effects of fast and slow nodding along with an interaction term, to compare it to M_4, which did not include the interaction term. This comparison showed that the inclusion of an interaction term did not significantly improve model fit compared to the less complex model. This would mean that, from a model comparison approach, on average for both speaking and listening, more fast nods and more slow nodding at the same time does not seem to be associated with more effective memory recall. In other words, fast and slow nodding does not seem to depend on each other.

Unfortunately, interpreting p-values for likelihood ratio tests in mixed models are not as straightforward as they are for the linear model. There are multiple approaches with differing opinions about which approach is the best and if there is even a correct way (Hox et al., 2010; Winter, 2013). The model comparison approach (Judd et al., 2008) interprets mixed models as a way to estimate the parameters of interest in each model that best supports or contradicts our proposed hypotheses using the Likelihood Ratio Test as a way to express an alternative model as significantly better fit than a null model without any fixed effects. However, researchers have also criticized this approach saying that the p-values obtained using the likelihood ratio test can be conservative when testing for the presence or absence of random-effects terms and anticonservative (i.e., high Type I error rate)

when testing for the presence or absence of fixed-effects terms (Hox et al., 2010; Winter, 2013). Considering such critique, the likelihood ratio tests may reflect overall goodness of fit but does not show why one model is better than another. For example, the model that includes fast nodding may be considered “less bad” than the model without fast nodding, but we have no way to tell whether the models have a better fit due to increased or reduced nodding, or some other reason.

A second approach is to fit a full mixed model and examine the parameter effects for the individual factors. Here (See Table 3-3, 3-4), we find that when the participants were speaking, neither fast nor slow nodding correlates with memory recall in any of the models. When the participants were listening, we similarly find that neither fast nor slow nodding correlates with memory recall. When comparing the fixed effects results with the scatter plots (See Figure 3-8), we observe a negative trend in the beta estimates of the slow nodding behaviour in listening trials compared to speaking trials, which seems also to be reflected by the trend of the blue line in the scatter plot. However, this trend is not statistically significant.

We can also observe that when the participants are speaking, we seem to get more ceiling effects than when they are listening. For example, a high proportion of the participants remembering 3 out of 3 facts correctly could make the memory discrimination at the top end of the measure impossible. Being more attentive while reading and having access to visual memory of seeing the words written on the cards might explain this. However, looking at the trends for when the participants were listening, we observe greater variance among the number of recalled facts compared to when they were speaking.

In the saturated model M_4, the inclusion of the fixed effects of fast nods and slow nodding coherence for both speaking and listening, was not significant. We also did

not find a significant interaction effect in model M_5, which shows that when the participants engaged in both fast and slow nodding at the same time, it does not seem to be associated with more effective memory recall.

Looking at the fixed effects from each model without comparing them using the model comparison approach, these results do not support our first hypothesis (**H_1**) that fast nodding as a listening behaviour and backchannel signal for the transfer of new information correlate with memory recall during conversation. We conclude that we have ambiguous results that would benefit from different data studies without ceiling effects, but that we can see some indications from the model comparison approach that head nodding might be related to memory recall during conversations. For example, from the perspective of the model comparison approach, we can say that fast nodding does not seem to be an isolated predictor for general memory recall, and that the results rather seem to reflect the importance of the overall engagement of interactive head nodding. In other words, the coordination of both fast nodding and slow nodding coherence during conversations can be associated with general memory recall. However, from a standard mixed-model approach, we find ambiguous results when looking at the individual models fixed effects estimates. Thus, we cannot conclusively claim that head nodding can be correlated with memory during conversations, but some analysis seems to reflect on the importance of head nodding patterns during unstructured conversations.

However, if future studies can provide further insight into this, then using head nods in a conversation can potentially signal to the other person involved that the conversation is working and that we are learning as we nod along. The other person can then use this as a prediction to help change the strategy of the interaction. These potential findings would lend support to the belief that changes in social

signals like head nodding can provide us with clues about our cognitive abilities, and that the different frequencies of head nodding behaviour can predict other cognitive outcomes, such as for example better memory performance. That could in turn lead to the investigation of what other kind of learning and memory processes are associated with various social behaviours. For example, in his book “Honest Signals” Pentland (2010) describes these subtle involuntary non-verbal patterns as not just a complement to language, but as a separate communication network which provides a window into our intentions, goals, and values. If we understand this old channel of communication, he claims, we can predict the outcomes of many social situations.

In the introduction to this chapter, we mentioned a few studies that have used social signals to predict outcomes in social interactions, for example during lectures (Chen et al., 2015; Pinzon-Gonzalez & Barba-Guaman, 2021; Sümer et al., 2021), and with job employment (Gifford & Wilkinson, 1985; McGovern et al., 1979). Researchers have also tried to use non-verbal behaviour to predict student learning and selling software used to provide feedback for instructors to change their teaching styles (Krithika & Priya, 2016). Much of this research rely on the concept of attention, since it is closely related to learning performance (Chen et al., 2015). The benefits of attention for engagement and learning are clear, but the term ‘attention’ is closely related to many different neurological and cognitive processes, which are beyond the scope of this thesis to cover. Thus, different forms of attention backchannels (i.e., gaze, head nods) can result in different information being received and encoded. For example, Richardson and Dale (2005) demonstrated a link between interpersonal coordination and eye gaze patterns, which influenced the type of information that was stored. These variations in different types of attention backchannels and levels of information, that can be stored within any given modality, is an area or research

that needs to be addressed to organise how information is sent and received differently depending on social context. We have investigated one such aspect in this study, namely if fast nodding is a listening and backchannel signal for the transfer of new information. Our results are ambiguous, which partly show that frequencies of head nods can be correlated with memory using the model comparison approach, but also that no such correlation exists when we examine the mixed effects model.

To find 'common ground' (Clark, 1996) in a conversation with the use of backchanneling, we must then acknowledge not just the multimodal nature of social interactions, but also what pre-determined knowledge people bring to the conversation. For example, as the participants in our study arrive with the predetermined knowledge that this is an experiment, they may go into the conversation less engaged and interested than what they would be if it was a real goal-directed conversation. This may in turn affect their level of attention, and consequently they may use fast head nodding just to signal fake engagement because they want to get the task done quickly and with as little effort as possible. Hence, both participants in a dyad may just "play along" with the interaction, which is one way to explain the non-significant results in our mixed effect model.

Even if the participants indeed are engaged and attentive to what is being discussed, how can we measure if they agree or not? Agreement occurs when they achieve some form of common ground, but the way this study was set up, it did not leave much room for the participants to have opinions about the facts presented until the second part of the task (i.e., dialogue). For example, one participant may not agree about facts unless the other participant challenges this as a belief with his or her own opinion. This opinion, determined by the participants beliefs and goals, is not shared but must be transferred to the other participant, which may or may not

receive this information as new to them. And since this opinion is not necessarily just about the fact, but embedded in their beliefs and goals, it can be difficult to know if the listening participant is nodding to signal being attentive, engaged, or if they agree with what is being said. In other words, it may be difficult to know if the listening participant nods because he signals receiving new information, or if he nods because he agrees with old information.

While verbal signals are defined with a set of rules needed for their understanding, non-verbal signals usually do not have a clearly defined vocabulary, which raises the uncertainty and ambiguous nature of decoding such signals (Vinciarelli et al., 2012). As such, we think that additional contextual information is needed to understand head nodding as an intrinsically ambiguous channel for social signals. However, we think the best way to do this is to measure multiple verbal and non-verbal signals extracted from multiple modalities and channels and find ways to integrate and examine relationships between them. This way we can find out whether a person dedicate several channels to the same source (e.g., listening and gazing at the same person). Within the scope of this thesis, we are presenting the analysis of a single modality (i.e., head nodding), but this signal is captured from a rich multimodal setup where we recorded other modalities (i.e., gaze, facial expressions, and speech). This means that our head nodding signals are performed and captured in a more naturalistic setting and social context than previous studies that only look at head movements in isolation. More research on the relation between memory and head nodding behaviour is needed, and since we have a rich multimodal dataset to work from, we are presented with the option to go back and analyze the relation between head nodding and gaze behaviour, facial expressions, or speech, in future studies.

3.7.2 Can Slow Nodding Predict Self-Related Memory Encoding?

Our second hypothesis (**H_2**) related to the coherence of slow nodding, is based on the idea that because slow nodding coherence is believed to be used to socially bond or get on well with others, either through mimicry or joint attention (i.e., the 'social glue hypothesis'), it should lead to a larger self-other overlap (i.e., increased feeling of closeness) between people. From this we predicted that slow nodding coherence should correlate with self-related effects and biases, measured by how biased they are in claiming that it was them reading a fact even if it was not (i.e., SBE). By examining the results in Figure 3-7, we can see patterns of SRE and SBE.

Self-Reference Effect (SRE). We quantified the relative contributions of self vs. other referenced information by dividing the general recall responses into speaking and listening trials. The results show that there is significantly more general memory recall when speaking compared to listening (Figure 3-7A), which indicates that the participants are more likely to remember information that is linked to themselves, giving support to the existence of the SRE for memory. However, it is also possible that this effect can come from being more attentive by having access to the visual memory of seeing the words written on the cards. This allows for more opportunities to make associations to new information thus facilitating more elaborate encoding. On the other hand, having access to visual memory can also act to relieve the burden of memory for speakers as they are prone to the experimental and social pressure of not making mistakes when reading as opposed to understanding the facts being read (i.e., passive reading). Since we used a surprise post-hoc memory tests instead of telling the participants that they were going to do a memory test at the end, this effect may have been especially prominent in the monologue part of the task. However, the participants knew they had to engage in dialogue eventually.

Some theories postulate that linking information to the self should be affected by goal-directed dynamics (Conway, 2005). That is, information is more likely to be remembered when receiving information consistent with long or short-term goals. In terms of the SRE then, facts that one is reading should be more memorable because information associated to the self is important and should not be forgotten.

Self-Bias Effect (SBE). The results from the mean memory performance (Figure 3-7B) show that the participants are good at identifying who was reading during the encoding process, with high response rates for correct responses (i.e., hits and correct rejections) and low response rates for incorrect responses (i.e., misses and false alarms). However, it is possible that the participants could have just responded randomly when they were not sure of who was reading during the encoding process. In addition, the speakers could also have been more attentive, similar as with the SRE, by having access to visual memory when reading the facts.

To test their sensitivity to *why* they gave a specific response, we calculated their ability to discriminate between their answers of 'Self' or 'Other', and the results from this analysis showed that the participants demonstrate very good sensitivity to the measure (Figure 3-7C). From the response rates, we also analysed the participants willingness to claim that it was them reading a specific fact when it was not. The results showed a positive response bias (i.e., criterion), which means that the participants were biased in answering 'Other' rather than 'Self'. This result is the opposite to what Russel and Jarrold (1999) found in their study, where they reported a significant bias towards the 'Self', where the participants were more likely to claim "I placed it" even if they did not. In contrast to their study, our results show that our participants identify themselves more with the 'Other' and have a more interdependent self-construal (Brewer & Gardner, 1996). People with an independent

self-construal generally feel closer to others (Aron et al., 1992; Holland et al., 2004). Studying aspects of an individual's self-construal can prove to be useful when exploring the relationship between mimicry and self-other overlap. One of the difficulties in much of the early work about the self was that it depended on subjective reports, like the Inclusion of Other in the Self Scale (IOS) (Aron et al., 1991). In this study, we used more behavioural measures of memory recall in relation to two well known memory effects associated with the idea of self-construal. This has proven to be useful and improves upon the subjective measures in the field.

We also wanted to test, in an exploratory fashion, whether real-world mimicry, or slow nodding coherence is associated with these self-related effects and biases. However, because the participants could only give a response if they already decided it was an Old fact, a separate index had to be calculated, which resulted in uneven trials between the two decisions, with some trials having fewer responses. Thus, because the two distributions had unequal variance, we were not able to perform a full trial-by-trial analysis on this data.

According to the Source Monitoring Framework (Johnson, Hashtroudi, & Lindsay, 1993), it is also possible that the SBE can be influenced by the SRE in the sense that recalling 'who' could also be related to how well the encoding process binds or encourages attention to self-associated features of the facts, such that they later can serve as cues for correct source identification. For example, if during encoding a specific self-association is created (reading a fact) it might later be activated when prompted with the question to recall the source during the encoding. This way, the SBE could instead of being identified as an action performed by either the self or the other, simply be identified as an action performed by either the self or *not*-the-self.

Furthermore, throughout this chapter, we have referred to the participants as ‘speakers’ or ‘listeners’ to either new or old facts about American states. This may have consequences for interpretation as speakers can be seen as ‘readers’ of facts that are ‘known’ to them. To clarify the distinction, a ‘speaker’ is the same as a ‘reader’ only that they are conveying the information to their partner by speaking. At this point during the participants’ task, all facts were considered ‘new’ to them. That is, we did not control for the participants knowledge of American states, even though we tried to limit this by designing the facts to be uncommon. When it was time for the memory test, both speakers and listeners were asked to make a judgement on a prompt to recall if a fact was Old (a fact they could remember from the task) or New (a fact they did not remember from the task). This is similar to what Rogers et al. (1977) did when they presented trait adjectives to participants who later were asked to recognize them from a list of words as either Old (i.e., a word on the list) or New (i.e., a word *not* on the list).

3.8 Limitations and Future Directions

Our data has some methodological limitations and implications. In relation to our first goal, episodic memory as we have used it in this study is an all-or-nothing measure, which could result in ceiling effects. For example, a high proportion of the participants remembering 3 out of 3 facts correctly could make the memory discrimination at the top end of the measure impossible. Memory tasks may also introduce primacy effects and recency effects (Morrison, Conway, & Chein, 2014) because participants may be more likely to remember the first or last fact than the middle fact when taking the memory test. This means that item order could have been included as a random effect in the models to account for the order that each fact was presented to the participants.

In relation to our second goal, there are some important design implications which we have learnt about in the present study, and which will help with future studies. As previously mentioned, it was not possible to explore the second hypothesis **H_2** using multilevel statistical modelling, which limited the interpretations of our results. Instead, we decided to focus our efforts on the main goal of testing **H_1**. This study could thus benefit from future studies on how to equate the number of responses for each trial to properly test **H_2**.

We should also not forget that what people remember is almost always a product not just of the original encoding process, but also the events happening in between the encoding and the recall process. Hence, any act of remembering must be viewed as having a social history. Such history often includes input and output from multiple modalities that is dynamically integrated over time. We employed the same multimodal data collection framework to the one we used in Chapter 2, with high-

resolution motion capture and wide synchronization of socially relevant data. In this study, we focus only on head nodding behaviours, but having a rich multimodal framework to work from make this dataset a valuable tool for potential future studies.

However, having such an extensive multimodal setup also introduces some degree of errors when it comes to labelling the data, but so does manual annotation methods to a greater extent. We excluded 7 dyads because of extensive errors. However, this loss can be compensated for by the large number of datapoints in the remaining sample. With the use of linear interpolation between the signals to remove NaNs so that the wavelet toolbox in Matlab could post-process the wavelet transforms, the errors in the remaining sample only affected approximately 500 out of 100.000 (0.5%) datapoints. Hence, by using this more automated method, we get a lot of data in the final analysis compared to manual annotation methods.

It is important to note that the hypotheses in this study are based around correlating measures. As such, we cannot claim any causal connections between the variables. For example, the AIC values are simply a way of ranking the models, and while a best model was found and a relationship between head nodding and memory performance was estimated using Likelihood Ratio Tests, we are unable to claim that learning causes nodding, or vice versa; that the absence of nodding equates to no learning. Analysis of the fixed effects coefficients also gives us conflicting results when it comes to making any firm conclusions regarding the role of fast and slow nodding behaviour and its relationship to memory performance, but the model comparison approach seems to reflect the importance of head nodding patterns during unstructured conversations.

Future studies can perhaps illuminate this issue, and our results could help generate computational models of more socially realistic virtual agents. It could also

be applied to improve the automatic detection of social signals, and to create an algorithm to artificially generate natural backchannels as behaviour rules. By programming head nodding signals to be used by virtual agents that can interact with participants, we can find ways to test how they respond to these behaviour rules to drive the development of new and improved psychological theories for social interaction. This will be the focus of the next study in this thesis (Chapter 4), where we continue to investigate the role of head nodding behaviour and memory by manipulating virtual interactive engagement. The aim is to find evidence to support the idea that changes in non-verbal social signals like head nodding can provide us with clues about our cognitive abilities, which in turn can lead to the ability to predict the outcomes of many social situations.

The experiment used in this study also utilizes non-structured (i.e., free flowing) conversation in addition to more structured conversations, whereas most of the previous work have used strictly structured conversations. Structured conversations, when the participants have predefined actions, have the benefit of better control of the conversational roles (i.e., speaker/listener), but results from such tasks are often not applicable to more free-flowing conversations between two people, and does not translate well to experimental setups using virtual agents. Unstructured conversations better represent how people interact and has the potential to capture different conversational dynamics between participants. In future studies we could consider *when* the nods occur during a conversation by going back to the recordings to manually transcribe verbal phrases from the participants. By doing this, we could look at what time a nod occurs during a phrase and if, for example, a nod at the end of a phrase signals the end of a turn. This may also account for the ambiguous results found in this study but was beyond the scope of this thesis.

Together with our high-resolution experimental setup used in Chapter 2 and here again in Chapter 3, we can capture these fine-grained non-verbal behaviours. Chapter 4 will give further insight into creating simple behavioural rules from head nodding signals and use these in virtual agents to manipulate the social interaction.

3.9 Conclusions

From the results in our previous study (Chapter 2), we posited that fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received this new information. In this study, we explored this idea by investigating whether we are more likely to engage in fast nodding when we recall more information from a conversation, by testing if different patterns of head nodding behaviour correlate with memory performance. The second goal was also to explore the link between slow nodding coherence and the idea that it is used to socially bond and get on well with others through a larger self-other overlap.

We present results based on these two hypotheses related to both head nodding patterns, and we find ambiguous results depending on what statistical approach we implemented. We cannot conclusively claim that head nodding can be correlated with learning and memory during conversations, but some analysis suggests that head nodding behaviours might be related to memory during conversations. Despite the limitations of not being able to fully explore the relationship between slow nodding coherence and self-other overlap, we found that people are more biased in claiming that their partner was reading a specific fact as opposed to themselves.

The findings of this study can provide useful insights into the natural parameters of head nodding behaviour in unstructured conversations, and its relationship to memory. We also demonstrate methodological solutions to identify a specific social behaviour, and link that to a fundamental cognitive function. This can in turn allow us to build interactive virtual agents who can simulate these natural backchannels to test how people respond to different types of interactions and drive the development of new and improved psychological theories of social interaction.

Chapter 4. Artificially Generating Head Nodding Signals in Virtual Agents to Test if Interactive Engagement Promotes Learning and Liking

4.1 Abstract

In Chapter 3, we found ambiguous results and could not conclusively claim that head nodding could be used to predict memory performance, but some analysis gave us indications that some head nodding frequencies might be related to memory during conversations. In this study we continue this investigation and demonstrate the benefits of using virtual reality to artificially generate behaviour rules in interactive virtual agents based on real-world behaviour data to test a causal link between interactive engagement and memory.

To test this, the participants interacted with both an interactive and non-interactive virtual agent. To drive the behaviour of the interactive agent, we programmed and generated measured head nodding parameters under different behaviour rules so that they could generate high resolution head nodding behaviour based on natural backchannels and mimicry behaviour when interacting with the participants. As a control condition, the other virtual agent was non-interactive, only driven by pre-recorded motion. This was further carried out as a Wizard of Oz (WoZ) experiment.

We obtain two outcome measures – learning facts about American states and liking ratings for the two virtual agents. We found no significant effect of agent-interactivity between interactive and non-interactive agents. However, we did find that people that are speaking are up to three times more likely (i.e., odds ratio) to remember information during a conversation compared to those who are listening. However, the effect seems less prominent than during real interactions. Further results show no significant differences in feelings of rapport between the interactive and non-interactive virtual agents, and no reliable correlations between ratings of liking and memory performance was found.

4.2 Introduction

In Chapter 3, we used high resolution motion capture, wavelet analysis, multilevel statistical modelling, and a model comparison approach to explore whether participants were more likely to engage in fast nodding when they recall more information from a conversation by testing if such patterns of head nodding behaviour could predict memory performance. We found ambiguous results and could not conclusively claim that head nodding could be used to predict memory performance, but some analysis gave us indications that some head nodding frequencies might be related to memory during conversations. However, we built this study on the preliminary results of Chapter 3, which initially showed a correlation between head nodding and memory. The aim of the present study was therefore to further investigate a causal link between interactive engagement with a virtual agent and memory performance. Further analyses did not support a correlation between fast nodding and memory performance, but for transparency we are presenting the idea of this chapter on the assumption that such a correlation exists, and that it was what initially led us to the methods of testing a more causal link between head nodding and memory. This is important as it is consistent with how we use our approach to reverse-engineering interpersonal coordination by *measuring* and *analysing* behaviour to determine which parameters (i.e., behaviour rules) are important, and then use these parameters or rules to engineer or *artificially generate* that behaviour in a virtual agent to test our hypotheses. For example, finding a correlation between two measures can guide our search for finding behaviour rules based on real-world behavioural data that we can then use to manipulate in virtual agent models to systematically test our hypotheses.

Using virtual reality, it is possible to create virtual agents who have different motion patterns and characteristics (Pan & Hamilton, 2018). Here, we can use the movement data from our previous studies to build a virtual human with high resolution head nodding behaviour based on natural backchannels and different behaviour rules. In conjunction with a Wizard of Oz (WoZ) system for conversation, this will allow a virtual agent to have the role of 'conversation partner' in the American-states discussion task used in Chapter 3. We can then evaluate if people learn more from virtual agents who show natural or interactive nodding behaviour.

Changing non-verbal behaviour of virtual characters, such as eye gaze and mimicry, has been shown to have significant effects on people's perception and attitude toward them (Bailenson & Yee, 2005). Thus, the distinction between our previously established fast and slow head nodding patterns can perhaps also be used to measure how much we like a virtual agent. The second aim of this study is therefore to investigate if interacting with an interactive virtual agent, driven by natural head nodding, can enhance feelings of rapport, or is preferred, over a non-interactive virtual agent with pre-recorded motion. Following this, we also want to answer the question of whether we learn better from virtual agents that we like, by investigating a link between feelings of rapport and memory performance.

This research can advance virtual agent technologies beyond previous studies (Bailenson & Yee, 2005; Pan, Gillies, & Slater, 2008; Verberne, Ham, Ponnada, & Midden, 2013), leading to more realistic simulations of virtual agents with a grounding in psychology. Hypothesis testing can then also benefit from seeing how people respond to these virtual agents. This highlight how new technologies and experimental designs can be used as a tool to further test and challenge our theories

in the field. In the next sections, we review how virtual head nodding are generated (4.2.1), and how to use interactive virtual agents to test our hypotheses (4.2.2).

4.2.1 Artificially Generating Head Nodding Signals

One of the most immersive experiences in virtual reality is one that involves social interaction with virtual agents. Such agents, although they are only simulating real appearances and behaviours, can interact with real people in a way in which they can appear to be alive. The rapid emergence of consumer virtual reality is driving a need for virtual agents, which can use social signals such as body movement and facial expressions. Probabilistic graphical models (Koller & Friedman, 2009) have traditionally been used to generate virtual agents that respond in a more natural way, corresponding to how a real person would interact during social interaction. The main challenge is to dynamically generate animation that is highly responsive without unnatural jerkiness, and which matches current models of coordination behaviour.

Previous studies have also used various techniques to create characters that can generate simple social behaviours such as body postures (Gillies et al., 2015), hand gestures (Kopp & Bergmann, 2013), and head nods (Huang et al., 2010). Such systems traditionally replay the measured movements with good timing but cannot yet create the detailed coordination found in real social interaction, nor have they been linked to psychological theory.

With high resolution motion tracking technology (See Section 1.3) we can capture the actual head nodding behaviour displayed by real people and use this to drive the head nods of virtual agents in a controlled manner. In combination with virtual reality technology, we then have the ability to manipulate the social interaction by changing

the visual cues or signals in order to represent the behaviour of the virtual agent that we want the participants to be able to observe. By using the positional head-trackers of the VR-headset, we can also program specific responses of the virtual agents that is conditional on the participants' behaviour – for example to copy a head nod.

However, the theories behind these virtual responses are often undefined, which makes it difficult to use these virtual agents as models for hypothesis testing. That is, when trying to implement natural backchannel responses most research papers focus on the timing on the behaviour, but do not describe the behaviour. As we have shown, differences in the frequency of head nods have different meanings and should thus be implemented under different behavioural rules to the virtual agent.

These behavioural rules, which govern the transition from the movement data to the generated virtual behaviour is covered in more detail in Section 4.4.2. In short, the decision on how the behaviour is implemented in the virtual characters is based on two rules: (1) If the virtual agent is listening or when the participant is speaking, it performs fast nodding to simulate a natural backchannel; and (2) based on research that we mimic head nods with a 600 ms delay (Hale et al., 2020), if the participant does a big nod, the agent will nod after a 600 ms delay simulating natural mimicry.

4.2.2 Hypothesis Testing using Interactive Virtual Agents

Along with a corresponding rise in "second-person" neuroscience, which studies the brains of two people in real-time interaction, the study of social neuroscience in more embodied and extended reality settings has been gaining in prominence in recent years (Schilbach et al., 2013). This field of research expands on the belief that is so common to experimental social psychology, that human social signals are not

restricted to the real world. For example, behaviours exhibited by virtual characters in virtual environments involve social signals, although perhaps of a different nature.

Blascovich et al. (2002) emphasise that a general assumption that seems conspicuous in all psychological research is that experimental manipulations of perceived (i.e., real-world) and imagined (i.e., written scenarios) stimuli are essentially equivalent for understanding psychological processes. Manipulating imagined stimuli costs less, requires less effort, and provides a greater degree of experimental control (i.e., precise manipulation of variables). However, a greater degree of experimental control often comes at the cost of ecological validity (i.e., the extent to which an experiment is like situations encountered in real life). Thus, a trade-off typically exists between experimental control and ecological validity. Recent technological advances in both motion capture and virtual reality systems have allowed researchers to lessen this trade-off by facilitating an increase in ecological validity without entirely sacrificing experimental control, or vice versa.

Studies using virtual reality as a tool have shown that people keep appropriate social distance from virtual agents (proxemics) (Bailenson et al., 2003), and mimic their behaviours (Vrijsen et al., 2010). Given these reactions, we can use VR to test conversational outcomes of social interaction. For example, a range of studies have successfully replicated psychological constructs with virtual reality, including trust (Hale & Hamilton, 2016b; McCall & Singer, 2015), prosociality (Gillath, McCall, Shaver, & Blascovich, 2008; Hale et al., 2020), and social anxiety (Pan et al., 2012).

Virtual reality has also been used to study behavioural mimicry. Mimicry emerges during infancy and has important roles for social learning and affiliation (Over & Carpenter, 2013). It is also widely believed that mimicking another person has positive consequences for social interaction and has therefore been described as a

‘social glue’ (Lakin et al., 2003). Bailenson and Yee (2005) have demonstrated positive effects of being mimicked in virtual reality. In their study, they tracked the participants’ head movements and programmed the virtual character to either mimic their movements or made head movements of their own that were previously recorded from another experiment. Participants who were mimicked rated the character as more likeable. However, they did not provide any explanation on how these mimicry ratings were weighted. In another study on virtual mimicry (Verberne et al., 2013), the authors found a positive effect of trust for one of their virtual agents that mimicked the participants’ head movements. However, with another agent the same mimicry manipulation did not lead to any positive effects. These studies provide mixed results between mimicry and liking, which follow the same patterns as ratings of liking in traditional research settings where human confederates were trained to mimic participants (Chartrand & Bargh, 1999; van Baaren et al., 2004). However, there are many reasons to be cautious of accepting these naturalistic studies of mimicry at face value. First, both effect sizes and experimental power in many previous studies have been small, and false positives may be present in the literature (Hale & Hamilton, 2016a).

A study by Hale and Hamilton (2016b) aimed to establish a similar paradigm in which participants could be mimicked by a virtual agent and use that to test the parameters that are important for social interaction, such as the timing and spatial form of mimicry, as well as various social characteristics of the mimicker. They reported mixed results in line with previous studies of mimicry using virtual reality in favour of no significant effects of virtual mimicry on rapport. This finding casts doubt over a strong version of the social glue theory, in which all types of mimicry have positive social effects. These results led them to the conclusion that we cannot make

the claim that being mimicked leads to changes in social evaluation, and we should take caution in accepting this dominant view and take note of its fragile effects.

In this study, similarly to what Hale and Hamilton (2016b) did, we want to follow up on previous results using virtual reality to see if there is a way of implementing our previously established head nodding patterns onto a virtual agent and use this as a model to test whether there is a causal link between interactive engagement and memory performance. We also want to investigate if interactive head nodding can be used to measure how much we like a virtual agent, by taking inspiration from Bailenson and Yee (2005), who has demonstrated that changing patterns of eye gaze and mimicry have significant effects on people's perception and attitude towards them. If we find support for this in terms of head nodding behaviour, we could use it to predict and quantify how much we like a virtual agent. Lastly, we want to follow up on this by trying to answer the question of whether we learn better from agents that we like by finding a link between feelings of rapport and better memory.

4.3 The Present Study

In Chapter 3, we found conflicting evidence regarding whether participants were more likely to engage in fast nodding when they recall more information from a conversation and whether this relates to memory performance. We found ambiguous results and could not conclusively claim that head nodding could be used to predict memory performance, but some analysis gave us indications that some head nodding frequencies might be related to memory during conversations. In this study we continue this investigation from preliminary results of Chapter 3, which initially showed a correlation between fast nodding and memory. In this study, we aim to test a causal link between the two previously correlated measures. Within a virtual reality paradigm, we can use the movement data we gathered in our previous studies to generate high resolution head nodding behaviour based on natural backchannels and different behaviour rules.

To test our hypotheses about social interaction, we set up two virtual agents to take part in a conversation about American states, matching the task used in Chapter 3 (See Section 4.4.2 for more details). In this task the participants, together with the virtual agents, alternated turns to read cards with facts on them relating to American states, presented on virtual tablets. The participants interacted with both the interactive and non-interactive virtual agent, in a counterbalanced within-subject design (See Section 4.4.3 for more details).

One agent was programmed to be fully interactive driven by the nodding behaviours found in our previous studies. As a control condition, the other virtual agent was non-interactive, only driven by pre-recorded motion. Both agents have the same pattern of gaze and lip-syncing behaviour, as well as pre-recorded synced-up speech segments from two voice actors. This means that we can systematically and

specifically manipulate the behaviour rules of head nodding behaviour, while keeping the other behaviours constant. To drive the behaviour of the interactive agent, we programmed and replayed the measured head nods under different behaviour rules so that they could perform social behaviours when interacting with the participants.

This was carried out as a WoZ experiment, which is a paradigm in which the participants interact with a virtual agent that they believe to be autonomous, but which is being partially controlled by a human being. In this study, the human is the experimenter, who is controlling the verbal component of the interactions to make the dialogue flow as naturally as possible. This paradigm comes with its own set of limitations, which will be discussed later. Overall, this approach means that we can create more naturally behaving virtual agents who are able to maintain a conversation on set topics and show gaze and head nodding behaviour. Our experimental design manipulates nodding behaviour, so that participants meet one agent with the natural nodding behaviour which is contingent on the participant's own speech and actions, while the control agent has pre-recorded head motion.

In this study, we obtain two outcome measures – learning facts about American states and liking ratings for the two virtual agents. These allow us to examine three hypotheses, which we will cover in the next section (4.3.1). We then compare how participants interact in a conversation with these virtual agents and measure their performance with a post-hoc memory test (See Section 3.4.3 for more details) to assess how many facts they remembered correctly from the conversation.

4.3.1 Aims, Hypotheses and Predictions

The first aim of this study is to find a way of implementing our previously established head nodding patterns on a virtual character, and use this character (i.e., interactive agent) as a model to test against the non-interactive agent whether there is a causal link between interactive engagement and memory. The second aim of this study is to investigate if interactive head nodding can be used to quantify how much we like a virtual agent, and to answer the question of whether we learn better from agents that we like. We will be testing three hypotheses based around these aims:

H_1: *Interactive engagement promotes the encoding of new information.*

H_2: *Interactive engagement promotes liking.*

H_3: *Liking is linked to learning.*

The first hypothesis (**H_1**) claim that interactive engagement promotes the encoding of new information and is based on a continued investigation of findings from Chapter 3, that some head nodding behaviours relate to learning new information during conversations. Testing this hypothesis, we predicted (**P_1**) that the conversation with the interactive virtual agent with interactive head nodding behaviour should lead to increased general recall on the memory test compared to the non-interactive virtual agent, measured by how many facts the participants remember. More specifically, if **H_1** is true, then we expect that participants will be more likely to encode new information from a conversation when they use head nodding as a signal and should reflect the degree to which we learn more through interactive engagement. Important to note is that due to the difficulty of manipulating head nodding in the participants, this is an indirect manipulation of head nodding

behaviour since we manipulate the virtual agent and its visual signals, rather than the participants' head nodding behaviour directly. By doing this, we change the dynamic of the conversation, and consequently the response (i.e., head nodding) from the participants. This is similar to what we did with the pseudo controls in Chapter 2 when we matched data from different trials within the same pairs of participants. Using pseudo-matching changed the dynamic of the conversation, and consequently how the participants behaved, while still maintaining the participants individual movement characteristics. Thus, we are both looking at the role of head nodding as visual feedback during conversation, but also investigating how head nodding movements impact the engagement in interaction between the human participant and the virtual agent.

The second hypothesis (**H_2**) claim that interactive engagement promotes liking and is based on the idea that changing non-verbal behaviour of virtual agents, such as eye gaze and mimicry, has been shown to have significant effects on people's perception and attitude (Bailenson & Yee, 2005). Testing this hypothesis, we predicted (**P_2**) that the virtual agent with interactive head nodding behaviour should lead to increased feelings of rapport compared to the non-interactive virtual agent, measured by the sum of the questionnaire ratings for each agent.

The third hypothesis (**H_3**) claim that liking is linked to learning, and states that we learn better from virtual agents that we like. Testing this hypothesis, we predicted (**P_3**) that there should be a link between feelings of rapport and increased general recall on the memory test, measured by correlating ratings of liking on the questionnaire with how many facts the participants remember. More specifically, if **H_3** is true, then we expect to be able to use measures of how much we like a virtual agent to predict how much we learn from them.

4.4 Methods

4.4.1 Participants

32 participants (M_{age}=24) were recruited from the UCL Psychology Subject Pool and the ICN Subject Database. Exclusion criteria included subjects that were not fluent in English. The participants did not have any previous experience with the tasks and were unaware of the purpose of the experiment. Ethical approval was arranged via the UCL Research Ethics committee, and all participants gave their written informed consent. All participants were informed about the potential side-effect of motion sickness from wearing the headset, and that they could end the experiment at any time. A monetary reimbursement was offered for participating at a rate of £7.50/hour.

4.4.2 Equipment

Virtual Reality System and Environment. All participants were engaged in a virtual conversation task with two virtual agents named Anna and Beth. Dr. Nadine Aburumman created the virtual environment and the characters using the Unity 3D graphics engine (Unity Technologies, 2020). Her report on the pilot study is published in (Aburumman, Gillies, Ward, & Hamilton, 2022). The participants sat at a desk in front of a projector screen (Figure 4-1A, B), and the virtual characters were programmed to appear seated on the other side of the desk facing them (Figure 4-1C, D). We used an HTC Vive Pro Eye VR Headset to display the virtual environment to the participants. The headset has built-in stereo headphones where they could hear the characters speak and get audio instructions.

Figure 4-1. Lab Setup. The experimental setup (A). Third person view of a participant wearing the HMD (B). The participants view of the virtual room with Anna (C) and Beth (D).

Two female virtual characters were created for the experiment, named Anna and Beth. The characters appearances (Figure 4-2) were rendered similar to the average age as the participant population. Apart from Anna having brown and Beth having blonde hair, both were similar in appearance, with big cartoonish eyes.

Figure 4-2. Virtual character appearances. Anna (left) has brown hair and brown eyes. Beth (right) has blonde hair and blue eyes. Both characters were programmed to be similar to the average age as the participant population (i.e., 24).

The reason for creating the characters in a cartoonish style comes from what we discussed in the introduction of the thesis about the ‘uncanny valley’ (Section 1.5.1). For example, how realistic can a virtual character be before we start perceiving them

as uncanny? Designing believable virtual characters is challenging, and in a study by Zell et al. (2015) the authors demonstrate how we can get out of the uncanny valley by relying on stylization to increase the appeal of a character by exaggerating or softening specific features (Figure 4-3). Thus, to make our characters appear as less uncanny, we chose to render them with a more cartoonish stylization.

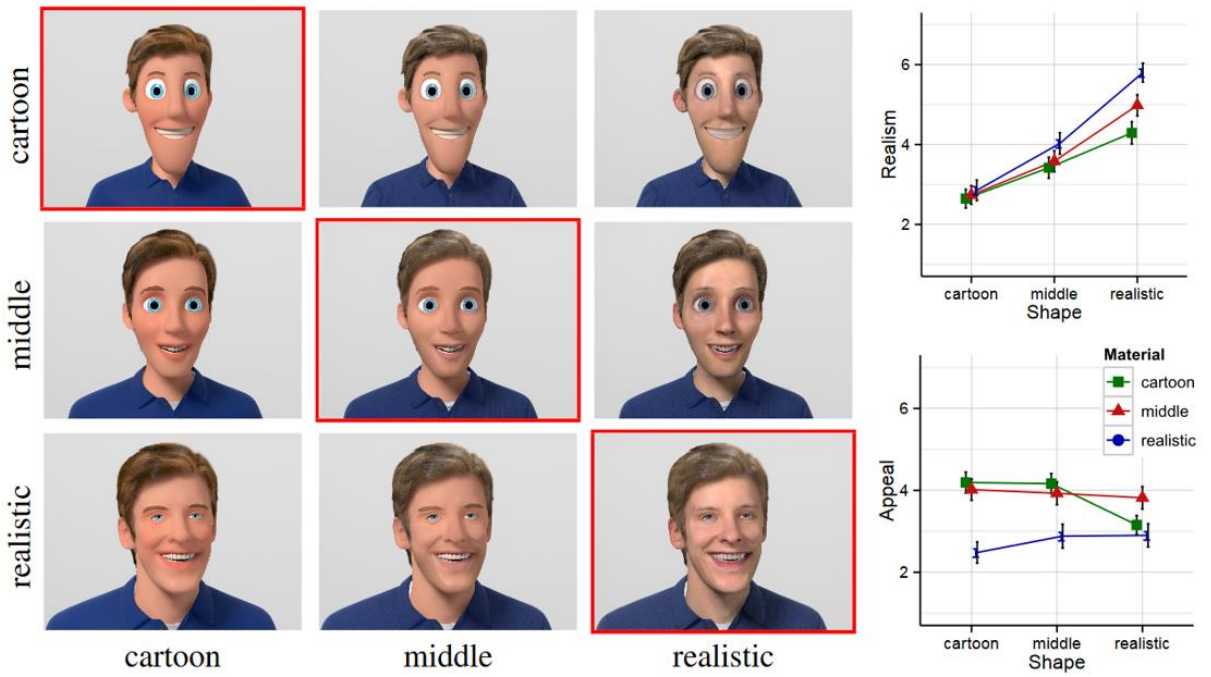


Figure 4-3. Three levels of stylization. Baseline styles are shown on the diagonal highlighted in red. Off-diagonal styles are created by mismatching shape and texture (material). The graphs show a decrease in the appeal of a character as they become more realistic. Borrowed from Zell et al. (2015).

Virtual Agents. After rendering the appearances of Anna and Beth, the characters were programmed to perform social behaviours when they interacted with the participants. In this way, Anna and Beth go from being just virtual 'characters' to becoming virtual 'agents' that can participate in conversations in a natural way. More specifically, both the interactive and the non-interactive agents were programmed to generate, in real time, both verbal and non-verbal social signals to effectively be able to have a realistic conversation with the participants. The non-verbal signals included eye gaze, blinks, lip-syncing, and head nodding. Eye-gaze and blinks were animated

as naturally as possible to enhance the realism of the agents. The lip movement were synced-up with the pre-recorded speech segments from the two voice actors.

For the verbal behaviour of the virtual agents, we scripted the monologue of each card, which were pre-recorded from two female voice actors with native British accents. We also pre-recorded several speech segments of potential dialogue options, which was triggered manually by the experimenter in Unity. This paradigm is commonly described as a 'Wizard of Oz' method because the verbal behaviour appears to come from the virtual agent but is actually monitored and driven by a human 'wizard' (Thórisson, 1994). Both agents were also programmed to move their lips according to the amplitude of the pre-recorded speech, so that it looked as if they were speaking. Lastly, the speech audio was presented from speakers behind the projector screen, so that the agents' voices came from their virtual location.

To drive the head nodding of the interactive agent, we generated head nods with appropriate timings under two different behaviour rules. For these to work, the interactive agent first must be able to detect the behaviour of the participant. Here we use the positional head-trackers or audio channels of the HMD. The HMD utilizes a sensor mounted on the participant's head, which allows for accurate head motion detection. The participant's speech was detected based on the audio from the microphone input of the HMD. Unity was used to program specific responses conditional on the participants' behaviour, as follows:

(1) If the interactive agent is listening or the participant is speaking, the agent periodically (approx. 15-20% of the total speaking time) performs high frequency fast nodding (4.5-6.5 Hz) to simulate a natural backchannel as visual feedback for the participants. Such a fast nod movement consists of 2-3 fast nodding peaks, with a where the velocity of the pitch rotation was in the interval [$r \cdot 16.33$, $r \cdot 40.84$], where r is

a radius, given by the distance from the center of the virtual agent's neck joint to the chin. This rule is based on our previous findings that participants tend to fast nod this way when listening (Hale et al., 2020; Chapter 2).

(2) if the participant does a "big" nod (which is any head movement where the velocity of the pitch rotation was in the interval [$r \cdot 1.25$, $r \cdot 6.91$]), then the interactive agent will nod 600 ms after the detected end of the head nod with the same value of pitch rotation velocity, to simulate natural mimicry (Hale et al., 2020). Their appearances were then used as either the interactive or non-interactive model, counterbalanced across participants.

To drive the head nodding behaviour of the non-interactive agent, we used pre-recorded movement data from a pilot participant which provides natural head motion that is not contingent on the participant's speech or actions in any way. Virtual agents with slightly different appearances were then used as either the interactive or non-interactive character, counterbalanced across participants.

4.4.3 Procedure

Participants arrived and were shown the equipment and informed of the procedures, after which they signed the informed consent. They were seated at a desk in front of a projector screen (Figure 4-1A, B) and given more detailed verbal instructions from the experimenter, after which they were fitted with the VR-headset. Inside the virtual room, the participants could see a virtual desk and a screen with written instructions in front of them. The participants were given a moment to become accustomed to the environment while watching a video example of the task. After watching the video, they were instructed they were going to meet two virtual people named Anna and

Beth and have a conversation with them. When ready, they were asked to press a button on one of the controllers to begin the task. Anna and Beth were both programmed to greet the participant (e.g., “Hi, I’m Anna, nice to meet you”).

Virtual Conversation Task. The participants, together with the virtual agents, took turns to read cards with facts on them relating to different American states (See Section 3.4.3), presented on virtual tablets. Whenever the agents spoke, they alternated between looking down at the card they were reading and looking up at the participant to meet their gaze. The difference in behaviour between the two agents was that the interactive agent was driven by the behaviour rules mentioned above, whereas the non-interactive agent acted as a control with pre-recorded head nodding from a pilot participant. The participants completed the task with both the interactive and the non-interactive agent one after the other, in a within-subjects design. Which virtual agent acted as the interactive one depended on which group each participant belonged to, which was based on what set of cards they were reading from (See Table 3-1). After completing 16 trials in two blocks (Anna and Beth), they were moved to a separate table to complete the memory test.

Likeability Questionnaire. After the participants had completed the memory test, they were given a questionnaire about their experience of Anna and Beth, containing 46 items. They indicated their agreement with statements on a Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). The first 4 statements assessed how real or immersive the participants felt the virtual environment was (“I felt that I was in the presence of another person in the virtual room”). The next 36 statements assessed the likeability (“I find Anna confident”) and realism (“Beth was aware of my existence”) of the virtual agents. The last 6 statements assessed direct likeability between the two virtual agents (“I felt that Anna was maintaining eye-contact with me

more than Beth"). At the end of the experiment, after finishing the questionnaire, participants were debriefed about their experience to determine whether they had noticed if either Anna or Beth had been mimicking them or otherwise guessed the purpose of the experiment ("Did you notice anything unusual about your interaction with Anna and Beth?", "Did you notice if Anna or Beth mimicked your behaviour in any way throughout the experiment?"). In the following section I will present how we analysed the data based on the three hypotheses that were stated.

4.5 Data Analysis

The first aim of this study, based on preliminary correlation measures, was to further investigate a causal link between interactive engagement with a virtual agent and memory performance. The second aim of this study was to test if head nodding can be used to quantify how much we like a virtual agent and follow this up by examining if there is a link between liking and learning.

4.5.1 Analysis of Memory Performance

We performed a basic analysis of general memory recall, similar to the 'real' interactions in Chapter 3. The memory decisions were categorized into one of four response categories based on the "Old/New" status of the fact and the given response: hit, miss, false alarm, and correct rejection responses on the memory test (See Section 3.5.1). The performance was calculated by considering hits and correct rejections and presented as a % of correct recall (hits + correct rejections divided by 2) across speaking vs. listening, and interactive vs. non-interactive engagement.

4.5.2 Multilevel Binary Logistic Regression Model Analysis

After the basic analysis on memory performance, we aimed to test if interactive engagement promotes the encoding of new information ($\mathbf{H_1}$) by testing whether the conversation with a virtual agent with interactive head nodding behaviour could lead to increased general recall on the memory test compared to a non-interactive virtual agent. We calculated the dichotomous outcome measure of memory on a trial-by-

trial basis whether the participants got the correct answer for each fact (0 = not correct and 1 = correct). The final sample was used to create a logistic multilevel model that consisted of 32 participants sorted into speaking and listening roles. There were 3 ‘real’ facts for each of the 16 trials sorted by Set (Table 3-1), which together with the ‘new’ facts resulted in 96 facts per participant, and a total of 3072 datapoints. The data used in the model [Memory ~ Agent + Role + (1|Participant)] included the memory recall (1 = correct, 0 = incorrect), and the dummy codes for Agent (1 = interactive, 0 = non-interactive), Role (1 = speaking, 0 = listening), and Participants (1-32).

We used a two-level model with Agent (interactive vs. non-interactive) and Role (speaking vs. listening) as predictors (Level 1) nested within participants (Level 2). We had no interest in analysing the grouping variable of ‘participant’ as a random effect but needed to factor this out for individual variation in the model parameters. We fitted this model to the data by performing a multilevel binary logistic regression. The aim of this analysis is to estimate the odds that the participants remember a fact correctly (1) or incorrectly (0) while taking the dependency of data into account for each level. Using this approach enable us to take both the between-stimuli and between-participant variations into account (Judd, Westfall, & Kenny, 2012). In other words, we could for example observe the probability or likelihood that a participant remembers a fact correctly when engaged interactively *and* while reading (rather than being told) a fact. This model is well suited for testing the relationship between our dichotomous outcome variable and our categorical predictors. The logistic model was evaluated using standard *F*-tests, and likelihood (odds ratio) tests for estimating the significance of each predictor. All tests were based on an alpha level of 0.05.

4.5.3 Likeability Questionnaire Ratings

After scoring the likeability ratings, we aimed to examine if interactive engagement promotes liking (**H_2**) by testing whether the conversation with the virtual agent with interactive head nodding behaviour could lead to increased feelings of rapport compared to the non-interactive agent. We calculated the outcome measure of the overall ratings for both Anna and Beth and matched that to whatever agent had the corresponding interactive or non-interactive behaviour for that session. We then compared the mean ratings between the agents with a paired-samples t-test.

In addition, based on a recent study (Aburumman, Gillies, Ward & Hamilton, 2022), we performed the same analysis for three specifically chosen questions which had shown positive results. These were the questions: Q1) "*I believe Anna/Beth was maintaining eye-contact with me*"; Q2) "*I felt that Anna/Beth's head movement was natural*"; and Q3) "*I believe Anna/Beth showed attention to what I was saying*". In this analysis, we similarly wanted to look at the differences between the interactive and non-interactive versions of these questions and compare the two means.

Analysing the likeability ratings further we also aimed to examine if liking is linked to learning (**H_3**) by testing a link between feelings of rapport and increased general recall on the memory test. We correlated the sum of the overall ratings for the interactive and non-interactive questions with the general memory recall which was divided into interactive and non-interactive recall rates across both speaking and listening trials. A Bonferroni correction was carried out for multiple comparisons.

4.6 Results

4.6.1 Mean Memory Performance

We carried out a basic analysis of the participants' general memory recall, presented as a percentage of correct recall across both speaking and listening trials. The mean results when speaking shows 92% correct recall for interactive, and 94% for non-interactive engagement. When listening, the results show 84% for interactive, and 82% for non-interactive engagement. The results are presented in Figure 4-4 below.

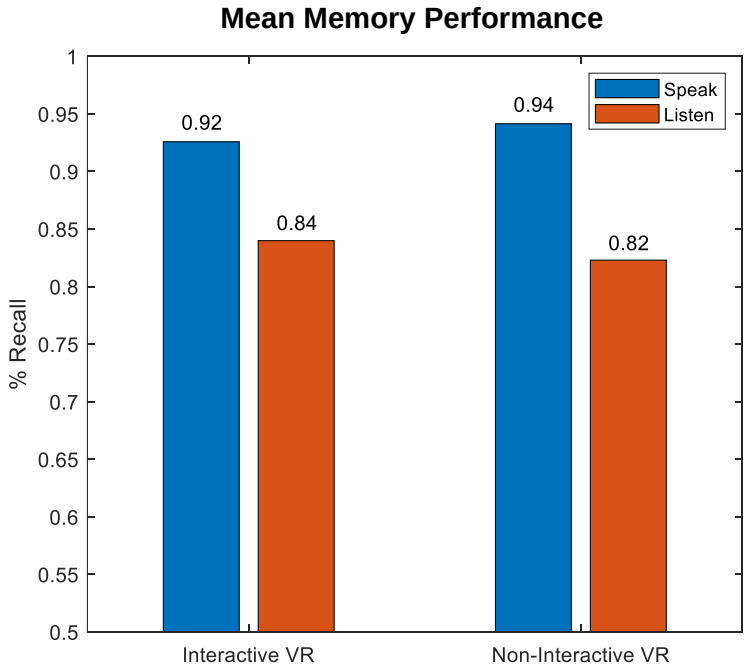


Figure 4-4. Mean Memory Performance. The mean general recall rates in percentages when speaking and listening across both interactive and non-interactive engagement in virtual reality.

4.6.2 The Effect of Interactive Engagement on Memory

We performed a multilevel binary logistic regression analysis to test whether having a conversation with a virtual agent with interactive head nodding behaviour could lead to increased general recall on the memory test compared to a non-interactive virtual agent. The results are presented in Table 4-1 below.

Table 4-1

Multilevel binary logistic regression results

Parameter Estimation	β	*SE* β	F	*df*	*p*	Odds Ratio
Overall Model Evaluation	–	–	21.369	2	.0001	–
Agent (1 = Interactive)	0.007	0.1360	0.002	1	.961	1.007
Role (1 = Speaking)	1.075	0.1661	*41.882*	1	.0001	2.929

Note. Memory ~ Agent + Role + (1|Participant). The random effect of participant was factored out in the regression (level 2) for individual variation in the model parameters.

The results of the overall model evaluation (p<.0001) shows that together the fixed effects are significant predictors of memory recall. The results for the individual predictors show no significant effect (p=.961) of agent-interactivity. Trials performed with the agent with interactive head nodding, are equally likely (odds ratio=1.007) to fall into the category of correctly being able to recall a fact as those interacting with the non-interactive agent. However, a significant effect (p<.0001) of role could be observed, and trials where participants read a fact and share the information with the agent are three times more likely (odds ratio=2.929) to fall into the category of correctly being able to recall a fact than trials where the participant is listening to one of the agents reading a fact. In other words, given the same interactive nature of the virtual agents, listening instead of speaking to a virtual agent during a conversation makes it less likely that you will remember something from the interaction.

4.6.3 The Effect of Interactive Engagement on Liking

We performed a paired-samples t-test of the overall likeability ratings between the interactive and non-interactive agent questions on the questionnaire in order to test if having a conversation with the virtual agent with interactive head nodding behaviour

can lead to increased feelings of rapport over that of the non-interactive agent. The results for the overall ratings show that there was no significant increase in feelings of rapport when interacting with the interactive virtual agent (*M*=87, *SD*=16.9) compared to the non-interactive agent (*M*=82.3, *SD*=18.1), *t*(31) = 1.12, *p=.27.* Moreover, based on previous reports (Aburumman et al., 2022), we also performed a paired-samples t-test for three specific questions (See Section 4.5.3 for details), which likewise were not significant, *t*(31) = 1.08, *p*=.29.

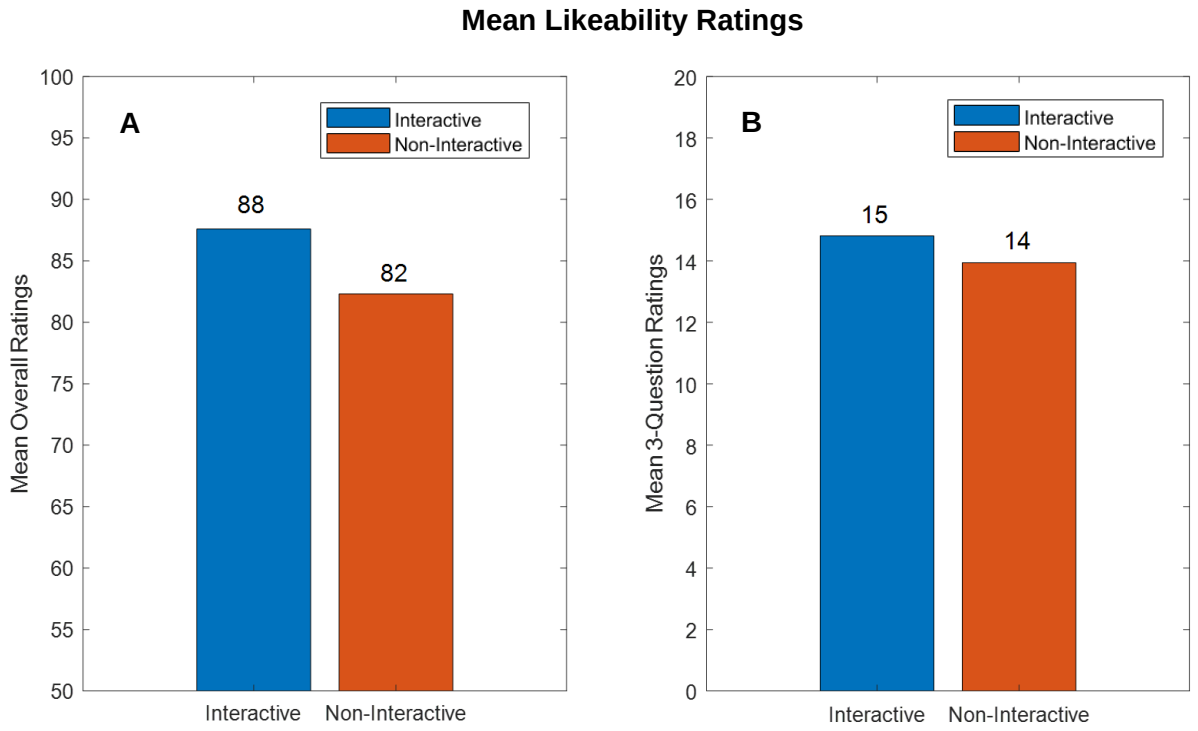
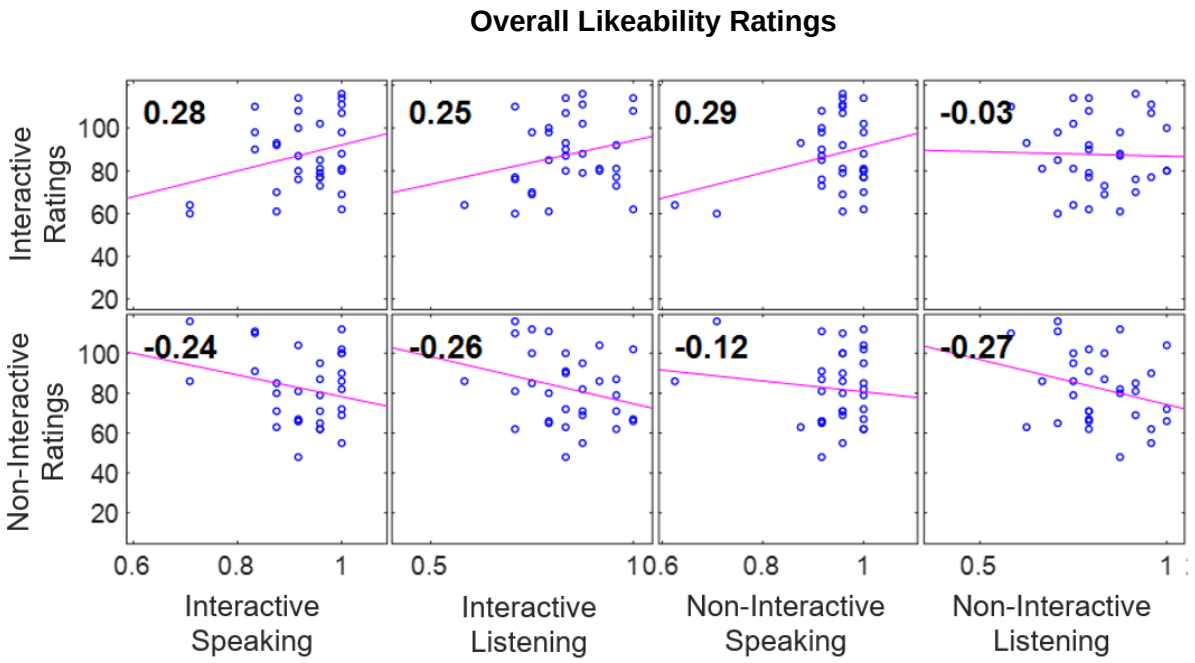


Figure 4-5. Mean Likeability Ratings. **(A)** shows the mean ratings across interactive (Score: 88) and non-interactive (Score: 82) conditions. **(B)** shows the mean ratings for the three specific questions (See Section 4.5.3) across interactive (Score: 15) and non-interactive (Score: 14) conditions.

4.6.4 The Effect of Liking on Learning

We further correlated the sum of the overall ratings for the interactive and non-interactive questions with the general memory recall, divided into interactive and non-interactive recall percentage across both speaking and listening trials to test if

there is a link between feelings of rapport and increased general recall. The four memory conditions include Interactive Speaking (IS), Interactive Listening (IL), Non-Interactive Speaking (NS), and Non-Interactive Listening (NL).



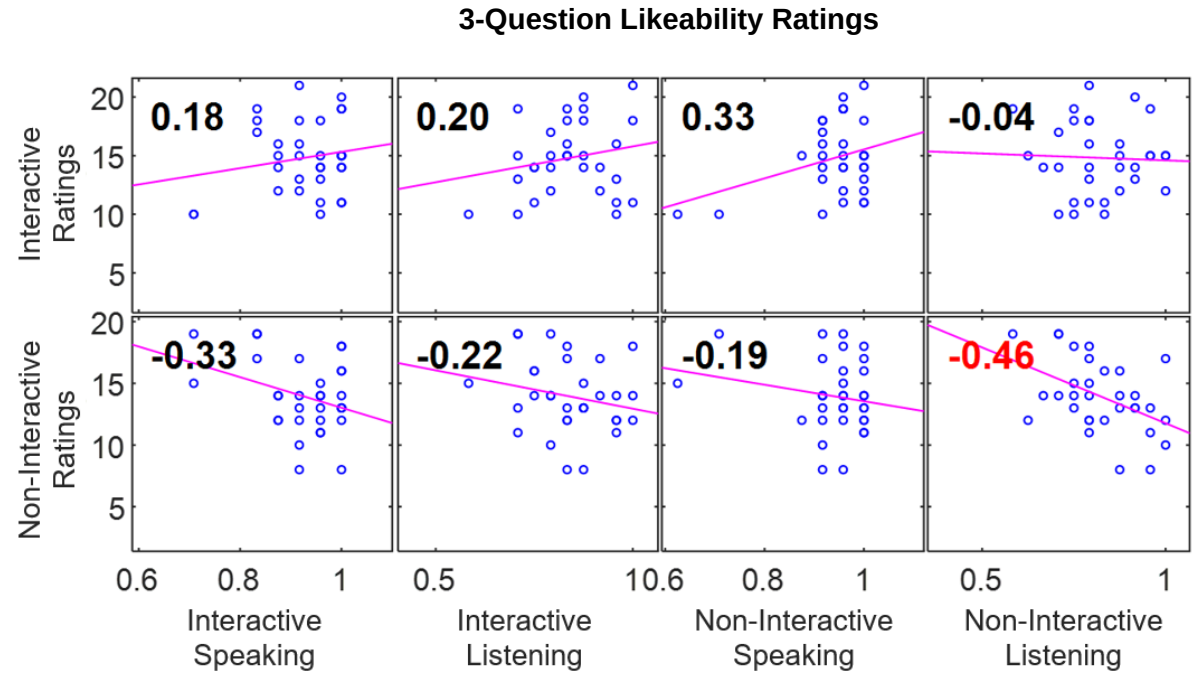


Figure 4-6. Correlation matrices. Red correlation scores indicate if the correlation is significantly (p<0.05) different from zero. The x-axis values represent the general memory recall in percentage. The y-axis values represent the sum of the question ratings. Each data point represents a participant.

For the overall ratings, the results presented in Figure 4-6 show that there is no reliable positive link between the interactive questions paired with any of the memory conditions (IS, p=.12; IL, p=.18; NS, p=.11; NL, p=.88). There was no reliable positive relationship between the non-interactive questions paired with any of the memory conditions (IS, p=.19; IL, p=.14; NS, p=.51; NL, p=.13).

For the 3-question ratings, the results presented in Figure 4-6 show that there is no reliable positive link between the interactive questions on the questionnaires paired with any of the memory conditions (IS, p=.33; IL, p=.27; NS, p=.07; NL, p=.82). Similarly, there was no reliable positive relationship between the non-interactive questions paired with any of the memory conditions (IS, p=.06; IL, p=.24; NS, p=.31), except for a negative correlation, r=-0.46, p=.008, in the non-interactive listening (NL) condition. However, this negative correlation did not survive a bonferroni correction for multiple comparisons.

4.7 Discussion

In Chapter 3, we found ambiguous results of whether we were more likely to engage in fast nodding when we recall more information from a conversation and use this to predict memory performance. Some analysis patterns gave us indications that some head nodding frequencies might relate to memory during unstructured conversations. In this chapter we aimed to continue this investigation and test if there is a causal link between interactive engagement and memory performance.

To test our hypotheses, we set up two virtual agents with eye-gaze, blinks, and lip-syncing non-verbal actions together with a WoZ speech system. One of these agents was programmed to show interactive natural head nodding behaviours based on our earlier findings. As a control condition, the head movements of the other virtual agent were non-interactive, only driven by pre-recorded motion. We then compared how the participants engaged in a conversation with these virtual agents and measured their performance with a post-hoc memory test.

The second goal of this study was to observe if head nodding could be used to change how much we like a virtual agent. For example, can interacting with interactive virtual agents, driven by natural head nodding behaviour, enhance feelings of rapport compared to non-interactive virtual agents? Following from this, we also looked at whether the participants learn better from virtual agents that they like, by examining a link between feelings of rapport and memory performance.

Next, we will explore the possible interpretations of the three hypotheses. We will then proceed to discuss the implications and limitations of the study, what this could mean for hypothesis testing using interactive virtual agent technology, and how future directions can extend the contribution of this approach.

4.7.1 Does Interactive Engagement Promote Learning?

With our first hypothesis (**H_1**), we predicted that the conversation with the virtual agent with interactive head nodding behaviour should lead to increased general recall on the memory test compared to the non-interactive virtual agent. The results show that there is no significant effect of agent-interactivity between the interactive and non-interactive agents, which fails to support our first hypothesis (**H_1**). We used the same task and measures from Chapter 3, with the difference that it was implemented in VR. Hence, the design is prone to the same kind of limitations. For example, there is the possibility of ceiling effects in the recall measure, with a high proportion of the participants remembering all facts correctly. In addition, the virtual agents and the participants are reading facts from a virtual tablet instead of a physical card, which results in the same limitations of having to interchangeably switch gaze direction between the tablet and one's virtual conversation partner.

Moreover, a conversation with a virtual agent comes with its own limitations as it neglects a lot of the parameters that are present in a naturalistic conversation. For example, we focus only on manipulating head nodding behaviours, while also controlling other highly important nonverbal behaviours like eye-blinking, gaze, and facial expressions in a simplistic fashion compared to real behaviour. This leaves out more detailed facial expressions, language parameters, posture, etc, which remain challenges and not very well supported with VR technology. Instead of trying to implement everything, we opted to delimit and focus on making the interaction between the virtual agent and the participant work as good as we could by designing all the non-essential features of the virtual agents, including the order of interacting, their visual appearance, and the voice lines, counterbalanced across participants so that we could investigate just the impact of the interactivity.

As we hypothesized in Chapter 3, it is likely to be the participant who is performing the fast nodding that have learned something from the interaction. However, it is very difficult to manipulate the head nodding of the participants directly. Thus, it is important to note that the first hypothesis in this study is an indirect manipulation of head nodding behaviour because we manipulate the virtual agent and its visual signals, rather than the participants' head nodding behaviour directly. This design decision impacts whether we can make a claim about interactive engagement promoting (i.e., causal) the encoding of new information (i.e., memory). However, by doing an indirect manipulation of the virtual agent's head nodding behaviour, we change the dynamic of the conversation, which in turn changes how the participants respond to the virtual agent with both fast and slow nods. This is similar to the pseudo controls we performed in Chapter 2 when we matched data from different trials within the same pairs of participants. Using pseudo-matching similarly changed the dynamic of the conversation, and consequently how the participants behaved, while still maintaining the participants individual movement characteristics.

This also brings us to the point about how we implement the behaviour rules. Since we are both looking at the role of head nodding as visual feedback and how it impacts the interactive engagement between the participant and the virtual agent, future studies could benefit from collecting the head nodding data from the HMD. For example, because the interactive virtual agent must detect the behaviour of the participant by using the HMD input for head motion and speech, future studies could use that data to analyse how it relates to memory performance, or some other measurable cognitive outcome.

Specific responses of the interactive virtual agent, based on behaviour rules, were programmed to be conditional on the participant's behaviour. The first rule was that

the agent performed high frequency fast nodding to simulate a backchannel as visual feedback for the participants. Each fast nodding event consisted of 2-3 fast nodding peaks, which is consistent with the theoretical definition of a nod from the background literature (Poggi et al., 2010) where it typically consists of multiple up and down movements, including an initial vertical movement, after a slight tilt up it, bending downward, and then going back to its starting point. The virtual agent performed these nodding events approximately 15-20% of the time the participant was speaking or when the agent was listening. This rule was based on our previous findings that participants tend to fast nod this way when listening (Hale et al., 2020; Chapter 2). The second rule that the virtual agents were programmed to perform was behavioural mimicry (Chartrand & Bargh, 1999). If the participants were listening and performed a nod within a certain range to count as a “big” nod, then the interactive agent would nod 600 ms after the detected end of the head nod. This 600 ms time lag was first revealed by Hale et al. (2020) and is used as a good rule to simulate natural mimicry behaviour.

A point worth mentioned is that the pre-recorded movements that were used for the non-interactive agent are captured from a single pilot participant. While we were careful to select these movements to be small and slow to not appear odd, having a larger sample for these movements available would be useful to ascertain the non-interactive agent as more representative of the population.

The WoZ design that we chose for the facial expressions and during the dialogue interaction has its limitations. However, most other studies that use similar virtual reality paradigms also use WoZ and a standard in the field (Jain, Pecule, Matsuyama, & Cassell, 2018). What it provides is a non-structured (i.e., free flowing) conversation in a VR setting. Using human controlled virtual characters (i.e., avatars)

or WoZ-controlled virtual agents enable us to better represent how people interact and has the potential to capture different conversational dynamics. In our experiment, both the virtual agents' voices were recorded by native English-speaking actresses, and we prepared a lot of phrases related to different conversational topics, including some stock phrases, and accompanying facial expressions, so that the wizard (i.e., experimenter), who was blind to the conditions, could provide a more or less natural conversation with the participant after some training.

However, even if all this preparation amounts to better experimental control, it still suffers from a lack of ecological validity, and the fact that participants will not behave in the same way when conversing with a virtual agent compared to a real person. That said, there is a growing body of research aimed at analysing social signals with the goal of building applications and interfaces, based on models of human behaviour, that limits the trade-off between experimental control and realism in VR (Burgoon, Magnenat-Thalmann, Pantic, & Vinciarelli, 2017).

While our results fail to support our first hypothesis, we do however observe a significant effect of 'role', showing that participants that are speaking, or reading, a fact are up to three times more likely (i.e., odds ratio) to remember the information. Thus, listening during a conversation makes it less likely that you will remember, or learn something from the interaction compared to when you are speaking. This gives us a good indication that these results replicate the self-reference effect (SRE) from the study in Chapter 3. However, this effect can still come from being more attentive, by having access to visual memory, which allows for more opportunities to make associations to new information. In Chapter 5 we will examine this further. Next, we will discuss if head nodding can be used to change how much we like a virtual agent, or if we learn better from virtual agents that we like.

4.7.2 Does Interactive Engagement Promote Liking?

With our second hypothesis (**H_2**), we predicted that having a conversation with the virtual agent with interactive head nodding behaviour should lead to increased feelings of rapport over the non-interactive virtual agent. The results from the questionnaire ratings show that there is no significant increase in feelings of rapport when interacting with the interactive virtual agent compared to the non-interactive agent, which fails to support our second hypothesis (**H_2**) that interactive engagement promotes liking during conversations.

This result is consistent with previous research using human and virtual mimickers, which have reported mixed results on the idea that mimicry of head nods can act as a social glue to increase affiliation and liking between people. For example, Bailenson and Yee (2005) demonstrated positive effects on participants' impression of being mimicked in virtual reality, whereas Verberne et al. (2013) used the same mimicry algorithm and found inconsistent results across a range of measures. It is also worth pointing out that ratings of liking show inconsistent effects of mimicry in traditional research settings where human confederates were trained to mimic participants (Hale & Hamilton, 2016a). In the study by Hale and Hamilton (2016b) they similarly found no significant effects of virtual mimicry on rapport, which cast doubt over a strong version of the 'social glue hypothesis', in which all types of mimicry have positive social effects. It could be that the positive effects of nodding are more subtle, but despite positive results from studies using interactive agents to support the social glue theory, we continue to find no significant effects in support for the idea that interactive engagement promotes liking towards a virtual agent.

We have extended previous WoZ paradigms using interactive virtual agents by implementing an unstructured and interactive conversation in which the agents follow

behavioural rules extracted from real world head nodding behaviour. Using this novel design, this study demonstrates that the 'social glue hypothesis' is difficult to replicate in virtual reality. Participants show no significant effects of liking when they are engaged in interactive head nodding behaviour with a virtual agent.

4.7.3 Do We Learn Better from Virtual Agents We Like?

With our third hypothesis (**H_3**), we predicted that increased feelings of rapport should be linked to increased general recall on the memory test. The results from the overall questionnaire ratings, as well as the three specific questions, show that there is no reliable positive link between the interactive and the non-interactive questions on the questionnaires paired with any of the memory conditions, which fails to support our third hypothesis (**H_3**) that liking is linked to learning. However, looking at Figure 4-4 we can observe an increasing trend on the interactive questions and a decreasing trend on the non-interactive questions. We like to emphasize that the data we have on this is based on a relatively small sample (i.e., 32 participants) and more conclusive research with larger sample sizes would be needed to make any recommendations about using measures of liking to predict learning.

4.8 Limitations and Future Directions

It could be argued that these findings reflect a failure of virtual interactions to achieve the same effects as real social interactions, which is something we will explore further in the next chapter of this thesis. We acknowledge that current virtual reality paradigms are far from perfect as participants in this study were aware they were interacting with a virtual agent and not a real person. However, studies have successfully replicated psychological constructs when interacting with virtual agents before, ranging from joint attention (Schilbach et al., 2009), proximity effects (McCall & Singer, 2015), and audience effects (Slater, Pertaub, Barker, & Clark, 2006) demonstrating that this approach can generate socially realistic interactions.

One issue with this study is the measures of likeability, which asks the participants to retrospectively rate how much they like a virtual agent on a likert scale. This is a well-known issue with using questionnaires since there are known dissociations between what people say about their own (and other's) behaviour and the factors influencing them (Nisbett & Wilson, 1977; Haidt, 2001). This could be improved on by including behavioural measures, like proximity effects (McCall & Singer, 2015).

A second issue concerning the verbal component of the interactions is that even though the WoZ method provides a degree of experimental control and some useful insights into how the participants interact, it is not without its problems. For example, because it is difficult to provide consistent responses across sessions, this method requires significant training so that the experimenter, or 'wizard', can respond in a way that is credible. We used several pre-recorded speech segments of dialogue options and behavioural instructions to minimize this effect. Since the experimenter played the role of the wizard, there is an added risk of the experimenter-expectancy effect, where the researcher's cognitive bias could have unconsciously influenced

the participant's behaviours. Playing the wizard is also exhausting, meaning that the researcher's reactions may change across sessions. Consequently, having more than one 'wizard' could have improved the experimental design of this study.

In this study we have demonstrated the benefits of *artificially generating* interpersonal coordination by generating behaviour rules in interactive virtual agents based on behavioural data from real social interactions to test our hypotheses. We currently lack detailed knowledge on many of these behavioural parameters during naturalistic interactions but using interactive virtual agents as models to for hypothesis testing provides us with a test of simple behaviours, because it allows us to specifically manipulate one type of body movement (head nods) while keeping all other body movements the same (eye-gaze, blinks, lip motion). However, real-world social interaction is complex and dynamic. Non-verbal behaviour is highly changeable and can depend on the topic of the conversation, the surrounding context, and the person with whom we interact. Current interactive virtual agents are not yet able to display this kind of range.

One way to improve and complement the lack of such advances is using virtual reality together with machine learning. Machine learning is a computing method associated with cognitive simulation, or artificial intelligence (Michalski, Carbonell, & Mitchell, 2013). It involves programming algorithms that can learn from and make predictions from datasets without having to manually code everything in detail. It allows researchers to automatically detect occurrences and model interactive patterns of data. In the context of modelling social interactions, this method can generate probabilistic models that predict behaviour over time of one person based on the other. For example, if the models derived by machine learning techniques are used to drive the behaviour of an interactive virtual agent, then the virtual agent will

be able to respond appropriately to the participant's behaviour in real-time. In this way, machine learning may be able to generate virtual agents that closely approximate real behaviour without the need to manually extract and code individual parameters (Gillies, 2009). This kind of data-driven approach to generating social interaction, in combination with the development of well-specified theoretical frameworks, will continue to require strong interdisciplinary collaboration.

4.9 Conclusions

Some analysis patterns from Chapter 3 gave us indications that some head nodding might be related to memory during unstructured conversations, and in this chapter, we aimed to continue this investigation to test if there is a causal link between interactive engagement and memory. We found no significant effect of agent-interactivity between interactive and non-interactive agents, which failed to support our hypothesis that interactive engagement promotes the encoding of new information. However, we did find that people that are speaking are up to three times more likely (i.e., odds ratio) to remember information during a conversation compared to those who are listening, effectively replicating the SRE during virtual conversations. However, the effect seems less prominent than during real interactions, which is something we want to explore further in the next chapter.

The second goal of this study was to observe if head nodding can be used to change how much we like a virtual agent. We found no significant differences in feelings of rapport between the interactive and non-interactive virtual agents, which failed to support our hypothesis that interactive engagement promotes liking during conversations. This result is consistent with previous research reporting that mimicry of head nods does not always act as a social glue to increase affiliation and liking.

Our third goal with this study was to observe if feelings of rapport is linked to increased general recall of the memory test. We found no reliable correlations between ratings of liking and memory performance, which failed to support our hypothesis that liking is linked to learning.

We have tried to demonstrate with this study the benefit of using virtual reality to *artificially generate* behaviour rules in interactive virtual agents based on real-world behavioural data, as models to test our hypotheses about memory and liking. The external validity and limitations of these virtual models is something we want to investigate further in the next chapter.

Chapter 5. Memory of Information Across Real, Virtual, and Video Interactivity: A Cross-Experimental Study

5.1 Abstract

In this study, we collected additional data where people acquire information in the context of video recordings, aimed at investigating the way that we learn new information depending on the level of interactivity from different mediums (i.e., real, virtual, and video). Across the three studies investigating learning in this thesis, Chapter 3 focused on 'real' interactions with a high level of interactivity; Chapter 4 on 'virtual' interactions with a moderate level of interactivity; and this chapter (5) on video recordings with no interactivity. In this cross-experimental study, we show that the level of interactivity supports memory and learning during conversations. First, the results confirm the existence of a 'video deficit' in memory compared to real interactions when speaking and show that real social interaction is important for learning and memory. Secondly, we also found a 'video deficit' in memory compared to virtual interactions, which shows that virtual agents can help enhance learning over video teaching. Lastly, this is supported by a surprising finding that we remember more when engaged with interactive virtual agents compared to real interactions when listening. In this study, we demonstrate the benefits of exploring social interaction in different online settings, which may prove to be valuable when creating tools for online education amid a changing landscape for teaching.

5.2 Introduction

In this new era of online social interaction following the recent coronavirus (COVID-19) pandemic, it is important to understand what makes an interaction work and how we can best implement the benefits of real social interactions in an online setting. So far, this thesis has examined what people remember from a conversation in a real-world face-to-face setting and in a virtual reality setting. Here, we collected additional data where people acquire information in the context of video recordings, aimed to explore the differences in memory across three experiments with varying levels of interactivity (i.e., real, virtual, and video). This will allow us to investigate the way that we learn new information depending on the level of interactivity from different mediums.

Most of the research on learning and memory examines either asocial learning (i.e., the student is alone) or observational learning (i.e., the student watches another individual but does not interact). Observational learning involves acquisition of information through passive exposure to the material (e.g., learning from a pre-recorded video) (Laland & Rendell, 2019). In contrast, interaction-based learning requires mutual feedback between student and teacher (Shamay-Tsoory, 2021). In other words, in observational learning, we learn *from* others, while in interaction-based learning we learn *with* others. Thus, these forms of social learning mainly differ based on the level of interactivity.

It has been shown that there are differences between learning from real-world interactions compared to video recordings of real people. This suggests that children learn better from real social interactions compared to when they watch the same person on video (Kuhl et al., 2003). In their study, nine and ten-month-old infants were assessed on their ability to learn a foreign language (Mandarin Chinese). Over

a course of 4 weeks, the children attended 12 sessions in which they were exposed to both Mandarin and English speakers. While some infants had direct or “real” interactions with the speakers, the two remaining groups where either exposed to the speaker through audio-visual recordings or could only hear the speaker through audio recordings alone. Following testing to determine how much the infants had learned, the authors concluded that phonetic learning could only be seen in the group that were directly exposed to the foreign language speakers. In fact, the learning of the other two groups was noted as similar to the English control group. This mismatch in learning has been defined in the literature as a “video deficit” (Anderson & Pempek, 2005). The basis of this concept is supported by studies with evidence that real social interaction is important for learning (Krcmar, Grela, & Lin, 2007; Roseberry, Hirsh-Pasek, Parish-Morris, & Golinkoff, 2009).

Although the evidence of a video deficit has been replicated in certain bird species (Baptista & Petrinovich, 1986; Eales, 1989), it is still unclear whether this effect is also seen in human adults, and why this should be true. In a study by Schreiber, Fukuta and Gordon (2010), they compare learning from students who attended a live lecture and those that watched a recording of the same lecture. A sample of 100 students were randomly assigned to two groups, and then performed the experiment in a within-subject design, switching between the two conditions. Following this, the students were given a test to see how much they learned from each version of the lecture. Unlike in young children where a video deficit was noted, no significant difference in learning was found between when the participants attended the live lecture and when they watched the video recordings. Similar results have also been observed in other studies with students either attending live lectures or watching digital lectures (Davis et al., 2007; Solomon, Ferenchick, Laird-Fick, & Kavanaugh,

2004; Vaccani, Javidnia, & Humphrey-Murto, 2016). In the study by Solomon et al. (2004), they found that despite the mean scores being equal between the two conditions, the variation among the scores of students who watch the digital lectures was almost twice as large as for students that attended the live lectures. It is possible that this effect could stem from some methodological inconsistencies in their study, like small and uneven sample sizes within and between each group; the fact that they used a non-validated local test to measure learning; technical problems in the presentation of the digital lectures; not controlling for exposure time (e.g., recorded material could be replayed multiple times, whereas the live session was only played once). These issues would make a direct comparison between learning from 'real' interactions versus from 'video' recordings difficult and in need of further investigation.

In a recent study by De Felice, Vigliocco, and Hamilton (2021), they investigated how adults learn information about unknown objects from live versus recorded lectures and found that interaction-based learning is more effective than observational learning. They also found that when the teacher's face and hands were fully visible, playing an active role in the interaction, improve learning over yoked observations of the same sessions. This shows that the presence of non-verbal signals affected learning new information differently depending on whether teaching was interactive or not.

Social interaction is cognitively demanding (Kourtis, Jacob, Sebanz, Sperber, & Knoblich, 2020) and could impact learning in different ways. It might impair learning by increasing cognitive load (Hertel, Brozovich, Joormann, & Gotlib, 2008), but it might also increase learning, as seen in children (Kuhl et al., 2003). As such, the relationship between the level of interactivity in the different tasks in this study and

learning could go in either direction. For example, a high level of interactivity (i.e., 'real' interactions) could increase cognitive load, and/or distract learners (Kajopoulos, Cheng, Kise, Müller, & Wykowska, 2021). In addition, non-verbal signals, such as eye-gaze (Morotta, Lupiáñez, Martella, & Casagrande, 2012), and gestures from a teacher (Wakefield, Novack, Congdon, Franconeri, & Goldin-Meadow, 2018) could benefit learning by facilitating interpersonal coordination between student and teacher. The importance of interactivity in social learning that de Felice et al. (2021) demonstrated not only raises the question of which aspects of the interaction contributed to learning, a question which we have examined in Chapter 3, but also what other types of social interactions provide a learning context that depends on the way in which information is delivered.

In addition to 'real' interactions, which elicit interaction-based learning, and 'video' recordings, which elicit observational learning, we can consider a third level of interactivity between these two types of learning. This level is 'virtual' interactions, which elicit interaction-based learning, but similar to videos are not real face-to-face interactions. According to popular wisdom, humans never relate to a computer or people on various forms of media in the same way they relate to another human being. However, research by Reeves and Nass (1996) showed that real communication transfers to human-media communication. For example, their study shows that people are polite to computers; that large faces on a screen can invade a person's personal space, and that motion on a screen affects physical responses in the same way that real-world motion does. They theorised that the human brain has not evolved fast enough to assimilate 20th century technology and believed that the mechanisms of social cognition might even transfer to "non-humans". Later studies using non-human, or computer animated characters, confirmed their external validity

with that of video recordings of real people (Bente, Krämer, Petersen, & Ruiter, 2001). In this study, the researchers recorded video of dyadic interactions and compared them with animated characters based on movement transcripts of the same interactions in a between-subject design. The participants socio-emotional impressions were assessed, and the data showed remarkable similarities between both conditions, indicating that most of the relevant social information available to them in the video recordings was also conveyed by the animated characters. However, the generalizability of virtual reality to the real world is something that has not been explored in much detail. Hence, we decided to investigate this in this cross-experimental study.

In the next section we will present the current study. In this study we aim to improve upon previous studies in the field both in the methods in which the video recordings are presented, as well as introduce data from three tasks together and compare three different levels of interactivity and their effects on memory.

5.3 The Present Study

Previous studies have not systematically examined the level of interactivity as a contributing factor in learning. Taking these clues from previous research we can build a hypothesis which states that there should be differences in the way that we learn new information depending on the level of interactivity from different mediums. To test this, we collected new data from video recordings to compare memory performance on a post-hoc memory test across three experiments in this thesis that have utilized different levels of interactivity (i.e., real, virtual, and video).

The ‘real’ interactions are presented in Chapter 3 and constitute interaction-based learning with a high level of interactivity, where the participants interact with another human being in real time. The task in that study contains both a monologue and dialogue part of the conversation. The ‘virtual’ interactions are presented in Chapter 4 and constitute interactive-based learning with a moderate level of interactivity, where the participants interact with virtual agents. In that study we found no differences in memory performance between interactive and non-interactive virtual agents. The task in that study also contain both a monologue and dialogue part of the conversation. The aim of the present study is to investigate ‘video’ recordings, which constitute observational learning with no interactivity, where the participants watch pre-recorded videos. Consequently, the task in this study only contains the monologue part of the conversation. The overall aim of this study is then to contrast all three levels of interactivity, where ‘real’ interactions involve a high level of interactivity between participants; ‘virtual’ interactions which involve a moderate and artificially generated level of interactivity; and ‘video’ recordings which involve no interactivity, in a cross-experimental design to investigate the way that we learn new information depending on the level of interactivity from different mediums. Understanding how learning is affected by different types of social interactions is important for education and training in many contexts. This has become even more important during the COVID-19 pandemic, where social interactions have been constrained across all domains of our lives. This research could be beneficial when creating tools for online learning and to understand how we can best implement the benefits of real social interactions in an online or virtual setting. In the next sections we will describe the tasks used to present the ‘video’ condition in more detail.

5.3.1 Conversation Task and Memory Test

In this study we employed the same task that was used in Chapter 3 about American state facts (See Section 3.4.3 for details). But instead of a real-world conversation it was performed on an individual basis where participants took turns to read cards when engaged with pre-recorded video clips of other participants doing the task (See Section 5.4.3 for details). Following this task, the participants were asked to complete a surprise memory test on a separate computer to assess how many facts they remember correctly from the conversation (See Section 3.4.3 for details).

5.4 Methods

5.4.1 Participants

36 participants (M_{age}=26) were recruited from the UCL Psychology Subject Pool and the ICN Subject Database. Exclusion criteria included subjects that were (1) not fluent in English (2) not between the ages of 18-35, and (3) did partake in the study presented in Chapter 3. The participants did not have any previous experience with the tasks involved and were unaware of the purpose of the experiment. Ethical approval was arranged via the UCL Research Ethics committee, and all participants gave their written informed consent. A monetary reimbursement was offered for participating in the study at a rate of £7.50/hour.

5.4.2 Equipment

In this study, we employed the same task that was used in Chapter 3 (See Section 3.4.3 for details). However, this time it was performed on an individual basis where participants took turns to read cards to camera or to watch pre-recorded video clips of other participants doing the task. The video clips were recorded when we tested the first 12 dyads from the study presented in Chapter 3. A video camera was positioned out of sight behind the yellow participant. The camera was positioned to provide participants in this study with a similar camera angle to what the yellow participant could see. In addition, the frame of the camera only included the blue participant's head and shoulders, restricting the social signals that may have aided learning. As only the blue participant was video recorded, the set of 8 cards given to blue was alternated from session to session the same way as in Chapter 3.

Prior to testing, the 12 video recordings of the blue participants were edited by MSc student Jessica Kankram, using Microsoft Photos. The videos were shortened to 8 clips of 1 minute, which each contained the blue participant reading the facts from each card (the monologue section from the task used in Chapter 3). These clips were then combined to make a longer video. After each clip, two screens instructing the participants what to do were seen before the next clip began to play (Figure 5-1).

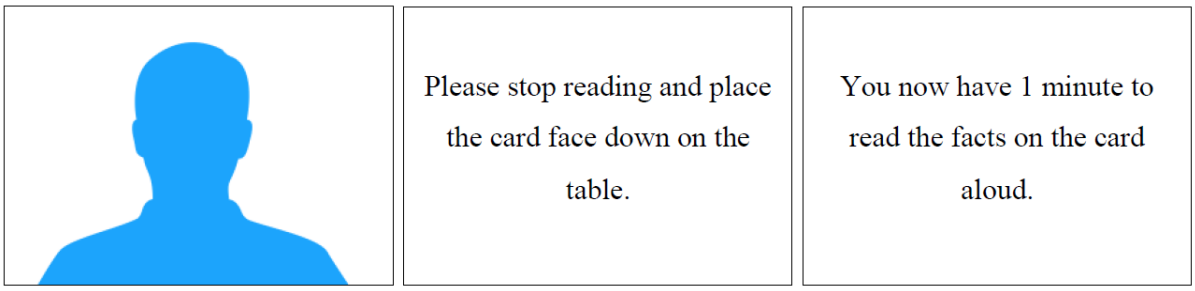


Figure 5-1. Storyboard for the video. The first image (left) represents the pre-recorded video of the blue participant reading the facts on a card. The second box shows the screen that followed (middle). This screen was accompanied by an audio beep which signaled the beginning of the 1 minute. The final screen (right) was again accompanied by an audio beep. This storyboard was repeated 8 times with a different card read by the blue participant each time and compiled into one video.

5.4.3 Procedure

Participants arrived at the lab and were informed of the procedures, after which they signed the informed consent. Participants were seated at a desk in front of a laptop with 8 cards placed face down in front of them. The edited audio-visual recording was played in full screen mode, with onscreen instructions guiding the participant.

Video Task. The task in this study followed a similar format to that in Chapter 3 but performed individually with pre-recorded video clips. Like the study in Chapter 3, each participant was assigned 8 cards. However, instead of going through the procedure with a real conversation partner, or a virtual agent as in Chapter 4, the participants were asked to watch one of the edited audio-visual recordings. The recordings were divided on a 3 to 1 basis, such that each blue participant from Chapter 3 was watched by 3 participants in this experiment, for a total sample of 36 participants. This means that our results should generalise across different videos and are not specific to the use of one particular video recording as a stimulus. The participants taking part in this experiment thus alternated between reading their assigned cards aloud (i.e., speaking) and watching pre-recorded clips of the monologue section from 12 of the blue participants in the study presented in Chapter 3 (i.e., listening). No time was assigned for discussion. In summary, the video began by playing a 1-minute section of the clip of a blue participant reading a card while the participant listened. The participant was then given 1 minute to read one of their assigned cards from the pile. Following this, the video returned to the blue participant reading the next card and so forth until all 16 cards had been covered. After completing all trials, the participant was moved to a separate table to complete the memory test (See Section 3.4.3).

5.5 Data Analysis

5.5.1 Analysis of Memory Performance

We performed a basic analysis of the participants' general memory recall, like the 'real' interactions in Chapter 3, and the 'virtual' interactions in Chapter 4. The memory decisions were categorized into one of four response categories based on the "Old/New" status of the fact and the given response: hit, miss, false alarm, and correct rejection responses on the memory test (Section 3.5.1). The participants' ability to discriminate Old from New facts on the memory test was calculated by considering the number of hits and correct rejections (i.e., the sum of Old/New responses correctly identified as Old/New) and presented as a percentage of correct recall (hits + correct rejections divided by 2) across both speaking and listening trials.

5.5.2 Mixed-Effects Model Analysis

After performing the basic analysis of general memory recall, we aimed to test the way that we learn new information depending on the level of interactivity from different mediums. The final sample across all chapters was used to create three models (M_1–M_3) that consisted of 114 participants in total. The data included in the mixed-effects models were the percentage of memory recall responses (0-1), and the dummy codes for Experiment (1 = Real Interactions; 2 = Nodding VR Interactions; 3 = No nodding VR Interactions; 4 = Video Interactions), Role (1 = speaking, 0 = listening), and the unique code for each participant across all three experiments (1-146, *note:* the 32 participants in the nodding VR interactions also participated in the no nodding VR interactions):

M_1: *Recall ~ Role* + (*1* | *Participant*)

M_2: *Recall ~ Experiment* + *Role* + (*1* | *Participant*)

M_3: *Recall ~ Experiment * Role* + (*1* | *Participant*)

We estimated the data of the sample using multilevel statistical modelling, and a model comparison approach (Judd et al., 2008). Prior to model comparison, we performed a linear multilevel regression for all models. We used two-level models with Experiment (real, interactive VR, non-interactive VR, video) and Role (speaking, listening) as predictors (level 1) nested within participants (level 2). We had no interest in analysing the grouping variable of 'participant' as a random effect but needed to factor this out for individual variation in the model parameters. The dependent variable was the memory performance (% correct recall). Model M_1 was used as a baseline model to compare the goodness of fit of M_2 and M_3. We combined the two factors into a single saturated model M_2, using both Experiment and Role to predict memory performance. The final model, M_3, was created to test for an interaction between Experiment and Role on the dependent variable. We follow the same reasoning as we did in Chapter 3 for the mixed-effects model analysis patterns (See Section 3.5.2 for more details). We present two major analyses of our data (i.e., the multi-level modelling and the model comparison approach) for fairness and completeness. Furthermore, instead of analysing the fixed effects parameter estimates of the mixed-effects model using dummy codes for our categorical variables like we did in Chapter 3 with the continuous variables, an ANOVA was performed on M_3 to explore the factors Experiment (real, interactive VR, non-interactive VR, video) and Role (speaking, listening) on memory recall, as well as any factor interactions.

5.6 Results

5.6.1 Mean Memory Performance

We performed a basic analysis of the participants' general memory recall, presented as a percentage of correct recall across both speaking and listening trials. The results can be seen in Figure 5-2, together with the results from Chapters 3 and 4.

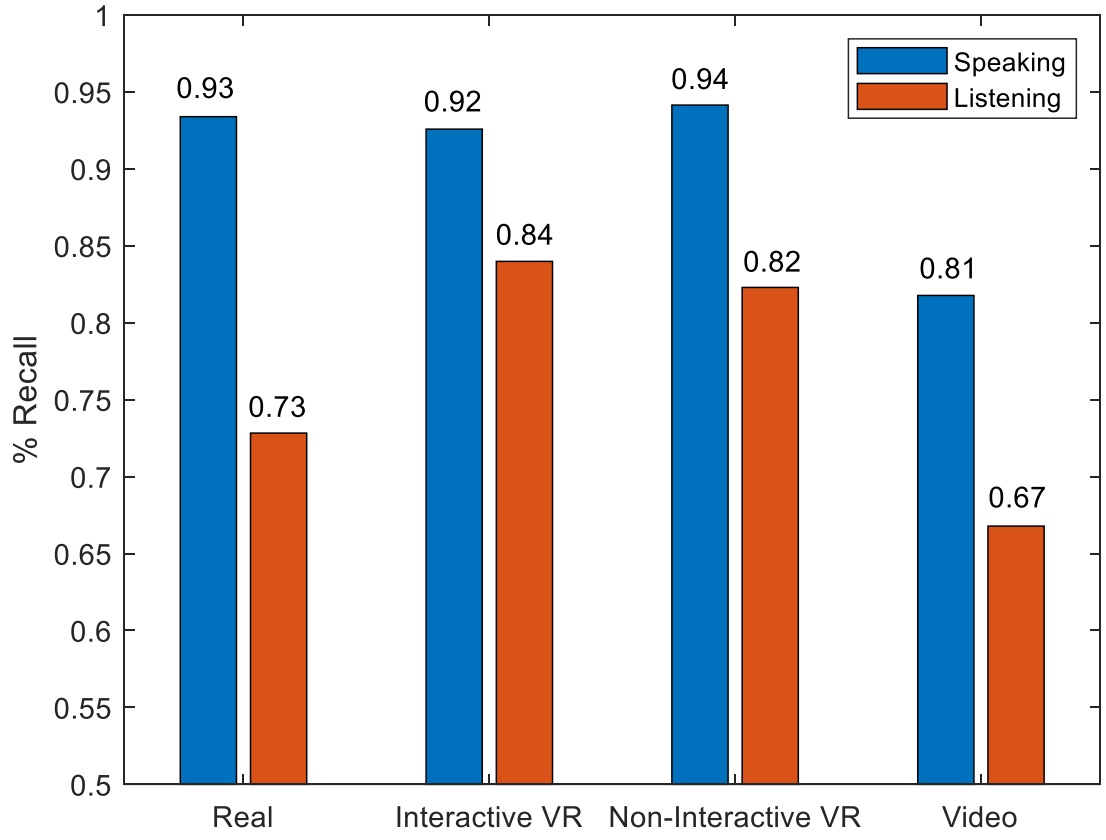


Figure 5-2. Mean memory performance across experiments. The barplot shows the percentage of general memory recall for both speaking and listening conditions across the three experiments. The 'virtual' interactions from Chapter 4 are further split into interactive and non-interactive engagement.

We can observe a higher recall rate when speaking compared to listening across the experiments, indicating that participants are more likely to remember information when reading aloud instead of listening. Independent-samples t-tests were conducted to compare memory recall across experiments. The results show that

when speaking there is a significant increase in memory recall in real interactions (M = 0.93, SD = 0.08) compared to the video condition (M = 0.81, SD = 0.11), $t(80) = 5.25$, $p<.0001$. However, no significant difference was observed between the real interactions (M = 0.73, SD = 0.19) and the video condition (M = 0.67, SD = 0.13) when listening, $t(80) = 1.64$, $p=.1$. These results confirm the existence of a 'video deficit' compared to real interactions, but only when speaking which indicates that participants recall less when talking in the video condition.

We found a significant increase in memory recall in both virtual interactions (M = 0.93, SD = 0.08) compared to the video condition (M = 0.81, SD = 0.11) when speaking, $t(66) = 4.47$, $p<.0001$. Here we also found a significant increase in memory recall in virtual interactions (M = 0.83, SD = 0.1) compared to the video condition (M = 0.67, SD = 0.13) when listening, $t(66) = 6.05$, $p<.0001$. These results confirm a 'video deficit' when compared to virtual interactions as well.

This gives us reason to think that virtual interactions have more in common with real interactions than the video condition. Comparing them reveal no significant increase in memory recall in the virtual interactions (M = 0.93, SD = 0.08) compared to the real interactions (M = 0.93, SD = 0.08) when speaking, $t(76) = 0.42$, $p=.674$. However, a significant increase in memory recall was observed in the virtual interactions (M = 0.83, SD = 0.1) compared to the real interactions (M = 0.73, SD = 0.19) when listening, $t(76) = 3$, $p=.003$. Interestingly, this shows that participants generally recall more information when listening to virtual agents compared to what they do when they listen to participants in real face-to-face conversations.

5.6.2 Memory Across Experiments and Roles

To give a more detailed analysis of memory performance across the experiments, we used multilevel statistical modelling and a model comparison approach. Prior to model comparisons, we performed a linear multilevel regression on all models. We used two-level models with Experiment (real, interactive VR, non-interactive VR, video) and Role (speaking, listening) as predictors (level 1) nested within participants (level 2). The results from the analysis are presented in Tables 5-1 and 5-2.

Table 5-1

Mixed-effects model comparisons for general recall.

Model Comparison	Mixed-Effects Models
M_1  M_2 ****  M_3 ***	Recall ~ Role + (1 | Participant) Recall ~ Experiment + Role + (1 | Participant) Recall ~ Experiment * Role + (1 | Participant)

Coloured arrows represent the best fit model for each comparison measured by differences in AIC-scores, and if the alternative model was accepted (Green) or rejected (Red) in favour of the compact model measured by the likelihood ratio, *$p < 0.05$; **p < 0.01; ***p < 0.001; ****p < 0.0001. *Note:* M_3 includes the fixed effects of Experiment and Role, as Matlab automatically calculates those when running the interaction model.

When comparing the baseline model 'Role' (M_1) with the saturated model (M_2), the results show that the inclusion of 'Experiment' (Arrow A, diff-AIC=32.76, p<000.1) significantly improved model fit. Furthermore, when comparing the saturated model (M_2) with model (M_3) which added the interaction effect between Experiment and Role in addition to the main effects of Experiment and Role, the results showed that there is an interaction effect between the level of interactivity of the experiment and what role the participants had in terms of speaking or listening (Arrow B, diff-AIC=8.92, p=.001). The ANOVA on model (M_3) is listed in Table 5-2 below.

Table 5-2

Results of the ANOVA for the generalized linear mixed model M_3

Model	Term	FStat	df1	df2	*pValue*
M_3: Recall ~ 1 + exp * role + (1 | ppt)	Intercept	1231.2	1	284	2.92e-105
	Experiment	4.1871	3	284	.006
	Role	107.53	1	284	1.41e-21
	Exp * Role	5.2386	3	284	.001

The ANOVA show a significant main effect of Experiment, $F_{3, 284}$ =4.187, p=.006, and Role, $F_{1, 284}$ =107.5, p<.0001, on memory recall. A significant interaction effect, $F_{3, 284}$ =5.238, p<.001, was also observed.

5.7 Discussion

In this study, we collected new data where participants acquire information from video recordings, aimed to explore the differences in memory across three experiments. This has allowed us to investigate the way that we learn new information depending on the level of interactivity from different mediums (i.e., real, virtual, and video). Chapter 3 focused on real interactions with a high level of interactive-based learning; Chapter 4 on virtual interactions with moderate level of interactive-based learning; and this chapter on video recordings with observational learning with no interactivity. In this cross-experimental study, we found significant differences between the experiments and the level of interactivity, as well as who spoke and listened during the conversation.

Consistent with results from Chapter 3, we observe an overall higher recall rate when speaking compared to listening, indicating that people are more likely to remember information when they are reading the facts aloud instead of listening. This is supported by a significant main effect of role (Table 5-2), and that the speaking condition is driving the effect (Figure 5-2). The model comparisons also reveal that the inclusion of 'Experiment' significantly improved model fit compared to a baseline model without this factor (Table 5-1). This is supported by a significant main effect of experiment (Table 5-2) which shows that there are differences between the experiments and the level of interactivity we engage with. These results suggest that the level of interactivity from the different mediums, as well as if the participants are speaking or listening, influences how much the participants remember from the conversation. Furthermore, the model comparisons reveal that the inclusion of an interaction term significantly improved model fit compared to the saturated model without this term (Table 5-1). This is further supported by a significant interaction effect between the two factors (Table 5-2) which shows that we must be careful when interpreting the main effects of the level of interactivity and if the participants were speaking or listening since the two factors depend on each other. In the following paragraphs we discuss the effects between the experiments.

Real > Video Interactivity. Although we made no specific predictions on the direction of these effects, a closer look reveals that participants engaged in real social interactions were seen to remember more than those engaged with the video recordings when speaking, but not when listening. Consistent with previous studies (Anderson & Pempek, 2005; Krcmar et al., 2007; Roseberry et al., 2009) our results help confirm the existence of a 'video deficit' and the evidence that real social interaction is important for memory and learning. This is also consistent with studies

that both children (Kuhl et al., 2003) and adults (de Felice et al., 2021) learn better from real social interactions compared to when they watch the same person on video. However, we only find this effect when speaking and not when listening. Thus, when listening to video recordings, these results are more in line with previous work on adults that shows mixed results with no difference between interaction-based learning and observational learning (Davis et al., 2008; Schreiber et al., 2010; Solomon et al., 2004; Vaccani et al., 2016).

One explanation for the video deficit during speaking is that the participants in real social interactions can better connect with their conversation partners by increasing active participation and the multimodal coordination of both verbal and non-verbal social signals. As a result, active engagement in real interactions may have enhanced learning compared to video. Moreover, the added social pressure by having another person physically present during real interactions may have also led to an increase in the attention on the task. This may in turn have led to enhanced learning compared to the pre-recorded video task where there is no such social pressure. For example, as we saw in Chapter 3, when participants take part in interactive-based learning, they may engage in joint attention or finding common ground to connect with their conversation partners, which may allow information to be shared more effectively.

Virtual > Video Interactivity. We also find that participants engaged in virtual interactions were seen to remember more than those that engaged in the video interactions, both when speaking and listening. These results also confirm the existence of a ‘video deficit’ compared to virtual interactions and suggests that learning via interaction with virtual agents is more like that in real interactions. This makes sense and might be attributed to the increased ecological validity of the

interaction compared to pre-recorded videos. In other words, because virtual agents are responsive, they have more natural verbal and non-verbal social signals compared to video stimuli, which suggests that having a conversation with a virtual agent is closer to engaging with naturalistic stimuli. Because interaction matters, and people have more sensorimotor experiences with real human stimuli over non-human stimuli during development, human perceptual features may elicit stronger responses than stimuli with non-human perceptual features.

Virtual > Real Interactivity. Lastly, we find that participants engaged with interactive virtual agents remember more when listening compared to real interactions. By the logic of the previous argument, this finding is surprising in the sense that real interactions should elicit stronger responses than stimuli with non-human perceptual features. There may be several possible explanations for this finding, but the most obvious ones relate to differences in social pressure between the two types of interactions. First, virtual interactions may have fewer social distractions compared to real interactions. For example, in real interactions we often ask ourselves questions such as “does she like me?” or “is he judging me?”. This type of social pressure might add distractions to the conversation, which consequently make it more difficult to remember what is being said. Secondly, instead of the real interactions adding social pressure and contributing to an increase in attention to the task, another type of pressure arises when engaging with virtual agents because they can sometimes behave awkwardly, and the participants often focus more on the conversation to monitor such interactions. What this means is that, as the participants work harder to monitor the conversation, they pay more attention to the task and remembers more from what is said when they feel they are not getting any help from the other person to drive the conversation forward. This

increased monitoring of the virtual agent is cognitively demanding as the participant is carrying more of the load. This goes against the theory of least collaborative effort which asserts that people in a conversation try to minimize the total effort spent on the interaction, both when speaking and listening (Clark & Wilkes-Gibb, 1986).

The results of this study support the conclusion that interaction-based learning is more effective than observational learning, and that the level of interactivity of the medium we engage with enhances memory and learning in social contexts. This suggests that different neural and cognitive mechanisms may support interaction-based and observational learning (Rice & Redcay, 2016; Seuren, Wherton, Greenhalgh, & Shaw, 2021).

5.8 Limitations and Future Directions

A factor which may have accounted for the differences in memory performance between the experiments is that while previous studies have studied intentional learning, the studies in this thesis test unintentional learning. As participants were unintentionally learning when asked to read the facts, their attention may have been focused on not making mistakes while reading as opposed to trying to understand the facts being read (i.e., passive reading). Similarly, when the participants were listening, they may have been able to focus more of their attention on what was being said (i.e., active listening).

A second issue concerning the structure of the tasks between the experiments is the lack of a dialogue section in the video task. The addition of a dialogue section in the real and virtual conversational tasks meant that the participants conversation

partner could have reminded them of some of the information and consequently strengthened their memory of certain facts. Moreover, the conversation during the dialogue section of the tasks in the real and virtual interactions could not be controlled. This could have created an unfair bias with participants remembering the more interesting and/or more discussed facts over the others.

However, the video task used in this study goes beyond previous studies by using a carefully matched design aimed to increase its ecological validity, which allows the participants to feel like they are exchanging information by speaking and listening to another participant, while still lacking the interactive element from engaging with video recordings. This makes this observational learning task a 'no-interactivity'-level condition, where the learner is passively decoding an interaction that feels present and engaging. It may be that watching video recordings accentuates both the sense of engagement and the sense of disengagement, depending on whether a listener feels the speaker is directly interacting with them, and vice versa.

Understanding how learning is affected by social interaction is important for education and training in many contexts. This has become even more important during the recent coronavirus (COVID-19) pandemic, where social contact has been constrained across all domains of our lives and online education has moved learning online. The findings from this study that learning can vary in terms of the level of interactivity, will thus prove to be beneficial when creating tools for online learning that can implement the benefits of real social interactions in an online setting. For example, this study has demonstrated that virtual agents can help enhance learning over video teaching, and that there is even some untapped potential using virtual agents over real interactions. These findings can help educational institutions to

effectively transform formal education into online education with the help of virtual classes and other online tools in this shifting educational landscape.

However, further studies are needed to disentangle what mechanisms from real social interactions extend to virtual interactions (e.g., turn-taking, rapport building, and information sharing). Furthermore, seeing as there are different types of people who learn differently depending on several factors (e.g., visual learners, social learners), future studies could look at individual differences across different levels of interactivity. For example, some people might not be as affected by the social pressure of having a real interaction, while others learn better from video recordings.

In real-world education, teaching typically occurs in classes with larger groups of people. In these contexts, the teacher does not actively engage with each student throughout the lecture. However, interaction-based learning requires mutual feedback between student and teacher (Shamay-Tsoory, 2021). Learning in a classroom – either offline or online – implies learning in the presence of others. In other words, you learn *with* others, instead of *from* others. Consequently, the role of attention may be even more impactful in online teaching since it is fundamental to successfully acquire new information. However, just as learning in a classroom setting in the presence of other students can modulate arousal, attentional, and motivational processes (Guerin, 1986), so can the feeling of presence with another participant lead to improved learning (Lytle, Garcia-Sierra, & Kuhl, 2018), or make it more difficult (Skuballa, Xu, & Jarodzka, 2019). That said, given that our design only involves two participants in dyadic social interactions, the results in this study do not generalize beyond the dyad to larger groups or classroom settings.

Online interactions and learning are also tightly connected to an aging population, where the elderly often have difficulties with both memory and new technology.

While this study did not include a demographic for age, we could benefit from future studies investigating if age plays a role when learning in video or virtual settings and if there is an interaction between the nature of the interaction and memory when it comes to an aging population. These kinds of questions are often explored in the field of social robotics (Cross, Hortensius, & Wykowska, 2019) to assess the quality of assistive care and psychological well-being of the elderly.

5.9 Conclusions

In this study, we collected additional data where people acquire information in the context of video recordings, aimed to explore the differences in memory across three experiments with varying levels of interactivity (i.e., real, virtual, and video). This allowed us to investigate the way that we learn new information depending on the level of interactivity from different mediums.

We found significant differences between the experiments and their level of interactivity, as well as who spoke and listened during the conversation. When exploring the direction of these effects, we were able to confirm the existence of a 'video deficit' in learning compared to real interactions when speaking and conclude that real social interaction is important for memory and learning. We were also able to find a 'video deficit' in learning compared to virtual interactions, which shows that interactive virtual agents can help enhance learning over video teaching. This is supported by a surprising finding that we remember more when engaged with

interactive virtual agents compared to real interactions when listening, which could be due to the added social pressure of having an awkward non-human present in need of help to drive the conversation, which results in closer monitoring of the conversation and better memory.

We have shown that the level of interactivity supports memory and learning during conversations and improves information transfer across people. These findings contribute to our understanding of human adult learning and the importance of interaction-based learning over observational learning. Further exploration of the differences between social interaction in different online settings may prove to be valuable when creating tools for online learning amid a changing landscape for teaching due to the recent coronavirus (COVID-19) pandemic.

Chapter 6. General Discussion

6.1 Summary of Experimental Chapters

The convergence of research questions in both psychology and computing sets the scene for the studies presented in this thesis, which draws together these diverse research areas, combining cognitive and psychological hypothesis testing with new advances in motion capture, signal analysis, and virtual reality to provide a new level of understanding of dyadic social interaction. Our methodological aim has been to *measure*, *analyse*, and *artificially generate* dyadic social interactions in real time.

However, to advance in the field it is important that these methods are guided by precise and well-defined psychological or cognitive theory to work. The overall cognitive aim of this thesis has been to understand and learn more about head nodding behaviour during face-to-face dyadic conversations in a naturalistic setting.

In Chapter 2, the aim was to capture two patterns of head nodding signals – fast nods and slow nod coherence – and determine what they mean and how they are used across different conversational contexts. We find that fast nodding is present in contexts when new information is exchanged and slow nodding only in a structured one-way information sharing type of conversation. This provides initial evidence that fast head nods are a signal of having received new information and that it has different meaning to that of slow nodding coherence. A secondary aim in Chapter 2 was to understand if there are reliable individual differences in social signalling behaviour. We find that nodding is consistently driven by context but is not a useful measure of individual differences in social skills.

In Chapter 3, based on the main findings from Chapter 2 that fast nods may be a signal of having received new information, the aim was to further investigate if head nodding between two people in a conversation might be related to measurable outcomes of the conversation, including how much the two people remember new

information and how they relate to each other in terms of self-other processing. This study provided initial hints that there might be a relationship between head nodding behaviour and memory performance, but further analyses were less clear.

In Chapter 4, we built on the preliminary results of Chapter 3 that initially showed a correlation between fast nodding and memory and aimed to further investigate a causal link between interactive engagement with a virtual agent and memory performance. We created a virtual agent who could show our head nodding behaviour rules and tested how much people remember from a conversation with the agent. The results from this study demonstrate no causal link between the interactivity of the agent and memory performance. In addition, in this study we also aimed to investigate if interactive head nodding can be used to measure how much we like the virtual agent, and whether we learn better from virtual agents that we like. We report no significant results between measures of liking with no reliable correlations to head nodding or memory.

In Chapter 5, we analysed data from Chapters 3 and 4 together with new data from a video-based conversation task. We aimed to summarise how the level of interactivity in different contexts (i.e., conversation with a real human, conversation with a virtual agent, task with an unresponsive video) impacts on the memory performance of the participants. The results show that the level of interactivity supports memory and learning during conversations.

In the following sections of this chapter, we will widen the scope and discuss the methodological (Section 6.2) and theoretical (Section 6.3) implications and developments of this thesis, as well as its limitations.

6.2 Methodological Implications and Developments

A running theme throughout this thesis has been to use new methods for *measuring*, *analyzing*, and *artificially generating* social interaction. To advance the field and overcome the limitations associated with traditional studies of interpersonal coordination, we need to constantly adapt and refine new methodologies. In the next three sections we will discuss the implications of this thesis in terms of methodological improvements from previous studies and its limitations from the perspectives of *Measuring* (Section 6.2.1), *Analysing* (Section 6.2.2), and *Artificially Generating* (Section 6.2.3) interpersonal coordination.

6.2.1 *Measuring* Interpersonal Coordination

The first step in our approach was to *measure* interpersonal coordination in real time using high resolution motion capture. In this thesis, we have tried to demonstrate that current motion capture technologies can be valuable in collecting rich datasets, because they provide many opportunities for investigating interpersonal coordination at a variety of levels. In the following paragraphs we point out some of the improvements and limitations with our experimental setup.

High Resolution Data Capture. Capturing behaviours in high resolution is the first step in reverse-engineering social interactions and picking apart the behaviours to determine which parameters are important to advance current theories (Hadley, et al., 2022). We use a motion capture system to identify potential behaviour rules based on the timing and frequency properties of head nodding from two different interpersonally coordinated behaviours – slow and fast nods.

This is a clear improvement on traditional approaches of using manual annotation methods, which often yield low resolution data and is very time consuming (Condon & Ogston, 1966; Kendon, 1970). Motion capture also provides more objective and detailed data in 3D about an interaction compared to previous studies using automated image-processing (i.e., 2D) methods like 'motion energy analysis' (Fujiwara & Daibo, 2016 Paxton & Dale, 2013; Ramseyer & Tschacher, 2010; Schmidt et al., 2012) or computer vision analysis (Dunbar et al., 2014).

Motion capture studies often focus on recording specific body movements in different scenarios, and the most relevant study for this thesis has been the study by Hale et al. (2020). Here the researchers recorded head and torso coordination in dyadic interactions using a Polhemus magnetic motion tracking device. By only using a single sensor or marker placed on the participant's forehead to record head movements in 60 Hz, they were able to detect two types of frequency head nods (i.e., fast and slow). The studies presented in this thesis improve upon this methodology to investigate dyadic social interaction using higher resolution full-body motion capture technology (Optitrack, NaturalPoint Inc., v.1.10), consisting of eight cameras (4 × Prime 13 and 4 × Prime13W) with a sampling frequency of 120 Hz. Instead of a single marker, each participant wore an upper-body suit with a set of 25 pre-determinately placed magnetic markers. 3 out of those 25 markers are placed on a cap and triangulated to account for the individual shape of the participant's head. This allows us to record head nodding in rich detail together with other body movements like hands and torso in a 3D space. This is important since head nodding should ultimately not be viewed and addressed in isolation from other behaviours.

Multimodal Data Capture. Most current research focuses on trying to capture and analyze a single modality only, with very few approaches presently trying to integrate more than one channel. However, taking into consideration multiple modalities can help to relieve ambiguity that is typical of unimodal communication. Within the scope of this thesis, we are presenting the analysis of a single modality (i.e., head nodding), but this signal is captured from a rich multimodal setup where we simultaneously recorded other body movements, along with gaze behaviour, facial expressions, and speech. What this means is that our head nodding signals are performed within the context of other modalities, both verbal and non-verbal, which makes them more naturally performed compared to a more constrained experiment. It also means that we could go back and analyze in future studies the data from different modalities, like we have done with gaze patterns (Dobre, Gillies, Falk, Ward, Hamilton, & Pan, 2021). Recording multimodal data also allow us to integrate different modalities and explore the relationships between them, and future studies need to address this challenge to fully understand social interaction.

From this, various multimodal questions can arise. For example, what is the relationship between head nods and verbal turn-taking? In fact, nodding does not occur on its own. The relation between head movements and speech has been investigated numerous times (Duncan, 1972; Hadar et al., 1983; Kendon, 1970), and both non-verbal and verbal channels must be considered together to understand conveyed meanings during conversations. For example, a large body of research shows that gestures generated during speech are, along with the words, part of an integrated speech production system (Goldin-Meadow & Alibali, 2012). We believe that studying multimodal aspects of social interaction is an essential next step for this research area, and our multimodal data capture setup enable us to do that.

Synchronized Dyadic Data Capture. Synchronizing multimodal data intra- and interpersonally faces problems not found in traditional research areas studying isolated participants responding to stimuli or interacting with confederates. Recent research has called for a 'second-person neuroscience' aimed at understanding naturalistic social interaction, involving multiple modalities across different contexts and timescales (Heerey, 2015; Schilbach et al., 2013).

In this thesis, although we analyze one specific modality in various scenarios, we are responding to this call by establishing a synchronized data collection protocol that can record multimodal data in naturalistic dyadic interactions. This setup is flexible enough to capture and *synchronize* body movements, eye-gaze, and speech between two or more people interacting.

This is not the first dataset to address multimodal data collection, and there are many corpus studies out there aimed at providing information about the way different modalities shape the structure of a social interaction and convey the speakers' cognitive and affective state in any given moment (including backchannel responses) (e.g., Vogel & Koutsombogera, 2018; Vinciarelli et al., 2012). So far, much effort has been devoted to recording facial expressions, body movement and speech data, but head movements have received less attention in relation to multimodal integration (Heylen, 2006). The importance of this signal in the turn-taking process and close relation with speech makes them important to include in the recording protocol, especially if aimed at recording real face-to-face conversations. The few multimodal corpora studies that include head movements (Aubrey et al., 2013; Carletta, 2007) do not capture the signals in high resolution within an experimental setting where speaking-listening roles or turn-taking structure can be manipulated.

In this thesis, we improve upon these previous studies by creating a high-resolution multimodal data collection protocol with wide synchronization of socially relevant data, where we account for the temporal characteristics between the different modalities, as well as the coordination of these modalities between two people. This will allow for each modality to produce unique signals within the context of other signals in a more natural way and analyse how they are coordinated between people. However, even though we have an ambitious framework to work from, investigating multimodal integration is considered beyond the scope of this thesis, and instead we focus our analysis on the investigation of how a single modality (i.e., head nodding) is coordinated in dyadic social interactions.

One of the difficulties in devising a multimodal data collection protocol is properly controlling for behavioural variables to ensure that the effects seen are not merely driven by effects created by the setup or the equipment itself. Another disadvantage is that a multimodal data capture system is very obtrusive and requires specialized equipment and a dedicated recording space. It can also be very costly and time-consuming, but we believe that these more complex and contextually dependant experimental setups will help us in understanding the interactive dynamics of interpersonal coordination by leading us to more informed decisions when identifying our behaviour rules. However, these setups also provide us with new challenges in analysing the data.

In the following section, we will discuss how we utilize advances in wavelet analysis to understand social coordination with better temporal detail and how we use our synchronized data to improve how we quantify the interpersonal coordination of head nodding between people.

6.2.2 *Analysing* Interpersonal Coordination

The second step in our approach was to *analyse* interpersonal coordination and its temporal progression in relation to dyadic conversation using advanced wavelet analysis methods. This method is being used in studies on dyadic interaction data with interesting results (Issartel et al., 2015; Schmitd et al., 2014), and has allowed us to tease apart the frequencies within the head nodding signals as the interaction unfolds. In the following paragraph we discuss some improvements and limitations with our approach to analyzing the time-frequency coordination of our data.

Wavelet Analysis. Social interaction can be seen as a dynamic system where each part or modality must be working in parallel with each other and where each part is subject to previous influence, as well as shapes future parts of the interactive system. Thus, social interactions gain their structure through a process of continuous adaptation in which the various parts involved, behavioural and cognitive, mutually influence and constrain each other within and across conversational partners. The word ‘dynamic’ can be used to describe such changes over time.

A major challenge for modelling the dynamics of interpersonal coordination is how to integrate the data from different modalities and from different people across different timescales and frequencies (Fujiwara & Daibo, 2016; Grinsted et al., 2004; Issartel et al., 2006). The previous section should make clear that social interaction involves various dynamic patterns and that analyzing any social signal, or collection of signals, poses significant challenges. One method, which we have tried to demonstrate with the studies in this thesis, is to translate the data into a form that allows the analysis of the degree of similarity between two sets of data – or time-series, and the progression in time of this interaction using cross-wavelet coherence analysis (CWC). With CWC we may be better prepared to explore how different

verbal and non-verbal communication channels change over the course of an interaction. While it is a relatively new tool for studying social interactions (Issartel et al., 2015), several proof-of-concept studies using CWC have found that it can be applied to time-series data on body movements in a dyadic interaction (Sofianidis et al., 2012; Varlet et al., 2011; Washburn et al., 2014). However, most of these studies involve interactions that have structured turn-taking behaviour. Fujiwara & Daibo (2016) followed this up with a CWC study that did not involve a specific task but focused on unstructured conversations. In this study the researchers also implemented a pseudo-pairing paradigm proposed by Bernieri and Rosenthal (1991), where they reordered video clips in a random order and then compared them to the real pairs with the idea that there would be more coordination in the real pairs than in the pseudo-pairs.

Hale et al. (2020) improved upon the study by Fujiwara and Daibo (2016) by (1) instead of recording motion frequency with 2D video, they recorded in 3D using motion capture to isolate the different movement coordinates of the head (i.e., pitch, yaw, roll); and (2) instead of testing whether their coherence pattern was present in the pseudo interactions *within* the pairs, they tested *between* pairs, which provide a stronger test. Their results revealed two types of coherence patterns, one positive coherence pattern at low frequencies, and the other an unexpected lower than chance coherence at higher frequencies. This is similar to a finding by Healey et al. (2014) where they found that people in dialogue systematically diverge from one another in their use of syntactic constructions. Hale et al. (2020) hypothesized that the two patterns could be different signals carrying distinct social information.

This thesis has explored the hypothesized distinction between these two coherence patterns, as it was not clear whether they were two distinct social signals.

The quantification of these two types of signals would also be useful for allowing us to identify them as two separate behaviour rules based on different frequency ranges of head nodding. When exploring this distinction, we improved on the study by Hale et al. (2020) by having a larger sample of dyads interacting in a full-body motion capture space to achieve greater (1) High resolution data capture; (2) multimodal data capture; and (3) dyadic data capture. We also used their improved pseudo-matching paradigm to get a stronger test where the pseudo pairs have the same general movement characteristics as the real pairs. In other words, by using within-dyad pseudo-trials, we can control for the unique behaviour of each individual and identify only the coordination patterns which are specific to the live interaction of the participants. We believe this is a strong analysis that should be used more often with data from dyadic interactions.

However, compared to the previous studies (i.e., Fujiwara & Daibo, 2016; Hale et al. 2020), we only analysed the coherence measure (R^2) which tells us if two people move at the same frequency within the same time-window, but not the phase measure (i.e., time lag) between the participants. This limits us in terms of looking at specific time lags of mimicry, for example. We initially aimed at performing cross-correlations between each participant's head pitch to compensate for the lack of phase measures, however, for reasons we discuss in more detail in Chapter 2, we decided not to further investigate either using a phase measure nor cross-correlations, with or without absolute values.

In this thesis, we have tried to integrate the benefits of *measuring* high resolution synchronized recordings with *analyses* that can tease apart the frequencies in the data. We believe that this approach opens exciting possibilities for new experimental paradigms and can help to *artificially generate* interpersonal coordination.

6.2.3 *Artificially Generating* Interpersonal Coordination

The third and final step in our approach was to *artificially generate* interpersonal coordination using our behaviour rules in virtual agents that can enact and engage in conversation with participants. This method is then used to explore how the participants respond to different agents showing different behaviour rules to test our hypotheses. In the following paragraphs we discuss some improvements and limitations with our approach to artificially generating behaviour rules in the agents.

Experimental Control. We currently lack detailed knowledge on many of the behavioural parameters during naturalistic interactions, and even less so when it comes to the interplay between multiple modalities. Hömke et al. (2017, 2018) provided a good example of a modern approach to studying interpersonal coordination. In their studies, they first *measure* real-world blinking, and *analyse* blinks in relation to dyadic conversation, and then they *artificially generate* a virtual agent who could blink to test how people responded to blinks. This thesis has followed the same kind of approach. Both Hömke's studies and this thesis has relied on the idea that specific social behaviours can be identified and understood in terms of 'behaviour rules' (Hadley et al., 2022).

Common to many experimental studies of social interaction is finding the right balance of ecological validity and experimental control; for example, by restricting tasks or assigning speaker and listening roles in advance. Using virtual agents as models provide us with strong manipulations of isolated behaviours to test our hypotheses, because it allows us to specifically control and manipulate one type of body movement (i.e., head nods) while keeping other body movements the same (eye-gaze, blinks, lip movement). In support of the two distinct frequency patterns of head nodding that Hale et al. (2020) found, we showed in Chapter 2 that these

patterns are indeed two different signals with different meanings. This is useful for allowing us to identify two plausible behaviour rules with different characteristics. These are (1) show fast nods to simulate natural backchanneling behaviour, and (2) show slow nods to simulate natural mimicry behaviour. A behavioural rule like this is then relatively easy to implement in a virtual agent to make it testable (Bailenson & Yee, 2005). Previous studies have mostly focused on the timing of these behaviours (e.g., backchannel responses). For example, Hömke et al. (2018) implemented behaviour rules as a WoZ method, where the experimenter controlled the timing of the inputs to the virtual character. In this thesis, we improved on these methods to increase experimental control by (1) programming the head nodding rules, along with all our other behaviours, into the virtual agents. This allows us to use both the timing and frequency properties of head nodding to build highly responsive agents based on real-world behaviour; and (2) by using the positional head sensors and audio feedback from the HMD, we also program the behaviour rules to be conditional on the participants behaviour. This allows us to implement the behaviour rules with full interactivity (i.e., “If the participant does X, the virtual agent should do Y at time Z”). Using this method, we also lessen the trade-off between the experimental control of being able to manipulate the behaviour of the virtual agents and its ecological validity in terms of how natural or close to real the social interactions feel to the participants.

Ecological Validity. Creating believable virtual characters and generating interactive behaviour is a great challenge for computer scientists (Pan et al., 2012; Rizzo & Talbot, 2016), and are the two most important areas in which researchers can improve the ecological validity of their virtual reality experiments. In this thesis, we have tried to improve in both these areas of research by (1) creating our virtual characters based on a study by Zell et al. (2015) where they demonstrate how we

can get out of the ‘uncanny valley’ by relying on stylization to increase the appeal of a character by exaggerating or softening specific features (See Figure 4-3). Thus, to make our characters appear as less uncanny, we chose to render them with a more cartoonish stylization; and (2) the task on discussing American states was entirely rendered in a virtual environment using fully immersive virtual reality HMDs. This is an improvement in terms of greater ecological validity from earlier studies using 3D projector screens (e.g., Hale & Hamilton, 2016b).

Having more natural behaviour represented by the virtual agents also leads to improved realism for the participants interacting with them, which is a good example for why virtual reality is a good tool to use in psychological experiments. In this thesis, we have improved upon the interactive behaviour of the virtual agents from previous studies in the field (Hömke et al., 2017, 2018) in two ways: (1) we have made the virtual agents behaviour conditional on the participant’s behaviour (e.g., to copy their head nods) and fully programmed into the virtual agents; (2) both the interactive and non-interactive agents were programmed to generate verbal and non-verbal behaviour to enhance the realism of the conversation for the participants. These behaviours included gaze, blinks, lip-syncing, head movements, and speech. Speech was the only behaviour that we designed as a WoZ, where the experimenter controls what the agents are saying. For this we used a scripted monologue for the monologue part of the task, and speech segments of potential dialogue options for the dialogue part of the task. Both were pre-recorded from two female voice actors. Important to note is that head nodding was the only behaviour that was experimentally manipulated, and all other behaviours, verbal and non-verbal, were included to enhance the ecological validity of the interactions but did not change between different experimental conditions. The WoZ system has its limitations, but it

is standard in the field (Jain et al., 2018) and provide a way for the participants to have an unstructured conversation that better represent how people interact naturally. We prepared a lot of general phrases and responses related to different conversational topics with accompanying facial expressions for the virtual agents. This method required significant training so that the 'wizard' could respond in a natural way, but this comes with the risk the of experimenter-expectancy effect, fatigue effects and human error. Thus, having more than one person as the 'wizard' could have improved this study. Future studies could also benefit from collecting the head nodding data from the HMDs to use when analysing relational data.

Experimental frameworks in virtual reality typically specify a certain behaviour or focus only on a small subset of the behaviours compared to real-world social interaction (e.g., Hale & Hamilton, 2016b; Hömke et al., 2018). This is due to the high computational demand of simulating a fully responsive multimodal and dynamic human social interaction. For example, in this thesis we have left out more detailed facial expressions, a multitude of language parameters, and posture to name a few. This is often decided based on what the current VR technology supports, which constrains what we can measure. Instead of trying to include everything, we chose to limit the behaviours to the ones that made the interaction between the agents and the participants work as naturally as possible. The VR method we have used in this thesis focus on one single modality to manipulate (i.e., head nodding), and does not make any claims beyond increasing ecological validity for the other verbal and non-verbal behaviours that is generated. Nevertheless, even very basic forms of social interactions can be enough to be perceived as realistic, which makes the use of simple behaviour rules ideal for using in virtual reality experiments as they are both perceivable and testable. For example, both behavioural (Wilms et al., 2010) and

neuroimaging (Schilbach et al., 2011) studies show that very basic but contingent eye gaze behaviour can be enough to elicit a sense of realism for participants interacting with virtual agents. However, future studies would benefit from including more interactive behaviours into their virtual agents.

Real-world social interaction is complex and dynamic. Non-verbal behaviour is highly changeable and can depend on the topic of the conversation, the surrounding context, and the person with whom we interact. Interactive virtual agents are not yet able to display this kind of dynamic range. Moreover, these behaviours are not static or temporally isolated events but evolve along with every interaction. Thus, the question we must ask is: How do we generate something that is constantly evolving? One way to approach this question is, as discussed, to perform integrated analysis on multiple levels. In this thesis we have been able to demonstrate how to analyse dyadic interaction (i.e., multi-person) and capturing data from a variety of interpersonal behaviours, verbal contributions, and contextual variables (i.e., multi-modal). The next challenge is to address how to analyse the integrative and dynamic aspects to understand how these behaviours coordinate together over time.

Another way to improve the ecological validity of virtual agents is to use machine learning methods. Machine learning is a computing method associated with cognitive simulation, or artificial intelligence (Michalski et al., 2013). It involves programming algorithms that can learn from and make predictions from datasets without having to manually code everything in detail. It allows researchers to automatically detect occurrences and model interactive patterns of data. In the context of modelling social interactions, this method can generate probabilistic models (Morency, de Kok, & Gratch, 2009) that predict behaviour over time of one person based on the other. For example, if the models derived by machine learning techniques is used to drive the

behaviour of an interactive virtual agent, then the virtual agent will be able to respond appropriately to the participant's behaviour in real-time. In this way, machine learning may be able to generate virtual agents that closely approximate real behaviour without the need to manually extract and code individual parameters like we did in Chapter 4 of this thesis (Gillies, 2009).

Other directions for future studies might involve looking at the neural correlates of social behaviours, because only when we can understand, model, and manipulate the system dynamics at the behavioural level can we begin to understand the roles that brain activity play in the creation of social signals. Two exciting neuroimaging methods that are being used to measure real-world social interaction include wearable neuroimaging systems like functional near-infrared spectroscopy (fNIRS) and brain-to-brain hyperscanning. fNIRS allows researchers to capture patterns of brain activation and mechanisms while people engage in naturalistic tasks like theatre acting (Hamilton, Pinti, Paoletti, & Ward, 2018). Several recent studies have also begun to investigate the interdependence of neural processes between two people as they interact using hyperscanning imaging methods (Konvalinka & Roepstorff, 2012). This allows researchers to monitor the brain activity of dynamically interacting participants, which will bring the research field closer to the study of real-world second person neuroscience (Schilbach et al., 2013).

Measuring, *analysing*, and *artificially generating* all aspects of behaviour remains a challenging problem, but all technical limitations, big and small, impose critical constraints on what psychology studies can be carried out (Pan & Hamilton, 2018). In the following section, we will discuss the theoretical developments and limitations of this thesis within the context of the methodological approach that we have used.

6.3 Theoretical Implications and Developments

With all these technological advances, it is sometimes easy to forget how essential it is to begin with developing a stable cognitive framework to ground our ideas in. As with any new method, it is best used to guide our research questions. Moreover, given the growing evidence of social coordination at different levels of analysis, we need to build coherent theoretical models that can explain the same phenomena. For example, motion capture or virtual reality data will require analysis that is different from that required by fNIRS or simple video analysis. However, the phenomenon to be explained is the same. A deeper understanding will arise as we bring together our methodological approach in conjunction with a good approach to theory. Together, method and theory will build on each other to create new testable hypotheses, which will lead to the development of new theories that can be challenged.

The step-by-step theoretical approach (See Figure 6-1) we have taken in this thesis has been to first *measure* and *analyse* different head nodding patterns in different contexts to identify different behaviour rules based on correlations between social context and nodding (Chapter 2). Based on such correlational findings, we can continue to build hypotheses linking a cognitive process and a particular head nodding pattern and explore this hypothesis by making correlational predictions on an outcome measure (e.g., memory) (Chapter 3). In the last step, we *artificially generate* behaviour rules to manipulate and test this correlation for causality (Chapter 4). This is like imposing experimental manipulations before the conversation, such as giving participants a goal to try to remember or setting up expectations about how their partner interacts (e.g., anti-social or outgroup member). We could then test if these manipulations change the head nodding behaviour during the conversation.

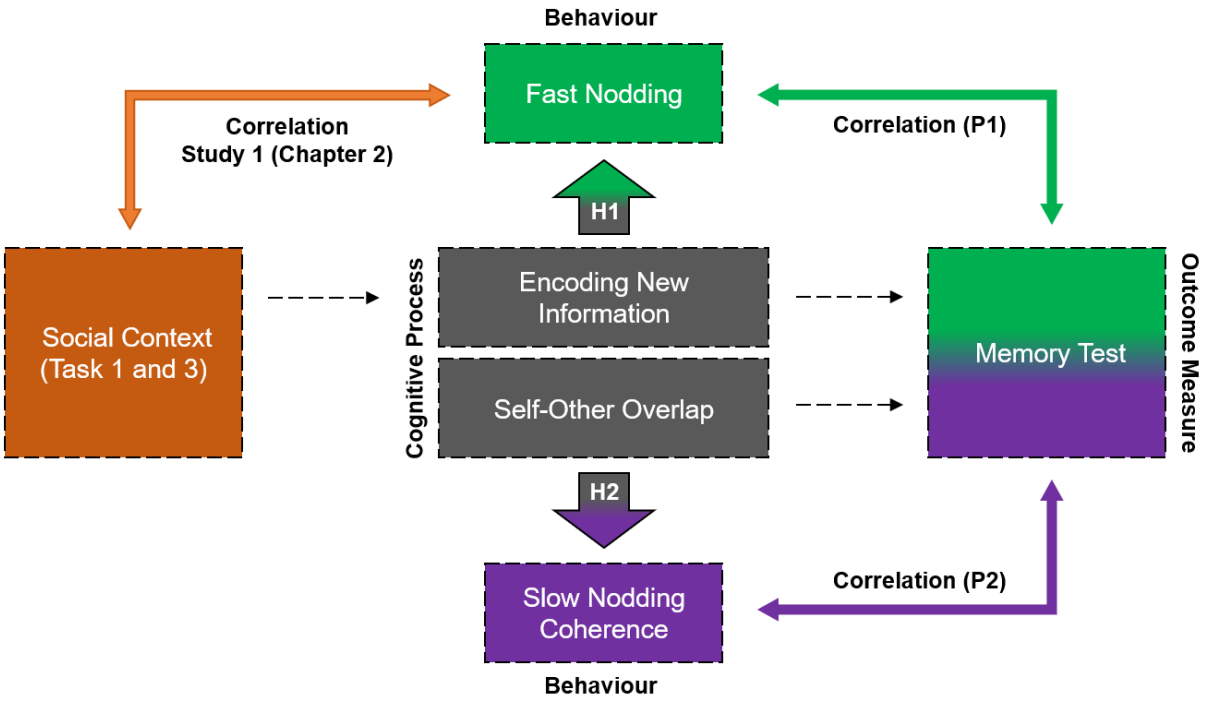


Figure 6-1. Cognitive approach. This box diagram, presented in Chapter 3, provides an example of the step-by-step theoretical approach we use to ask informed questions and hypotheses based on correlational findings and predictions that can then be tested using virtual agents to find a causal link.

The main objective of this thesis is to understand and learn more about head nodding behaviour during face-to-face dyadic conversations in a naturalistic setting. Specifically, we want to understand what two different type of head nodding patterns – fast nodding and slow nodding coherence – mean during different conversational contexts, how they are used, and if they relate to conversational outcomes like memory and liking. In the next two sections we will discuss the implications and larger developments of our findings for the field of interpersonal coordination from the perspectives of Fast Nodding and Backchanneling Signals (Section 6.3.1) and Slow Nodding and Behavioural Mimicry (Section 6.3.2).

6.3.1 Fast Nodding and Backchanneling Signals

In previous research Hale et al. (2020) identified patterns of high frequency fast nodding and low frequency slow nodding coherence in conversations but did not test whether these were distinct social signals carrying different social meanings. They proposed, however, that the fast nodding pattern might be a backchannel signal related to listening. Consistent with Hale et al., (2020) we were able to replicate these two frequency patterns using higher resolution methods. However, for these measures to be meaningful and used to quantify features of an interaction, it is important that we understand *why* people engage in these types of head nodding behaviours. In Chapter 2, we were able to test for the presence of fast nodding across different conversational contexts to determine if it really is being used as a backchannel signal. This hypothesis made sense to us since research have long recognized that both verbal (Sacks et al., 1974) and non-verbal behaviour (Argyle, 2013; Clark & Krych, 2004; Duncan & Fiske, 1977) from listeners could be backchannels that works as feedback to a speaker. McClave (2000) has further recognized head nodding specifically as a potential backchannel. In addition to this, Clark (1996) has reported that backchannel nodding is usually considered to be much faster type of nodding. It is a quick head nod that is visible but very subtle, and usually not something we notice. With our first study in this thesis, we had the opportunity to test if fast nods might be a backchannel signal.

Context and Meaning. Much of the research with non-verbal backchannels has been dedicated to identifying the social meaning behind them. For example, Hömke et al. (2017) investigated long blinks as an additional type of backchannel that could serve as a social signal. Boholm and Allwood (2010) considered the meaning of head nods with co-occurring speech and found that their main function was to

provide communicative feedback. Attributing social meaning to behaviours implies that behaviours are meaningful, understandable signals. However, the meaning of a social signal can be different depending on a lot of factors, which can make the behaviour ambiguous. Head nodding, for example, is particularly sensitive to conversational demands and convey several different meanings (Poggi et al., 2010). For example, some researchers suggest that the meaning of head nodding is to give feedback via backchannels that one has understood what the speaker is saying (Allwood & Cerrato, 2003; Duncan, 1972; Yngve, 1970), while others go further and claim that the listener must also accept or agree with what the speaker is saying (Heylen, 2006; McClave, 2000).

Context provides an important way to understand the meaning of social signals, because we would expect some signals to remain constant across contexts, while other might change. For example, if nodding is a social signal, then the meaning of the signal can be inferred by how nodding changes across such contexts. In other words, the meaning of a signal may change with context like the way the ambiguity of a word is generally overcome by considering the context it was uttered in.

In Chapter 2, we created three different conversational contexts with varying turn-taking structure (i.e., structured vs unstructured conversation) and how the information was shared between the participants in the dyad (i.e., shared information vs new information). Our results seem to indicate that fast nods might be a signal of having received new information during a conversation, as indicated from a significant increase in fast nodding behaviour in the contexts where there was a transfer of new information between participants. Such an exchange of new information may signal to the other that you have received their message and are paying attention to what they are saying. In the task where there was shared

information between participants, they may already have mutual understanding, or "common ground" (Clark, 1996) and find no use to signal "message received" to their partner since they share the same set of beliefs and knowledge. In a similar way, in the tasks where there was new information to be received, the participants may have used backchannel nodding to update their common ground or grounded the new information from their partner to build mutual understanding. However, communication is a complex multimodal system, and one of the areas this thesis has not touched upon is detailed verbal turn-taking. For example, an alternative explanation could be that in the task where the participants share information, they could have more equal turn-taking behaviour, which in turn might mean they do not need to produce as many nods as backchannels since producing a relevant next turn (i.e., verbalisation) also indicates understanding. However, looking at Figure 2-1B, we can see indications that the turn-taking behaviour is relatively varied.

Hale et al. (2020) examined fast nodding only in the context of a structured conversation (i.e., picture description task). In this context, participants naturally move their heads when alternating between looking down at the picture and at the other participant. This reflects a joint attention behaviour because the speaker has an important gaze target (i.e., the picture) (Emery, 2000). This behaviour is a potential confound since it produced the same kind of nodding movement as the one we were trying to measure, namely a vertical head movement in which the head, after a slight tilt up, bends downward and then goes back to its starting point (Poggi et al. 2010). Thus, the picture description task is unable to discriminate between a head nod and simple gaze following or joint attention. In Chapter 2, we improved upon the design by Hale et al. (2020) and included two additional tasks with unstructured conversations with different information sharing context (i.e., shared

information vs. new information). This way we could try to tease apart what a head nod means in different contexts. A similar problem could occur from our experimental setup, in which there are moments where the participants turn their head to look at the empty projector screen where they last saw the video in the video discussion task. We did not control for if nods were present during these head turns by segmenting the head trajectory, but this could have combined or partially overlapped two kinds of movements with different meanings that usually occur on their own.

A head nod can also integrate with other modalities for it to change its meaning. For example, a backchannel may not necessarily be the same if I nod while speaking (Boholm & Allwood, 2010), smiling, or shifting gaze (Evinger et al., 1994). Hömke et al. (2017) also noted that long blinks often co-occurred with other modalities. Furthermore, the meanings may be more complex depending on the way that the non-verbal signal is produced, for example if it is a single nod or repeated, or with different magnitudes (Poggi et al., 2010).

The way people move their head when they speak is also an interesting confound because of the diversity of meanings that could be present. This is not the same as when speech co-occurs with a nod (Boholm & Allwood, 2010). While head nods may not be *caused* by speaking, a head nod may be a motoric consequence of speech production (McClave, 2000). Hadar et al. (1983) found that when speaking the head moved a lot compared to when someone was listening, and that there was a correlation between head movements and verbalisations. In addition to this, they also reported that head movements play an important role in putting emphasis on important parts of the speech utterance. However, because the experiments in this thesis were designed to test why *listeners* use fast nodding backchannels or slow nodding mimicry rather than *speakers*, this may not be a problem. However, this is a

confound to always look out for when it comes to analysing listener-speaker roles and turn-taking behaviour (Duncan & Fiske, 1977), especially in unstructured conversations where the turns happen quicker. Patterns of head movement in both the speaker and listener have been found in the turn-taking process. For example, Duncan (1972) observed that speakers turn away their head at the start of an utterance while they turn to their interaction partner to hand over the turn.

In this thesis, we are not presenting any data on turn-taking behaviour (except from three graphs to highlight the difference between the structured and unstructured conversations in Figure 2-1). We have, however, collected data on listener and speaker turns in the speech part of our dataset, where each participant's voice was recorded on two separate audio channels with machine-specific timestamps. Thus, turn-taking behaviour might be an interesting direction to take in future studies. For example, we can revisit the corpus data to examine speaking-listening roles in relation to speech. Furthermore, being able to examine turn-taking within the contexts of the different tasks would be interesting, seeing as the failure to follow the rules on how to decide who is speaking can also be considered a social signal. For instance, interrupting someone else's turn may signal aggressiveness and dominance, whereas turn overlapping may signal competitiveness, or signal to people outside the conversation that it is becoming conflictual (Pesarin et al., 2011).

Memory and Learning. Based on the results from investigating fast nodding in different contexts, we showed that it is represented more in conversational contexts in which there was a transfer of new information between participants. From these results, we can ask the question if fast head nods could be a backchannel signal from the listener in the conversation to inform the speaker that they have received

this information? And if so, might fast nods be seen more on trials when participants are learning more, which means fast nods might correlate with later memory scores.

In Chapter 3, we used a new memory task to test and quantify this relation by trying to link the cognitive process (e.g., memory) to a particular head nodding pattern (e.g., fast nodding). From this we then can make correlational predictions on the outcome in terms of how much the participants remember from the information that was discussed. This exploratory study provided initial hints that there might be a relationship between head nodding behaviour and performance on a later memory test, though further analyses were less clear. It is important to emphasize that this hypothesis is based around correlating measures between the head nodding behaviour and a conversational outcome. There is not much research related to head nodding and 'item memory' as a learning outcome. However, several studies have tried to link non-verbal behaviours in conversation to a range of learning outcomes (Chen et al., 2015; Pinzon-Gonzalez & Barba-Guaman, 2021; Sümer et al., 2021). In these studies, non-verbal behaviour is taken as an index of 'attention' or 'listening' (in contrast to boredom or mind wandering), and thus is expected to predict learning outcomes (i.e., memory). Seeing as head nodding is sensitive to conversational demands, such as signalling attention and understanding (Hadar et al., 1983), this links to our interpretation of fast nods as a potential backchannel that signals listening or understanding (i.e., comprehension). Backchannels to indicate comprehension are a fundamental component of communication between people (Kendon, 2002). For example, Richardson and Dale (2005) showed that the more coordinated a listener's eye gaze was with the speaker in a conversation, the better the listener did on a comprehension test. However, they did not investigate if this was the case for speakers as well, or in unstructured conversations. Furthermore,

most research on learning examines isolated participants in front of computer screens, or through observational learning. However, learning new information often occurs in social contexts, and in this thesis, we aimed to improve on this and include a behavioral measure of acquired information and factual knowledge (i.e., 'item memory') taken from a real-world dyadic interaction to gain a better understanding of the natural parameters of fast head nodding and its relation to memory. This can be especially important for real-world settings such as education or psychotherapy.

An implication of measuring memory with our task is that participants are probably being more attentive while reading a fact by having access to visual memory and seeing the words written on the cards. This may create ceiling effects where participants remember all facts correctly. Also, relating back to our example of the two poker players, attention can also be easy to fake, and our participants can easily have signaled fake engagement just to get the task done. Thus, more research is still needed on the relation between head nodding behaviour and memory as a conversational outcome.

Virtual Backchannels. In Chapter 4, we built on the preliminary result of Chapter 3 and aimed to test if there is a causal link between interactive engagement and memory performance. To test this hypothesis, we created two virtual agents who could show our head nodding behaviour rules and test how much participants remember from a conversation with the agent (See Section 6.2.3 for more details on the limits of these agents). The results from this study show that there is no significant effect of agent-interactivity between the interactive and non-interactive agents, which demonstrate that there is no causal link between interactive engagement and memory. In the study by Hömke et al. (2018), they developed a similar experimental paradigm using virtual agents to selectively manipulate blink

duration in a virtual listener to test how participants would react to the visual feedback, or backchanneling, of either a short blink (~200 ms) or a long blink (~600 ms). However, the researchers implemented the behaviour rules as a WoZ method, where the experimenter controlled the timing of the inputs to the virtual character. We improved upon this method and instead programmed the fast nodding backchannel behaviour, along with all our other behaviours, into the virtual agents to be conditional on certain behaviours from the participants. The only exception in our study was speech, which similar to what Hömke et al. (2018) did with blinks, was implemented as a WoZ method. However, since blinking was the variable that they manipulated in their study, and speech was only used to increase the ecological validity in our experiment, we believe this design choice may improve future studies on interpersonal coordination using virtual agents.

However, when we manipulate a behaviour rule, either by programming it into the agents or using a WoZ method, and use this to test psychological hypotheses, we must be careful with how we build the hypotheses so that we are clear that what we are measuring relates to the participant and not the agents. This is crucial in terms of what conclusions we can make. For example, in our experiment, it is likely to be the participant who is performing the fast nodding that have learned something from the interaction. Since we cannot manipulate head nodding backchannels directly in participants, it is important to highlight that our manipulation of fast nodding backchannels is an indirect manipulation of the visual signals that the participant is getting 'back' from seeing the head nod. This in turn changes the interactive characteristics of the conversation and how the participants respond (e.g., nodding). This is similar to studies examining the virtual mimicry of 'being mimicked' rather than measuring the participant who is doing the 'mimicking' (Hale & Hamilton,

2016b). As such, in our experiment on memory and learning, we set out with the aim to investigate a causal link between memory and fast nodding under the assumption that this was an indirect manipulation of the interactivity of the virtual agent, and not a direct manipulation of the participants head nodding.

While our results fail to support this hypothesis, we do however observe that participants that are speaking, or reading, a fact are up to three times more likely (i.e., odds ratio) to remember information from the interactive agent. This is likely a good indication that these results replicate the SRE with virtual agents, but this effect can still come from being more attentive and having access to visual memory, which allows for more opportunities to make associations to new information.

In Chapter 5, we analyzed data from Chapters 3 and 4 together with new data from a video-based conversation task, where the aim was to summarize how the level of interactivity in different contexts (i.e., real, virtual, video) impacted on the memory performance of the participants. Previous studies had not systematically examined the level of interactivity as a contributing factor in learning, and this thesis had examined memory and learning in the context of real-world interactions and in a virtual reality setting. We found that the level of interactivity significantly changed memory and learning during conversations, and we confirmed the existence of a 'video deficit' in both real and virtual interactions. This is consistent with previous studies showing that real social interaction is important for memory and learning (Anderson & Pempek, 2005; Krcmar et al., 2007; Roseberry et al., 2009), and that both children (Kuhl et al., 2003) and adults (de Felice et al., 2021) learn better from real interactions compared to when they learn from video. These results further support the conclusion that interaction-based learning is more effective than

observational learning due to a greater level of interactivity with natural verbal and non-verbal social signals in real and virtual interactions.

Surprisingly, we also found that we remember more when engaged with interactive virtual agents compared to real interactions when listening. This effect may be attributed to the added social pressure of having another person physically present during real-world interactions, which can lead to increased attention and learning. On the other hand, real-world social presence can also lead to more distractions compared to virtual interaction and make it more difficult to focus on the task. Along these lines, there is also the possibility that the virtual agents can be perceived as awkward in their behaviour. This can make the participants work harder on the interaction with the virtual agents to make the conversation work without being awkward, which means that they will carry more of the load and remember more from the conversation.

By design, the video-based task did not have a dialogue section, and apart from this being a non-interactive condition of observational learning, the addition of a dialogue section in the real and virtual interactions may have influenced the participants to remind each other of some of the information involved in certain facts, and consequently created some common ground that facilitated the memory of those facts. However, this study improved on the video-based task by using a matched design aimed to increase the ecological validity and allow the participants to feel like they are performing the task and exchanging information with someone via video link. In other words, compared to for example video lectures where the learner is passively acquiring information, the learner in this video-based task is decoding an interaction that feels present and engaging, but still lacks a real interactive element.

Understanding how learning is affected by social interaction is important for education and this kind of research can help in artificially generating virtual agents who can teach participants new information and act as a virtual tutor. The task on American states facts is well-suited to learning studies of this kind because it includes a set of uncommon facts that can be tested after the experiment as unintentional learning. There has been a growing body of research aimed at analysing social signals with the goal of building tools, applications, and interfaces for humans, based on models of human behaviour (Burgoon et al., 2017; Vinciarelli et al., 2009). This study contributes to this research by showing that learning can vary depending on the level of interactivity of tools for online learning that can implement the benefits of real social interactions in online educational settings and video conferencing using virtual agents over video teaching. However, we recognize that this could be driven by a series of cognitive processes (e.g., attention, cognitive load, backchanneling etc.) that may be absent in the non-interactive video task.

6.3.2 Slow Nodding Coherence and Behavioural Mimicry

Like the high frequency fast nodding pattern found by Hale et al., (2020), we were also able to replicate their low frequency slow nodding pattern and show that this signal carried different social meaning to fast nodding. Head movements occurring at <1Hz have sometimes been linked to behavioural mimicry, or the chameleon effect (Chartrand & Bargh, 1999; Stel, van Dijk, & Oliver, 2009). Hale et al. (2020) were able to use wavelet phase measures to determine the precise parameters and timing of this slow nodding mimicry behaviour. They reported that this nodding behaviour is generated by a mechanism with a 600 ms time lag between the speaker and listener

in a dyad. This is consistent with a reactive mimicry mechanism in which one participant sees the other's head nod and then responds to it (Heyes, 2011).

Context and Meaning. In Chapter 2, we aimed to build on the work of Hale et al. (2020) to explore mimicry across contexts and found that it changes based on the conversational context. We considered two possible theories for contextual effects on slow nodding. First, the 'Social Glue Hypothesis' states that low frequency slow nodding, or mimicry, is closely related to social bonding and the desire to get on well with others (Lakin et al., 2003), and should therefore create the same motivation to form a social bond across all three conversational contexts. Second, the 'gaze following' hypothesis suggests that slow nodding reflects the fact that people follow the gaze of their partner only in contexts where there is a potential gaze target (e.g., where one person is holding a picture, and not in other contexts). Some studies which score mimicry behaviour on observation of interpersonal coordination may not distinguish between gaze following and mimicry (Salazar-Kämpf et al., 2017). However, we suggest that it can be useful to make this distinction, because the two actions could have different social meanings. For example, in the raw motion capture data, the joint attention or gaze following pattern might look like a nodding action, which means it is important to consider the location of potential gaze targets when interpreting nodding behaviour. If we interpret this slow nodding in terms of joint attention, we suggest that the picture provided a gaze target for one participant who would alternate gaze between their partner's face and the picture in their hands (i.e., an up-down head movement) while the partner shared their attentional state and thus copied their head movements with a delay, leading to coordinated slow nodding. If this interpretation is correct, then conversations in a different context without the picture should not show coordinated slow nodding behaviour.

Individual Differences in Nodding Behaviour. Collecting nodding data across contexts allowed us to explore if the behaviour of individual participants is consistent from one context to another, that is, do some people always engage in a lot of nodding regardless of context while other rarely nod? We were particularly interested to test if there are robust individual differences in slow nodding behaviour since the individual differences underlying mimicry remain largely unexplored. Previous studies have shown that although people have a general tendency to mimic each other (Chartrand & Bargh, 1999), certain features of the individuals involved influence how much they mimic each other (Chartrand & Lakin, 2013; van Baaren et al., 2009). In Experiment 3 from Chartrand and Bargh (1999) they also showed that empathic individuals exhibit mimicry to a greater extent than do other people (i.e., a social chameleon). It has also been suggested that some people show more spontaneous mimicry than others and that this is related to increased liking (Salazar-Kämpf et al., 2017). There is not much data provided to quantify this, and we believe this is important to explore if we are to use nodding measures in clinical assessments. For example, if the amount of nodding someone engages in is to be used as a clinical measure, it should be robust across different conversational contexts as well as consistent within an individual.

In Chapter 2, we tested this and found evidence to suggest that slow nodding mimicry (and fast nodding) is driven by context and show no reliable individual differences, which suggests that head nodding measures may have limited clinical validity. If we would have found individual differences, this would have motivated us to test if the tendency to nod reflects broader social skills in future studies. However, these results can be important because there have been recent attempts to use automated analyses of interactive behaviour to identify and diagnose disorders of

social interaction such as autism (Georgescu et al. 2019). The results could also support the development of automated methods that could discriminate personality.

A limitation with our study is that each person only appears in one dyad, so we were not able to analyse each person's behaviour independently of their conversation partner as (Salazar-Kämpf et al., 2017) did in their study. Then there are the typical limitations regarding data from subjective questionnaire measures in dyadic studies which have characteristics that are dependent on the way people coordinate and adapt their behaviour over time. This presents a challenge for researchers who want to understand individual differences, and future studies should design such questionnaires to be sensitive to these kinds of interactional parameters. For a subjective measure, the sample size in our study is relatively small (n=62). However, the slow nodding pattern that we have identified presents us with a way of disentangling the meaning behind the behaviour by studying it in different contexts. This enables a more comprehensive understanding of how mimicry is related to other conversational outcomes.

Liking and Self-Other Overlap. Based on the results from investigating slow nodding in different contexts, we showed that it changes across different conversational contexts in favour of being a form of joint attention rather than a mimicry behaviour. From these results, seeing as both joint attention and mimicry can be interpreted as a form of social glue that make people feel closer to each other, we were interested to explore the potential relationship between a larger self-other overlap in terms of how close we feel to the other person, and mimicry. In Chapter 3, we tested this prediction with a source memory test to discover how biased the participants are towards themselves compared to others.

The ‘social glue hypothesis’, which predicts that coordination should be related to liking and affiliation, gives us an explanation for *why* mimicry may be an effective means of feeling closer to someone, but less research has focused on explanations of *how* such consequences of mimicry might occur. Ashton-James et al. (2007) proposed that mimicry helps increase the interdependence of one’s self-construal (i.e., more ‘other’ focused), which lead to a larger self-other overlap between people and more positive social outcomes. Other researchers have joined in to test this idea on participants that are either mimicking or being mimicked (Hale & Hamilton, 2016a; Hove & Risen, 2009; Stel et al., 2011; Wiltermuth & Heath, 2009). However, these studies either use confederates to mimic participants, use low-resolution methods or subjective measures. Moreover, few of these studies test how we perceive ourselves during a conversation – that is, how mimicking or being mimicked can affect our self-construal, or our sense of identity relative to the other person.

In Chapter 3 we improved on these aspects by using (1) real-world dyadic conversations (2) high resolution methods, and (3) behavioural outcome measures related to two known memory effects (i.e., SRE and SBE). Since learning new information often occurs in social contexts, using these two memory effects as outcome measures in a dyadic interaction allowed us to link the recollection of the facts to real episodes or experienced events taken from the conversation instead of the facts being merely a semantic recollection.

Our results first show that the participants remember more when speaking compared to listening, which indicates that they are more likely to remember information that is linked to themselves. This would effectively replicate the SRE for memory and is consistent with previous studies that find support for the SRE (Cunningham et al., 2008; Klein, 2012; Macrae et a., 2004; Maki & McCaul, 1985;

Powell et al., 2010; Rogers et al., 1977; Symons & Johnson, 1997). Secondly, our results unexpectedly show that when prompted to remember *who* read a fact, our participants were more likely to claim "the other said it" even if they did not. In other words, the participants may have a more interdependent self-construal (i.e., 'other' focused) and identify more with the 'Other' (Brewer & Gardner, 1996). This result is the opposite of what Russel and Jarrold (1999) found in their study, in which their child participants showed a significant memory bias toward the 'Self'. The difference between ours and their study is that our participants were adults involved in unstructured dyadic conversations, which might affect their self-construal in a way that is more interactive and interdependent, leading to a feeling of closeness to others (Holland et al., 2004) or more self-other overlap.

These results can prove to be useful when exploring the relationship between mimicry and self-other overlap. We initially intended to perform a similar analysis to what we did with the fast nodding behaviour. However, we encountered problems with uneven numbers of trials since the participants could only give a self-other response if they remembered a fact and this forced us to calculate a separate index with fewer responses. This design is something that can be improved upon in future studies. Overall, the mechanisms linking mimicry and self-other overlap are still debated and in need of further research.

Virtual Mimicry. In Chapter 4, we aimed to test if interacting with interactive virtual agents, driven by natural mimicry behaviour, enhance feelings of rapport compared to non-interactive agents, and if we learn more from agents that we like. In other words, can slow nodding mimicry be used to change how much we like and learn from a virtual agent? To test this hypothesis, we created two virtual agents who could show our head nodding behaviour rules and test how much participants liked

and learned from the agents (See Section 6.2.3 for more details on the limits of the agents). First, the results related to liking the agents showed that there is no significant increase in feelings of rapport when interacting with the interactive agent. Secondly, the results related to learning from the agents show that there is no reliable positive link between agent-interactivity and memory performance.

These results are consistent with previous research showing mixed results for the effect of virtual mimicry on rapport (Hale & Hamilton, 2016b; Verberne et al., 2013), and traditional research settings with human confederates (Chartrand & Bargh, 1999; van Baaren et al., 2004). This casts doubt over a strong version of the 'social glue hypothesis'. Bailenson and Yee (2005) who first demonstrated the positive effects on participants' impressions of being mimicked in virtual reality, report that participants who were mimicked rated the agent as more likeable. However, the researchers of this study did not provide any explanation on how their mimicry ratings were weighted. Furthermore, both effect sizes and experimental power in many previous studies have been small, and false positives may be present (Hale & Hamilton, 2016a). Our results provide further support to the conclusion that we should use caution in accepting the social glue hypothesis in a virtual mimicry setting since it is difficult to replicate.

The advantage with using virtual agents to test mimicry is that we get high experimental control over manipulating the precise time lags of the nods, as well as the benefit of making the nod conditional on the head nodding behaviour of the participants wearing the HMDs. This can also lead to the disadvantage that, given the importance of timing in dynamic social interactions, even a minor gap in the analysis of the signal or when generating the code to implement the behaviour rule in the virtual agent can prevent the correct signals to be sent. For example, an

incorrectly sent signal can be interpreted by the receiving participants as social ignorance or incompetence (Vinciarelli et al., 2009). In two influential studies (Bailenson & Yee, 2005; Gratch, Wang, Gerten, Fast, & Duffy, 2007) the researchers specifically addressed this issue of movement dynamics and the timing of non-verbal feedback for creating a feeling of rapport between participants and virtual agents. This issue makes the combination of high resolution motion capture and virtual reality especially effective since it produces the most realistic representations of movements (i.e., high ecological validity).

Behaviour rules are an excellent starting point for the study of social interaction but might still be too simple to account for the richness of human social behaviour. Virtual agents that are governed only by simple behaviour rules will at some point begin to diverge from human behaviour. Thus, a critical question is when a behaviour rule breaks down. Virtual mimicry provides a good experimental setting for testing when a behaviour rule breaks down since researchers can systematically manipulate the time lag (e.g., 600 ms to 1200 ms) and observe when participant responses change. Future studies could also try to time virtual mimicry to the appropriate points of a participant's speech to examine its temporal relationship.

6.4 Concluding Remarks

This kind of data-driven approach to *measuring*, *analysing*, and *artificially generating* social interaction, in combination with the development of well-specified theoretical questions, will continue to require strong interdisciplinary collaboration. The concluding remarks, or the argument we like to make, is not that studying isolated behaviours is an invalid approach because of the absence of contextual and dynamic factors, but rather that we should use these findings to build more complex models that are able to handle dynamic multimodal interactions within the realm of two-person neuroscience. The current challenge is to come full circle and bring social neuroscience “out of the laboratory” to replicate complex and dynamic real-world social interaction. Such an approach of the ‘big-data’ of social coordination will be critical in creating a new understanding of our everyday social behaviour and the mechanisms that support it.

References

Abney, D., Paxton, A., Dale, R., & Kello, C. T. (2015). Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. *Cognitive Processing*, *16*(4). doi: 10.1007/s10339-015-0648-2.

Aburumman, N., Gillies, M., Ward, J. A., & Hamilton, A. F. de C. (2022). Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. *International Journal of Human-Computer Studies*, *164*, 102819.

Allwood, J., & Cerrato, L. (2003). A study of gestural feedback expressions. In Paggio et al. (Eds.), *Proceedings of the First Nordic Symposium on Multimodal Communication*, Copenhagen.

Anderson, D. R. & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505-522.

Argyle, M. (2013). *Bodily Communication*. Routledge.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. New York: Cambridge University Press.

Argyle, M., & Trower, P. (1979). *Person to person: Ways of communicating*. New York: HarperCollins.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology, 63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality and Social Psychology, 60*(2), 241–253. https://doi.org/10.1037/0022-3514.60.2.241

Ashenfelter, K. T., Boker, S. M., Waddell, J. R., & Vitanov, N. (2009). Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. *Journal of Experimental Psychology*, *35*(4), pp. 1072–91.

Ashton-James, C. E., van Baaren, R. B., Chartrand, T. L., Decety, J., & Karremans, J. (2007). Mimicry and Me: The Impact of Mimicry on Self–Construal. *Social Cognition*, *25*(4), 518–535. http://doi.org/10.1521/soco.2007.25.4.518

Aubrey, A. J., Marshall, D., Rosin, P. L., Vandeventer, J., Cunningham, D. W., & Wallraven, C. (2013). Cardiff conversation database (CCDb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Portland, Oregon, USA, 277–82.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal Distance in Immersive Virtual Environments. *Personality and Social Psychology Bulletin*, *29*(7), 819–833. https://doi.org/10.1177/0146167203029007002

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, *16*(10), 814–819. doi.org/10.1111/j.1467-9280.2005.01619.x

Baltrusaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), (IEEE), pp. 1–10.

Baptista, L. F., & Petrinovich, L. (1986). Song development in the white-crowned sparrow: social factors and sex differences. *Animal Behaviour*, *34*(5), 1359-1371.

Bargh, J. A. (1982). Attention and automaticity in the processing of self-relevant information. *Journal of Personality and Social Psychology, 43*(3), 425–436. https://doi.org/10.1037/0022-3514.43.3.425

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of social interaction. In E. Higgins & R. Sorrentino (Eds.), *Handbook of Motivation and Cognition* (Vol. 2, pp. 93–130). New York: Guilford Press.

Bavelas, J. B., & Chovil, N. (1997). *Faces in dialogue.* In J. A. Russell & J. M. Fernández-Dols (Eds.), *Studies In Emotion and Social Interaction, 2nd Series. The Psychology of Facial Expression* (pp. 334–346). Cambridge University Press.

Bavelas, J., & Healing, S. (2013). Reconciling the effects of mutual visibility on gesturing: A review. *Gesture*, *13*(1), 63–92. https://doi.org/10.1075/gest.13.1.03bav

Bekkering, H., & Prinz, W. (2002). Goal representations in imitative actions. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation In Animals and Artifacts* (pp. 555–572). Cambridge: MIT Press.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology Section A*, *53*(1), 153–164. https://doi.org/10.1080/713755872

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, *57*(1), 289–300. doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bente, G., Krämer, N., Petersen, A., & De Ruiter, J. (2001). Computer animated movement and person perception: Methodological advances in nonverbal behavior research. *Journal of Nonverbal Behavior*, *25*, 151-166. 10.1023/A:1010690525717.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of nonverbal behavior,* 12 (2), 120–138.

http://dx.doi.org/10.1007/BF00986930

Bernieri, F. J., Gillis, J. S., Davis, J. M., & Grahe, J. E. (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, *71*(1), 110–129. https://doi.org/10.1037/0022-3514.71.1.110

Bernieri, F. J., & Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In R. S. Feldman & B. Rimé (Eds.), *Studies In Emotion & Social Interaction. Fundamentals of Nonverbal Behavior* (pp. 401–432). New York: Cambridge University Press.

Bernieri, F. J., Steven, J., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology*, *54*(2), 243–253. doi.org/10.1037/0022-3514.54.2.243

Birdwhistell, R. L. (1970). *Kinesics and Context: Essays on Body-Motion Communication*. Philadelphia, University of Pennsylvania Press.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, *13*, 103–124. doi.org/10.1207/S15327965PLI1302_01

Boholm, M., & Allwood, J. (2010). Repeated head movements, their function and relation to speech. In. Kipp et al. (Eds.), *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010.

Boker, S. M., Xu, M., Rotondo, J. L., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*. *7*(3), 338-355.

Bouaziz, S., Wang, Y., & Pauly, M. (2013). Online modeling for realtime facial animation. *Proceedings of SIGGRAPH 2013 – ACM Transactions on Graphics*, *32*(4). doi.org/10.1145/2461912.2461976

Bourgeois, P., & Hess, U. (2008). The impact of social context on mimicry. *Biological Psychology*, *77*(3), 343–352. https://doi.org/10.1016/j.biopsycho.2007.11.008

Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2359–2367. https://doi.org/10.1098/rstb.2009.0066

Brewer, M. B., & Gardner, W. (1996). Who is this "We"? Levels of collective identity and self representations. *Journal of Personality and Social Psychology, 71*(1), 83–93. https://doi.org/10.1037/0022-3514.71.1.83

Burgoon, J. K., Bonito, J. A., Ramirez, A., Dunbar, N., Kam, K., & Fischer, J. (2002). Testing the Interactivity Principle: Effects of Mediation, Propinquity, and Verbal and Nonverbal Modalities in Interpersonal Interaction. *Journal of Communication*, *52*(3), 657-677. doi:10.1093/joc/52.3.657.

Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., & Vinciarelli, A. (Eds.). (2017). *Social signal processing*. Cambridge University Press.

Cappella, J. N. (1981). Mutual influence in expressive behaviour: Adult-adult and infant-adult dyadic interaction. *Psychological Bulletin*, *89*(1), 101–132. doi.org/10.1037/0033-2909.89.1.101

Carletta, J. (2007). Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, *41*, 181-190. https://doi.org/10.1007/s10579-007-9040-x

Cassell, J., & Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence, 13*(4-5), 519-538. doi:10.1080/088395199117360

Catmur, C., Gillmeister, H., Bird, G., Liepelt, R., Brass, M., & Heyes, C. (2008). Through the looking glass: counter-mirror activation following incompatible sensorimotor learning. *The European Journal of Neuroscience*, *28*(6), 1208–1215. https://doi.org/10.1111/j.1460-9568.2008.06419.x

Cerrato, L. (2005). Linguistic functions of head nods. In. J. Allwood & B. Dorriots, (Eds.), Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication, Gothenburg University, Sweden.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. doi.org/10.1037//0022-35er14.76.6.893

Chartrand, T. L., & Lakin, J. L. (2013). The Antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*(1), 285–308. https://doi.org/10.1146/annurev-psych-113011-143754

Chartrand, T. L., & van Baaren, R. (2009). Chapter 5 Human Mimicry. In M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. Volume 41, pp. 219–274). Academic Press.

Chen, C. M., & Huang, S. (2014). Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance. *British Journal of Educational Technology*, *45*, 959-980.

Chen, C. M., Wang, J. Y., & Yu, C. M. (2017). Novel attention aware system based on brainwave signals. *British Journal of Educational Technology*, *48*: 348-369. https://doi.org/10.1111/bjet.12359

Cheng, C. M., & Chartrand, T. L. (2003). Self-monitoring without awareness: Using mimicry as a nonconscious affiliation strategy. *Journal of Personality*, *85*(6), 1170–1179.

Chovil, N. (1991). Discourse oriented facial displays in conversation. *Research on Language and Social Interaction*, *25*, 163–194. doi.org/10.1080/08351819109389361

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, H. H. (1996). Using Language. Cambridge, UK: University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). *American Psychological Association*. https://doi.org/10.1037/10096-006

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*(1), 62–81. https://doi.org/10.1016/j.jml.2003.08.004

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time-series: An R package. *Frontiers in Psychology*, *5*. 510. doi: 10.3389/fpsyg.2014.00510

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science, 210*(4466), 207–210. https://doi.org/10.1126/science.7414331

Condon, W. S., & Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease*, *143*(4), 338–347.

Condon, W. S., & Ogston, W. D. (1971). Speech and body motion synchrony of the speaker-hearer. In D. Horton & J. Jenkins, (Eds.), *The Perception of Language*, pp. 150–84. Columbus, OH:Charles E. Merrill.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*, 594–628.

Couzin I, D. (2009). Collective cognition in animal groups. *Trends in Cognitive Sciences*, *13*(1), 36-43. doi: 10.1016/j.tics.2008.10.002.

Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: Applying neurocognitive insights to human-robot interaction. Phil. Trans. R. Soc, *374*: 20180024. http://doi.org/10.1098/rstb.2018.0024

Cross, L., Turgeon, M., & Atherton, G. (2019). How moving together binds us together: The social consequences of interpersonal entrainment and group processes, *Open Psychology*, *1*(1), 273-302. https://doi.org/10.1515/psych-2018-0018

Cunningham, S. J., Turk, D. J., Macdonald, L. M. & Macrae, C. N. (2008). Yours or mine? Ownership and memory. *Consciousness and cognition*, *17*(1), 312-318.

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. C. (2013). The Self-Organization of human interaction. *Psychology of Learning and Motivation - Advances in*

Research and Theory, *59*, 43–95. https://doi.org/10.1016/B978-0-12-407187-2.00002-2

Davis, J., Chryssafidou, E., Zamora, J., Davies, D., Khan, K. & Coomarasamy, A. (2007). Computer-based teaching is as good as face to face lecture-based teaching of evidence based medicine: a randomised controlled trial. *BMC medical education*, *7*(1), 23.

De Felice, S., Vigliocco, G., & Hamilton, A.F.d.C. (2021). Social interaction is a catalyst for adult human learning in online contexts. *Current Biology*, *8;31*(21):4853-4859.e3. doi:10.1016/j.cub.2021.08.045

Demos, A. P., Chaffin, R., & Kant, V. (2014). Toward a dynamical theory of body movement in musical performance. *Frontiers in Psychology*, *5*, 477. doi: 10.3389/fpsyg.2014.00477

Dobre, G. C., Gillies, M., Falk, P., Ward, J., Hamilton, A., & Pan, X. (2021). Direct gaze triggers higher frequency of gaze change: An automatic analysis of dyads in unstructured conversation. In *Proceedings in International Conference on Multimodal Interaction (ICMI '21)*, Montreal, Canada, 735-739. 10.1145/3462244.3479962.

Dunbar, N. E., Jensen, M. L., Tower, D. C., & Burgoon, J. K. (2014). Synchronization of nonverbal behaviors in detecting mediated and non-mediated deception. *Journal of Nonverbal Behavior*, *38*(3), 355-376. https://doi.org/10.1007/s10919-014-0179-z

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283–292. https://doi.org/10.1037/h0033031

Duncan, S., & Fiske, D. (1977). *Face-to-Face Interaction: Research, Methods, and Theory*. London, Routledge.

Eales, L. A. (1989). The influences of visual and vocal interaction on song learning in zebra finches. *Animal Behaviour*, *37*(3), 507–508. https://doi.org/10.1016/0003-3472(89)90097-3

Ekman, P., & Friesen, W. V. (1972). Hand Movements. *Journal of Communication*, *22*(4), 353–374. https://doi.org/10.1111/j.1460-2466.1972.tb00163.x

Emery, N. J. (2000). The eyes have it: The neuroethology, function, and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, *24*, 581–604. doi.org/10.1016/S0149-7634(00)00025-7

Evinger, C., Manning, K.A., Pellegrini, J.J., Basso, M. A., Powers, A. S., & Sibony, P. A. (1994). Not looking while leaping: The linkage of blinking and saccadic gaze shifts. *Experimental Brain Research, 100***,** 337–344. https://doi.org/10.1007/BF00227203

Feese, S., Arnrich, B., Tröster, G., Meyer, B., & Jonas, K. (2011). Detecting posture mirroring in social interactions with wearable sensors. *Proceedings of the 15th International Symposium on Wearable Computers*, *USA*, *5959582*, 119–120. doi**:**10.1109/ISWC.2011.31

Forbes, P., Pan, X., & Hamilton, A. F. de C. (2016). Reduced mimicry to virtual reality avatars in autism spectrum disorder. Journal of Autism and Development Disorder, 46, 1-10. https://doi.org/10.1007/s10803-016-2930-2

Fujiwara, K., & Daibo, I. (2016). Evaluating interpersonal synchrony: Wavelet transform toward an unstructured conversation. *Frontiers in Psychology*, *7*, 516. doi.org/10.3389/fpsyg.2016.00516

Fusaroli, R., & Tylén, K. (2012). Carving language for Social Coordination: A dynamical approach. *Interactional Studies, 13*(1), 103-124. https://doi.org/10.1075/is.13.1.07fus

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Georgescu, A. L., Koehler, J. C., Weiske, J., Vogeley, K., Koutsouleris, N. & Falter-Wagner, C. (2019). Machine Learning to Study Social Interaction Difficulties in ASD. *Frontiers in Robotics and AI*, *6*, 132. doi.org/10.3389/frobt.2019.00132

Georgieff, N. & Jeannerod, M. (1998). Beyond consciousness of external reality: A "who" system for consciousness of action and self-consciousness. *Conscious Cognition, 7*(3), 465-77. doi:10.1006/ccog.1998.0367. PMID: 9787056

Gifford, C., & Wilkinson, M. (1985). Non-verbal cues in the employment interview: Links between application qualities and interviewer judgments. *Appl. Psychol. 70*(4). 729–736.

Giles, H., & Powesland, P. F. (1975). *Speech style and social evaluation*. Oxford, England: Academic Press.

Gillath, O., McCall, C., Shaver, P. R., & Blascovich, J. (2008). What can virtual reality teach us about prosocial tendencies in real and virtual environments? *Media Psychology*, *11*(2), 259-282. doi:10.1080/15213260801906489

Gillies, M. (2009). Learning finite-state machine controllers from motion capture data. *IEEE Transactions on Computational Intelligence and AI in Games*, *1*(1), 63–72. https://doi.org/10.1109/TCIAIG.2009.2019630

Gillies, M., Brenton, H., & Kleinsmith, A. (2015). Embodied design of full bodied interaction with virtual humans. *Proceedings of 2nd Int. Conference on Movement and Computing*, *Canada*.

Goldin-Meadow, S., & Alibali, M. (2012). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*, 10.1146/annurev-psych-113011-143802.

Goodwin, C. (1981). *Conversational Organization. Interaction Between Speakers and Hearers*. New York, Academic Press.

Gopie, N., & MacLeod, C. M. (2009). Destination memory: Stop me if i've told you this before. *Psychological Science*, *20*(12), 1492-1499. doi:10.1111/j.1467-9280.2009.02472.x

Grammer, K., Kruck, K. B., & Magnusson, M. S. (1998). The courtship dance: Patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Non-verbal Behaviour*, *22*(1), 3–29. doi.org/10.1023/A:1022986608835

Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating Rapport with Virtual Agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent Virtual Agents* (pp. 125–138). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-74997-4_12

Grèzes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: A meta-analysis. *Human Brain Mapping*, *12*(1), 1–19. https://doi.org/10.1002/1097-0193(200101)12:1<1::AID-HBM10>3.0.CO;2-V

Grinsted, A., Moore, J. C., & Jevrejeva, S. (2004). Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes in Geophysics*, *11*, 561–566. doi.org/10.5194/npg-11-561-2004

Guerin, B. (1986). Mere presence effects in humans: A review. *Journal of Experimental Social Psychology, 22*(1), 38–77. https://doi.org/10.1016/0022-1031(86)90040-5

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. *Human Movement Science*, *2*(1–2), 35–46. https://doi.org/10.1016/0167-9457(83)90004-0

Hadar, U., Steiner, T. J., & Rose, F. C. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, *9*(4), 214–228. https://doi.org/10.1007/BF00986881

Hadley, L.V., Naylor, G. & Hamilton, A.F.d.C. (2022). A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology, 1***,** 42–54. https://doi.org/10.1038/s44159-021-00008-w

Healey, P. G. T., Purver, M., & Howes, C. (2014) Divergence in dialogue. *PLoS ONE*, *9*(6): e98598. https://doi.org/10.1371/journal.pone.0098598

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hale, J., & Hamilton, A. F. de C. (2016a). Cognitive mechanisms for responding to mimicry from others. *Neuroscience & Biobehavioral Reviews*, *63*, 106–123. https://doi.org/10.1016/j.neubiorev.2016.02.006

Hale, J., & Hamilton, A. F. de C. (2016b). Get on my wavelength: The effects of prosocial priming on interpersonal coherence measured with high-resolution motion capture. *Open Science Framework*. Retrieved from osf.io/nex72

Hale, J., Ward, J. A., Buccheri, F., Oliver, D., & Hamilton, A. F. de C. (2020). Are you on my wavelength? Interpersonal coordination in naturalistic conversations. *Journal of Non-Verbal Behaviour, 44*, 63–83.

Hamilton, A. F. de C., Pinti, P., Paoletti, D., & Ward J. A. (2018). Seeing into the brain of an actor with mocap and fNIRS*. International Symposium on Wearable Computers (ISWC)*.

Hasson, U., & Frith, C. D. (2016). Mirroring and beyond: Coupled dynamics as a generalized framework for modelling social interactions. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 371(1693). B37120150366. https://doi.org/10.1098/rstb.2015.0366

Hedges, et al. (2013). Play, attention, and learning: How do play and timing shape the development of attention and influence classroom learning? *Annals of New York Academy of Sciences*, *1292*, 1-20. https://doi.org/10.1111/nyas.12154

Heerey, E. A. (2015). Decoding the dyad: Challenges in the study of individual differences in social behavior. *Current Directions in Psychological Science*, *24*(4), 285–291. doi.org/10.1177/0963721415570731

Heerey, E. A., & Kring, A. M. (2007). Interpersonal consequences of social anxiety. *Journal of Abnormal Psychology*, *116*(1), 125–134. https://doi.org/10.1037/0021-843X.116.1.125

Hertel, P. T., Brozovich, F., Joormann, J., & Gotlib, I. H. (2008). Biases in interpretation and memory in generalized social phobia. *Journal of Abnormal Psychology, 117*(2), 278–288. https://doi.org/10.1037/0021-843X.117.2.278

Hess, U., & Bourgeois, P. (2010). You smile-I smile: Emotion expression in social interaction. *Biological Psychology*, *84*(3), 514–520.

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, *137*(3), 463–483.

doi:10.1037/a0022288.

Heyes, C. (2012). What's social about social learning? *Journal of Computational Psychology, 126*(2):193-202. doi:10.1037/a0025180.

Heylen, D. (2006). Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics*, *03*(03), 241–267. https://doi.org/10.1142/S0219843606000746

Hoehl, S., Fairhurst, M., & Schirmer, A. (2020). Interactional synchrony: Signals, mechanisms and benefits. *Social Cognitive and Affective Neuroscience*, *16*(1-2), 5-18. doi: 10.1098/scan/nsaa024

Hogeveen, J., Chartrand, T., & Obhi, S. (2014). Social mimicry enhances Mu-suppression during action observation. *Cerebral Cortex*, *25*(8), 2076-2082. 10.1093/cercor/bhu016.

Holland, R. W., Roeder, U.-R., van Baaren, R. B., Brandt, A. C., & Hannover, B. (2004). Don't stand so close to me: The effects of self-construal on interpersonal closeness. *Psychological Science, 15*(4), 237–242. https://doi.org/10.1111/j.0956-7976.2004.00658.x

Hove, M. J., & Risen, J. L. (2009). It’s All in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, *27*(6), 949–960. https://doi.org/10.1521/soco.2009.27.6.949

Hox, J., Moerbeek, M., & van de Schoot, R. (2010). Multilevel Analysis: Techniques and Applications, Second Edition (2nd ed.). Routledge. https://doi.org/10.4324/9780203852279

Huang, L., Morency, L-P., & Gratch, J. (2010). Learning backchannel prediction model from parasocial consensus sampling: A subjective evaluation. In Allbeck, J., Badler, N., Bickmore., T., Pelachaud, C., & Safonova, A. (Eds.), *Lecture Notes*

in Computer Science: Vol. 6356. Intelligent Virtual Agents (pp. 159–172). doi.org/10.1007/978-3-642-15892-6_17

Huth, A., & Wissel, C. (1992). The simulation of fish schools in comparison with experimental data. *Ecological Modelling*, *75*, 135-146.

Hömke, P., Holler, J., & Levinson, S. (2017). Eye Blinking as addressee feedback in face-to-tace conversation. *Research on Language and Social Interaction*, *50*, 1-17. doi:10.1080/08351813.2017.1262143.

Hömke, P., Holler, J., & Levinson, S. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLOS ONE*, *13*, e0208030. 10.1371/journal.pone.0208030.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, *286*(5449), 2526–2528. https://doi.org/10.1126/science.286.5449.2526

Issartel, J., Bardainne, T., Gaillot, P., & Marin, L. (2015). The relevance of the cross-wavelet transform in the analysis of human interaction – a tutorial. *Frontiers in Psychology*, *5*(1566). doi.org/10.3389/fpsyg.2014.01566

Issartel, J., Marin, L., Gaillot, P., Bardainne, T., & Cadopi, M. (2006). A practical guide to time-frequency analysis in the study of human motor behaviour: The contribution of the wavelet transform. *Journal of Motor Behaviour*, *38*(2), 139–159. doi.org/10.3200/JMBR.38.2.139-159

Jain, A., Pecune, F., Matsuyama, Y., & Cassell, J. (2018). A user simulator architecture for socially-aware conversational agents. Proceedings of 18th AMC International Conference on Intelligent Virtual Agents. 10.1145/3267851.3267916.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. doi.org/10.1037/0033-2909.114.1.3

Judd, C. M., McClelland, G., & Ryan, C. S. (2008). *Data Analysis A Model Comparison Approach, Second Edition*. Routledge.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54-69.

Kajopoulos, J., Cheng, G., Kise, K., Müller, H. J., & Wykowska, A. (2021). Focusing on the face or getting distracted by social signals? The effect of distracting gestures on attentional focus in natural interaction. *Psychological Research*, *85*(2), 491–502. https://doi.org/10.1007/s00426-020-01383-4

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*(1), 22–63. doi.org/10.1016/0001-6918(67)90005-4

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica*, *32*, 101–125. doi.org/10.1016/0001-6918(70)90094-6

Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters.* Cambridge University Press.

Kendon, A. (2002). Some uses of head shake. *Gesture*, *2*(2). 147–182. https://doi.org/10.1075/gest.2.2.03ken

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511807572

Klein, S. B. (2012). Self, memory, and the self-reference effect: An examination of conceptual and methodological issues. *Personality and Social Psychology Review*, *16*(3), 283-300. doi:10.1177/1088868311434214.

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, *4*(1), 15–33.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press.

Konvalinka, I., & Roepstorff, A. (2012). The two-brain approach: how can mutually interacting brains teach us something about social interaction? *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00215

Konvalinka, I., Vuurst, P., Roepstorff, A., & Frith, C. D. (2010). Follow you, follow me: Continuous mutual prediction and adaptation in joint tapping. *Quarterly Journal of Experimental Psychology*, *63*(11), 2220–2230. doi:10.1080/17470218.2010.497843

Kopp, S., & Bergmann, K. (2013). Automatic and strategic alignment of co-verbal gestures in dialogue. In K. Dautenhahn., A. & Cangelosi (Eds.), *Advances in Interaction Studies: Vol. 6. Alignment in Communication: Towards a New Theory of Communication* (pp. 87–107). John Benjamins Publishing Company.

Kourtis, D., Jacob, P., Sebanz, N., Sperber, D., & Knoblich, G. (2020). Making sense of human interaction benefits from communicative cues. *Scientific Reports*, *10*(1):18135. doi:10.1038/s41598-020-75283-3.

Koutsombogera, M., & Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, L18-1466.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behaviour: Correcting a reductionist bias. *Neuron*, *93*(3), 480–490. doi.org/10.1016/j.neuron.2016.12.041

Krcmar, M., Grela, B. & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, *10*(1), 41-63.

Krithika, L. B., & Priya, G. G. (2016). Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. *Procedia Computer Science*, *85*, 767-776. 10.1016/j.procs.2016.05.264.

Kuhl, P. K., Tsao, F. M. and Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, *100*(15), 9096-9101.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, *20*(1), 54–72. https://doi.org/10.3758/s13423-012-0341-8

Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339. https://doi.org/10.1111/1467-9280.14481

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behaviour*, *27*(3), 145–162. doi.org/10.1037/e413812005-152

Laland, K.N., & Rendell, L. (2019). Social learning: Theory. In Choe, J.C. (Eds.), *Encyclopedia of Animal Behavior*, 380–386. Elsevier.

Liebowitz, M. R. (1987). Social phobia. In T. A. Ban., P. Pichot., & W. Pöldinger (Eds.), *Modern Problems of Pharmacopsychiatry*: *Vol. 22. Anxiety* (pp. 141–173). doi*:*10.1159/000414022

Liepelt, R., Cramon, D. Y. V., & Brass, M. (2008). What is matched in direct matching? Intention attribution modulates motor priming. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 578–591. https://doi.org/10.1037/0096-1523.34.3.578

Liu, D., Sun, P., Xiao, Y., & Yin, Y. (2010). Drowsiness detection based on eyelid movement. Education Technology and Computer Science, Second International Workshop IEEE, 2, 49-52.

Louwerse, M., Dale, R., Bard, E., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*, 10.1111/j.1551-6709.2012.01269.x.

Lytle, S. R., Garcia-Sierra, A., & Kuhl, P. K. (2018). Two are better than one: Infant language learning from video improves in the presence of peers. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9859–9866. https://doi.org/10.1073/pnas.1611621115

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren & C. Lewis (Eds.). *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 21–57). Hillsdale, NJ: Erlbaum

Macrae, C. N., Moran, J. M., Heatheron, T. F., Banfield, J. F., & Kelley, W. M. (2004). Medial prefrontal activity predicts memory for self. *Cerebral Cortex*, *14*, 647–654

Maki, R. H., & McCaul, K. D. (1985). The effects of self-reference versus other reference on the recall of traits and nouns. *Bulletin of the Psychonomic Society*, *23*(3), 169-172.

Maister, L., & Tsakiris, M. (2016). Intimate imitation: Automatic motor imitation in romantic relationships. *Cognition*, *152*, 108-113. doi:10.1016/j.cognition.2016.03.018.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, *1*(2), 320–339. https://doi.org/10.1111/j.1756-8765.2009.01022.x

Marwan, N. (2008). A historical review of recurrence plots. *The European Physical Journal*, *164*(1), 3-12.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. Physics Reports, 438(5), 237-329.

Maynard-Smith, J., & Harper, D. G., (2003). Animal signals. Oxford: Oxford University Press.

McCall, C., & Singer, T. (2015). Facing off with unfair others: Introducing proxemic imaging as an implicit measure of approach and avoidance during social interaction. *PLOS ONE*, *10*(2), e0117532. https://doi.org/10.1371/journal.pone.0117532

McCarthy, M. (2003). Talking back: "Small" interactional response tokens in everyday conversation. *Research on Language and Social Interaction*, *36*, 33-63. 10.1207/S15327973RLSI3601_3.

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878. https://doi.org/10.1016/S0378-2166(99)00079-X

McGovern, T., Jones, B., & Morris, S. (1979). Comparison of professional versus student ratings of job interviewee behaviour. *J. Couns. Psychol. 26*(2), 176–179.

Meltzhoff, A. N., Kuhl, P. K., Movellan, J., & Sejnowski, T. J. (2009). Foundations for a new science of learning. *Science, 325*(5938)*, 284-288.*

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (2013). Machine learning: An artificial intelligence approach. *Symbolic Computing*, Vol. 2.

Morency, L. P., de Kok, I., & Gratch, J. (2009). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems, 20***,** 70–84. https://doi.org/10.1007/s10458-009-9092-y

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, *19*(2), pp. 98-100. doi:10.1109/MRA.2012.2192811.

Marotta, A., Lupiáñez, J., Martella, D., & Casagrande, M. (2012). Eye gaze versus arrows as spatial cues: Two qualitatively different modes of attentional selection. *Journal of Experimental Psychology: Human Perception and Performance, 38*(2), 326–335. https://doi.org/10.1037/a0023959

Morlet, J. (1983). *Sampling and wave propagation*. New York, NY: Springer.

Morrison, A. B., Conway, A. R. A., & Chein, J. M. (2014). Primacy and recency effects as indices of the focus of attention. *Frontiers in Human Neuroscience, 8, 1662-5161.* doi: 10.3389/fnhum.2014.00006

Neumann, R., & Strack, F. (2000). “Mood contagion”: The automatic transfer of mood between persons. *Journal of Personality and Social Psychology*, *79*(2), 211–223.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Ochs, E., Schegloff, E. A., & Thompson, S. A. (1996). Interaction and Grammar. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620874

Optitrack, NaturalPoint Inc. (v.1.10). Motion Capture System. Retrieved from https://optitrack.com/

Oullier, O., de Guzman, G. C., Jantzen, K. J., Lagarde, J., & Kelso, J. A. (2008). Social coordination dynamics: Measuring human bonding. *Social Neuroscience*, *3*(2), 178–192. doi:10.1080/17470910701563392

Over, H., & Carpenter, M. (2013). The social side of imitation. *Child Development Perspectives. 7***,** 6–11.

Pan, X., & Hamilton, A. F. de C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395-417. https://doi.org/10.1111/bjop.12290

Pan, X., Gillies, M., Barker, C., Clark, D. M., & Slater, M. (2012). Socially anxious and confident men interact with a forward virtual woman: An experimental study. *PLoS One*, *7*, e32931. doi.org/10.1371/journal.pone.0032931

Pan, X., Gillies, M., & Slater, M. (2008). Male bodily responses during an interaction with a virtual woman. *Proceedings of 8th Int. Conference on Intelligent Virtual Agents*, *Japan*. doi:10.1007/978-3-540-85483-8_9

Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavioural Research Methods*, *45*(2), 329–343. doi.org/10.3758/s1342801202492

Pentland, A. (Sandy). (2010). *Honest Signals*. MIT Press.

Pesarin, A., Cristani, M., Murino, V., & Vinciarelli, A. (2011). Conversation analysis at work: detection of conflict in competitive discussions through semi-automatic turn-organization analysis. *Cognitive Process* 13**,** 533–540. https://doi.org/10.1007/s10339-011-0417-9

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioural and Brain Sciences*, *36*(4), 329–347. doi.org/10.1017/s0140525x12001495

Pinzon-Gonzalez, J.G., & Barba-Guaman, L. (2022). Use of head position estimation for attention level detection in remote classrooms. In: Arai, K. (eds) *Proceedings of the Future Technologies Conference (FTC) 2021*, *Volume 1*. https://doi.org/10.1007/978-3-030-89906-6_20

Poggi, I., D'Errico, F., & Vincze, L. (2010). Types of nods. The polysemy of a social signal. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valetta, Malta, 596.*

Poppe, R., Zee, S., Heylen, D. K. J., & Taylor, P. J. (2013). AMAB: Automated measurement and analysis of body motion. *Behavioural Research Methods*, *46*(3), 625–633. doi: 10.3758/s13428-013-0398-y

Powell, L. J., Macrae, C. N., Cloutier, J., Metcalfe, J., & Mitchell, J. P. (2010). Dissociable neural substrates for agentic versus conceptual representations of self. Journal of Cognitive Neuroscience, 22(10), 2186-2197. doi.org/10.1162/jocn.2009.21368

Prins, N., & Kingdom, F. A. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the palamedes toolbox. *Frontiers in Psychology*, *9*:1250. doi:10.3389/fpsyg.2018.01250.

Pupil Labs Inc. Eye Tracking Technology. Retrieved from https://pupil-labs.com/

Ramseyer, F., & Tschacher, W. (2006). Synchrony: A core concept for a constructivist approach to psychotherapy. *Constructivism in the Human Sciences*, *11*(1-2), 150–171.

Ramseyer, F., & Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference. In A. Esposito., N. Campell., C. Vogel., A. Hussain., & A. Nijholt (Eds.), *Lecture Notes in Computer Science: Vol. 5967. Development of Multimodal Interfaces: Active Listening and Synchrony* (pp. 182–196). doi.org/10.1007/978-3-642-12397-9_15

Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body-movements reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, *79*(3), 284–295. doi:10.1037/a0023419

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (Vol. XIV). New York, NY, US: Cambridge University Press.

Rice, K., & Redcay, E. (2016). Interaction matters: A perceived social partner alters the neural processing of human speech. *Neuroimage, 129*, 480–488. http://dx.doi.org/10.1016/j.neuroimage.2015.11.041

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-60.

Richardson, C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science*, *18*(5), 407–413. https://doi.org/10.1111/j.1467-9280.2007.01914.x

Richardson, D. C., & Kirkham, N. Z. (2004). Multimodal events and moving Locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General, 133*(1), 46–62. https://doi.org/10.1037/0096-3445.133.1.46

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76*(3), 269–295. https://doi.org/10.1016/S0010-0277(00)00084-6

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, *25*(1), 70–74. doi.org/10.1177/0963721415617806

Rizzo, A., & Talbot, T. (2016). Virtual reality standardized patients for clinical training. In C. D. Combs, J. A. Sokolowski, & C. M. Banks (Eds.), *The Digital Patient: Advancing Healthcare, Research, and Education* (pp. 257–272). Hoboken: John Wiley & Sons.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Roberts, S. (2011). *DipDap.* [BBC]. Ragdoll Productions.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*, 677–688.

Roseberry, S., Hirsh-Pasek, K., Parish-Morris, J. & Golinkoff, R.M. (2009). Live action: Can young children learn verbs from video? *Child development*, *80*(5), 1360-1375.

Rotondo, J. L., & Boker, S. M. (2002). Behavioral synchronization in human conversational interaction. *Mirror Neurons and the Evolution of Brain and Language*. 151-162.

Rouder, J., Engelhardt, C., Mccabe, S., & Morey, R. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, 23. 10.3758/s13423-016-1026-5.

Russell, J. & Jarrold, C. (1999). Memory for actions in children with autism: Self versus other. *Cognitive Neuropsychiatry*, *4*(4), 303-331.

Sacks, H., & Schegloff, E. A. (2002). Home position. *Gesture*, *2*(2), pp. 133-146.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, *50*, 696–735.

Salazar-Kämpf, M., Liebermann, H., Kerschreiter, R., Krause, S., Nestler, S., & Schmukle, C. (2017). Disentangling the sources of mimicry: Social relations analyses of the link between mimicry and liking. *Psychological Science, 29*(1), 131-138. doi:10.1177/0956797617727121

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413–422. https://doi.org/10.1093/scan/nsr025

Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers.* Houghton, Mifflin and Company.

Schain, C., Lindner, I., Beck, F., & Echterhoff, G. (2012). Looking at the actor's face: Identity cues and attentional focus in false memories of action performance from observation. *Journal of Experimental Social Psychology, 48*, 1201-1204.

Scheflen, A. E. (1963). Communication and regulation in psychotherapy. *Psychiatry: Journal for the Study of Interpersonal Processes, 26*(2), 126–136. http://dx.doi.org/10.1080/00332747.1963.11023345

Schilbach, L., Eickhoff, S. B., Cieslik, E., Shah, N. J., Fink, G. R., & Vogeley, K. (2011). Eyes on me: an fMRI study of the effects of social gaze on action control. *Social Cognitive and Affective Neuroscience*, *6*(4), 393–403. https://doi.org/10.1093/scan/nsq067

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second person neuroscience. *Behavioural and Brain Sciences*, *36*(4), 393–462. doi:10.1017/S0140525X12000660

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., … Vogeley, K. (2009). Minds made for sharing: Initiating joint attention recruits reward-related neurocircuitry. *Journal of Cognitive Neuroscience*, *22*(12), 2702–2715. https://doi.org/10.1162/jocn.2009.21401

Schmidt, R. C., O'Brien, B., & Sysko R. (1999). Self-organization of between-persons cooperative tasks and possible applications to sport. *International Journal of Sport Psychology*, *30*, p. 558–579.

Schmidt, R. C., Morr, S., Fitzpatrick, P., & Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behaviour*, *36*(4), 262–279. doi.org/10.1007/s10919-012-0138-5

Schmidt, R. C., Nie, N., Franco, A., & Richardson, M. J. (2014). Bodily synchronization underlying joke telling. *Frontiers in Human Neuroscience*, *8*(633), 1–13. doi.org/10.3389/fnhum.2014.00633

Schreiber, B. E., Fukuta, J. & Gordon, F. (2010). Live lecture versus video podcast in undergraduate medical education: A randomised controlled trial. *BMC medical education*, *10*(1), 68. http://dx.doi.org/10.1186/1472-6920-10-68

Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science*, *1*, 353-367. doi:10.1111/j.1756-8765.2009.01024.x

Seuren, L.M., Wherton, J., Greenhalgh, T., & Shaw, S.E. (2021). Whose turn is it anyway? Latency and the organization of turn-taking in videomediated interaction. *Journal of Pragmatics*, *172*, 63–78.

Shamay-Tsoory, S. G. (2021). Brains that fire together wire together: Interbrain plasticity underlies learning in social interactions. *The Neuroscientist*. doi:10.1177/1073858421996682

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology*, *29*(2), 326–332.

Skuballa, I. T., Xu, K. M., & Jarodzka, H. (2019). The impact of co-actors on cognitive load: When the mere presence of others makes learning more difficult. *Computers in Human Behavior, 101,* 30–41. https://doi.org/10.1016/j.chb.2019.06.016

Slater, M., Pertaub, D. P., Barker, C., & Clark, D. M. (2006). An experimental study on fear of public speaking using a virtual environment. *CyberPsychology & Behavior*, *9*(5), 627–633. https://doi.org/10.1089/cpb.2006.9.627

Smith, L. B., Colunga, E., & Yoshida, H. (2010). Knowledge as process: Contextually cued attention and early word learning. *Cognitive Science, 34*(7), 1287–1314. https://doi.org/10.1111/j.1551-6709.2010.01130.x

Sofianidis, G., Hatzitaki, V., Grouios, G., Johannsen, L., & Wing, A. (2012). Somatosensory driven interpersonal synchrony during rhythmic sway. *Human movement science*, *31*, 553-66. 10.1016/j.humov.2011.07.007.

Solomon, D. J., Ferenchick, G. S., Laird-Fick, H. S. & Kavanaugh, K. (2004). A randomized trial comparing digital and live lecture formats. *BMC Medical Education*, *4*(1), 27.

Steinmayr, R., Ziegler, M., & Träuble, B. (2010). Do intelligence and sustained attention interact in predicting academic achievement? *Learning and Individual Differences, 20*(1), 14–18. https://doi.org/10.1016/j.lindif.2009.10.009

Stel, M., van Dijk, E., & Oliver, E. (2009). You want to know the truth? Then don't mimic! *Psychological Science*, *20*(6), 693–699. doi.org/10.1111/j.1467-9280.2009.02350.x

Stel, M., Rispens, S., Leliveld, M., & Lokhorst, A. M. (2011). The consequences of mimicry for prosocials and proselfs: Effects of social value orientation on the mimicry-liking link. *European Journal of Social Psychology*, *41*(3), 269–274. http://doi.org/10.1002/ejsp.790

Sterelny, K. (2009). *Peacekeeping in the culture wars.* In K. N. Laland & B. G. Galef (Eds.), *The question of animal culture* (pp. 288–304). Harvard University Press.

Sümer, P., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2018). Multimodal engagement analysis from facial videos in the classroom, in *IEEE Transactions on Affective Computing*, doi:10.1109/TAFFC.2021.3127692.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology,* Vol. *25*, 1–65. Academic Press. https://doi.org/10.1016/S0065-2601(08)60281-6

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A metaanalysis. *Psychological Bulletin*, 121, 371–394.

Taylor, G. J. Ryan, D., & Bagby, R. M. (1986). Toward the development of a new self-report alexithymia scale. *Psychotherapy and Psychosomatics*, *44*(4), 191–199. doi:10.1159/000287912

Thòrisson, K. R. (1994). Face-to-face communication with computer agents. In *AAAI Spring Symposium on Believable Agents Working Notes*, 86-90. Stanford University, Stanford.

Thòrisson, K. R. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. In B. Granström, D. House & I. Karlsson (Eds), *Multimodality in Language and Speech Systems. Text, Speech and Language Technology*, vol. 19. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2367-1_8

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson, *Organization of memory.* Academic Press.

Tulving, E. (1984). Précis of elements in episodic memory. *Cambridge University Press*, *7*(2), 257-268.

Unity Technologies (2020). Real-time 3D development platform. Retrieved from https://unity.com/

Vaccani, J-P., Javidnia, H., & Humphrey-Murto, S. (2016). The effectiveness of

webcast compared to live lectures as a teaching tool in medical school. *Medical Teacher*, *38*:1, 59-63, doi:10.3109/0142159X.2014.970990

van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science*, *15*(1), 71–74. https://doi.org/10.1111/j.0963-7214.2004.01501012.x

van Baaren, R. B., Janssen, L., Chartrand, T. L., & Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2381–2389. https://doi.org/10.1098/rstb.2009.0057

Varlet, M., Marin, L., Lagarde, J., & Bardy, B. G. (2011). Social postural coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 473–483. https://doi.org/10.1037/a0020552

Verberne, F. M. F., Ham, J., Ponnada, A., & Midden, C. J. H. (2013). Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In S. Berkovsky, & J. Freyne, (Eds.). *Persuasive Technology*. *Lecture Notes in Computer Science*, Vol. 7822. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-37157-8_28

Vinciarelli, A., Pantic, M., Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743-1759.

Vinciarelli et al. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing, *1*. 10.1109/T-AFFC.2011.27.

Vrijsen, J. N., Lange, W.-G., Becker, E. S., & Rinck, M. (2010). Socially anxious individuals lack unintentional mimicry. *Behaviour Research and Therapy*, *48*(6), 561–564. https://doi.org/10.1016/j.brat.2010.02.004

Wakefield, E., Novack, M. A., Congdon, E. L., Franconeri, S., & Goldin-Meadow, S. (2018). Gesture helps learners learn, but not merely by guiding their visual attention. *Developmental science*, *21*(6), e12664. https://doi.org/10.1111/desc.12664

Walton, A. E., Richardson, M. J., Langland-Hassan, P., & Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00313

Wang, Y., & Hamilton, A. F. C. (2012). Social top-down response modulation (STORM): A model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience*, *6*(153). doi.org/10.3389/fnhum.2012.00153

Wang, Y., Ramsey, R., & Hamilton, A. F. de C. (2011). The control of mimicry by eye contact is mediated by medial prefrontal cortex. *The Journal of Neuroscience*, *31*(33), 12001–12010. https://doi.org/10.1523/JNEUROSCI.0845-11.2011

Washburn et al. (2014). Dancers entrain more effectively than non-dancers to another actor's movements. Frontiers in Human Neuroscience, 8, https://doi.org/10.3389/fnhum.2014.00800

Wilmermuth, S. S., & Heath, C. (2009). Synchrony and cooperation. *Psychological Science*, *20*, 1–5. http://dx.doi.org/10.1111/j.1467-9280.2008.02253.x

Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G. R., & Vogeley, K. (2010). It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for. *Social Cognitive and Affective Neuroscience*, *5*(1), 98–107. https://doi.org/10.1093/scan/nsq024

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *ArXiv:1308.5499 [cs.CL]*. https://doi.org/10.48550/arXiv.1308.5499

Won, A. S., Bailenson, J. N., Stathatos, S. C., & Dai, W. (2014). Automatically

detected nonverbal behavior predicts creativity in collaborating dyads. *Journal of Nonverbal Behavior*, *38*(3), 389–408. https://doi.org/10.1007/s10919-014-0186-0

Woodbury-Smith, M. R., Robinson, J., Wheelwright, S., & Baron-Cohen, S. (2005). Screening adults for Asperger syndrome using the AQ: A preliminary study of its diagnostic validity in clinical practice. *Journal of Autism Development Disorders*, *35*(3), 331–335. doi*:*10.1007/s10803-005-3300-7

WorldViz (2022). Virtual Reality for Training and Research. Retrieved from https://www.worldviz.com/

Yngve, V. (1970). On getting a word in edgewise. *Papers from the Chicago Linguistic Society*, *6*, 567-577.

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *Journal on Image and Video Processing,* 80. https://doi.org/10.1186/s13640-017-0228-8

Zell, E., Carlos, A., Jarabo, A., Zebrek, K., Gutierrez, D., McDonnell. R., & Botsch, M. (2015). To stylize or not to stylize? The effect of shape and material stylization on the perception of computer-generated faces. *ACM Transaction on Graphics 34*(6)*, SIGGRAPH.*

Appendix: Exploratory Analysis

This section has been added for completeness, and to show some added support for the third hypothesis (**H**$_3$) which stated that *Coherence of slow head nods is a product of joint attention or gaze following.* This is not part of what we consider the main analysis between real and pseudo pairings but is kept for exploratory purposes.

We performed a 2×3 repeated measures ANOVA with Real (Real, Pseudo) and Task (Picture, Video, Meal) as within-subject factors. Results are shown separately for each main effect together with the interaction effect (Figure A). Graphs A, B and C show the mean and standard error of coherence (R^2) of each effect. High coherence means a high degree of coordination, as it indicates that two people are moving with the same frequency. To assess the difference in coherence between all the conditions, we calculated the effect size as an indication of the % of variance between them. Graphs D, E and F show the effect sizes (partial eta-squared, ηp^2).

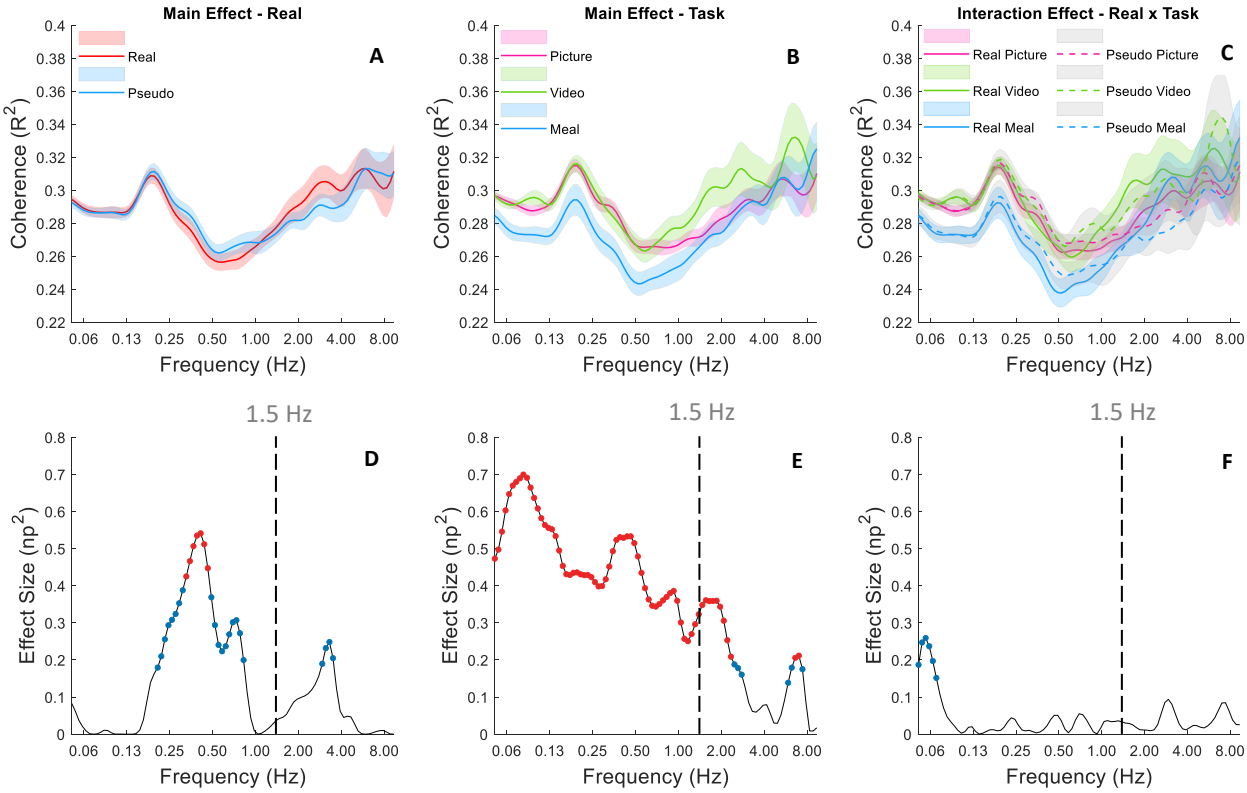


Figure A. ANOVA cross-wavelet coherence. Graphs A, B, and C show the mean and standard error of coherence (R^2) of each effect. Graphs D, E, and F show the effect sizes (partial eta-squared, ηp^2). The dotted line indicates frequencies where there is a significant difference of coherence. Red dots represent points on the frequency range that pass a $p<0.05$ FDR significance threshold, while blue dots represent significant differences that did not pass this threshold.

We observe two distinct patterns of coherence across the range of frequencies displayed. These patterns are divided into two frequency ranges, above and below 1.5 Hz, as indicated by the dashed vertical line (D, E, F). In the low frequency range (<1.5 Hz) we observe a significant main effect of real interactions, and a significant main effect of task. However, we do not observe a significant interaction effect between the factors. In the high-frequency range (>1.5 Hz), results show no significant main effect of real interactions, but we observe a significant main effect of task. Again, no significant interaction effect between the factors was found.

The results suggest that, averaged over the three tasks, slow nodding coherence during real interactions were significantly greater than pseudo interactions. Furthermore, results also showed that, averaged over real and pseudo interactions, slow nodding coherence across all three tasks were significantly greater. No significant interaction was observed between the two factors, indicating that the analysis of real vs. pseudo interactions did not lead to greater slow nodding coherence depending on which context they were in. This lack of interdependence can be valuable in informing our interpretation of this behaviour.

The observed main effects further strengthen the third hypothesis (**H_3**), as it shows that the participants are behaving differently across the tasks. Looking at the main effect of task (B) gives us an indication of the direction of this effect since we only compared the variance across all three tasks, but it looks as if we have less coherence in the Meal Planning Task than the other two tasks. The lack of an

interaction effect between the two factors indicates that it being a real or pseudo interaction did not necessarily lead to greater slow nodding coherence depending on which specific context they were in. A reason not to put too much weight on the interaction effect is perhaps that we see a much stronger main effect of task compared to real interactions. This can easily overshadow whatever is going on when estimating the interdependence between the two factors.

The results also showed that, averaged over real and pseudo interactions, fast nodding across all three tasks was significantly greater. Similar to the slow nodding behaviour, no significant interaction was observed between the two factors, indicating that the analysis of real vs. pseudo interactions did not lead to greater fast nodding depending on which specific context the participants were in.

Looking at the main effect of task (B) again gives us an indication of the direction of this effect, but for fast nodding it looks as if we have more coherence in the Video Discussion Task than the other two tasks. Based on the results from the real vs. pseudo interactions, this makes sense since the participants do not display the fast nodding pattern compared to the other two tasks, which consequently leads to the participants' having more coherence at higher frequencies relative to the other tasks. However, for the same reasons that we cannot put so much weight on the direction of the results for the slow nodding behaviour in terms of the main effect of task and the interaction between the two factors, we have to be equally cautious when interpreting the results of the fast nodding behaviour.