STAT489 Project

Roydon Goldsack

2021-07-26

Contents

Chapter 1

Introduction

The current report investigates the psychological concept of empathy using statistical modelling techniques. The project was split into two parts, firstly the data preparation that was completed as part of the Summer Research Scholarship I was awarded. The work was on the eMotion project during the 2020-2021 Trimester 3. The second part is using a subset of the prepared data to build a model of cognitive empathy. The report discusses both of these parts. The information provided here is intended to be used by the researchers of the eMotion study. The information includes, but is not limited to, a detailed description of the data generated by the eMotion study which is useful as a reference for researchers working with the data. An investigation of some key descriptive statistics is also be provided to increase familiarity with the data before the models are created. The introduction sets out the theoretical basis for the models provided later on and discusses the importance of the work being done.

1.1 Defining Empathy

Empathy is an important part of our daily lives and helps us move through our social world. Empathy can lead us to make decisions to help others or; reduce harm done to them (Bloom, 2017). Empathy also helps make us aware of what others are feeling and share these emotions (Ickes, 1993). Given the intuitive understanding of empathy based on these aspects and our experiences, one would expect consensus in defining empathy. However, there is no singular agreed-upon definition of empathy in the literature (Bloom, 2017; Eklund & Meranius, 2020; Hall & Schwartz, 2019) and some definitions

across studies even contradict each other (Hall & Schwartz, 2019). There are several reviews of the literature on empathy attempting to consolidate its many definitions (i.e. Eklund & Meranius, 2020; Hall & Schwartz, 2019). Definitions of empathy tend to vary across several f different contexts as pointed out by Hall and Schwartz (2019). Firstly, definitions depend on the populations being empathised for such as victims of sexual abuse vs. typical WEIRD (white, educated, industrialised, rich, democratic) research participants. Next, they differ on how empathy is measured. For example, a large number of studies used self-report measures of empathic traits such as the Interpersonal Reactivity Index (IRI; Davis, 1980), while a smaller number used state empathic measures that ask for empathy towards specific individuals (i.e. patients, a person in a vignette). Self-report measures can be problematic due to issues of demand characteristics or biases, for instance, participants may be biased towards reporting they are more empathetic due to social pressures (Coll et al., 2017).

The variation in definitions seems to be narrowing in towards agreement of empathy as a multidimensional concept, most frequently distinguishing between cognitive and affective empathy. The multidimensionality of empathy refers to the combination of multiple processes, types, features etc. For example, the commonly used empathy scale, the IRI, includes four different forms of empathy: perspective-taking, fantasy, empathic concern, and, personal distress (Davis, 1980). Cognitive empathy is inferring what another person is feeling. For example, inferring that a person is sad based on some, often behavioural, cues such as their head pointing down, shoulders slumped. Affective empathy is sharing another person's affect and emotions. For example, if your friend is very excited, you feel excited with them. Hall and Schwartz (2019) also point out that when the IRI scales are not used the most frequent choices were scales of empathic concern and perspective taking, representing again affective and cognitive empathy, respectively. They go onto discuss the large variation in the conceptual definitions of empathy used and then the variation in the number of defining features in these definitions. This variation is further evidence of the difficulty in defining empathy. Eklund and Meranius (2020) conducted a thematic analysis on 52 reviews of the empathy literature. They identified 13 sub-themes, four of which were found in all or almost all articles. These four themes were: understanding, feeling, sharing, and, self-other differentiation. They state, "Thus, in the empathy literature there is virtual consensus that the empathizer (1) understands, (2) feels, and (3) shares the other person's feelings (4) with self-other differentiation" (p. 3). This definition implies a consensus undetected by Hall and Schwartz

(2019) and is supported by the large body of literature analysed, however, their findings have yet to be tested empirically. Thus far we have focused on the definition of empathy. Finding that there is a general lack of consensus but the literature is slowly moving toward an agreed-upon definition. We now move our focus to the operationalisation of these definitions and the methods of measuring empathy.

1.2 Measuring Empathy

The measurement of empathy is directly linked with its definition and thus varies in the literature (Hall & Schwartz, 2019). Empathy has predominantly been measured using self-report measures such as the IRI, however, following the discovery of mirror neurons, a neuroscientific approach has been increasingly taken (Eklund & Meranius, 2020; Hein & Singer, 2010). Mirror neurons fire in the patterns of specific actions within an observer that are being performed by an external actor. For example, watching a person, the actor, raising their right hand and waving will cause the mirror neurons associated with the motor control of the right arm of the observer to fire. Due to their ease of application, and depending on the construct, good validity self-report measures are a mainstay in psychology. The issues with self-report measures have been discussed previously, but measures such as the IRI continue to be widely used. This allows us to compare our findings across other studies and be confident that our results will be reliable. The neuroscientific approach uses psychophysiological and brain imaging correlates of empathy to attempt to remove any potential biases introduced through self-report measures (Neumann & Westbury, 2011). Correlates are measures that are believed to represent and co-occur with the phenomenon of interest. For example, studies have used functional magnetic resonance imaging (fMRI) to investigate brain activation in response to empathy provoking stimuli, largely in the domain of pain (Hein & Singer, 2010). There has also been work on the effect of brain lesions on empathy (Decety, 2011). Overall these correlative measures have proved to be at least reasonably effective in measuring empathy and giving researchers another option other than self-report measures.

There have been many studies using psychophysiological methods to investigate empathy. These methods include, for example, measuring skin conductance and heart rate. Skin conductance is used as a measure of arousal, typically with higher amounts of conductance and thus sweat, indicating higher arousal. Heart rate is also used as a measure of arousal, typically

with higher heart rate indicating higher arousal. These are both measures of autonomic nervous system activity. As outlined by Neumann and Westbury (2011), skin conductance and heart rate are often measured as tonic or phasic. Tonic measures mean heart rate and variability over time. Phasic changes indicate short-term changes in state as a result of stimulus events. These two forms allow for closer investigation of empathy over time. There are several studies that combine both self-report and psychophysiological measures of empathy (Neumann & Westbury, 2011). These studies have found significant correlations between the two measurements indicating psychophysiological measurements of empathy have good validity. Recently, motion capture has been introduced to research in affective sciences and is used to predict various outcomes such as emotion (Li et al., 2016). Li et al. (2016) used motion capture data of human gaits to categorise various emotions. Their model was able to predict when an individual was angry or happy with decent accuracy solely based on their gait. The possibilities of motion capture in psychology are just beginning to be seen and thus the body of literature is still small.

One of the most important aspects of empathy is that it happens between individuals (Schilbach et al., 2013). So far the studies outlined have focused on self-report measures from one individual, or psychophysiological measures of an individual as a response to stimuli. Therefore, forgetting empathy's basis in interactions. The study of empathic accuracy attempts to partly address this. The study of empathic accuracy investigates how accurate the two people are at judging the emotional state of the other (Ickes, 1993). Empathic accuracy gives evidence for cognitive empathy. The evidence for cognitive empathy comes from how accurate the dyad of individuals is at judging the state of the other. This can be extended to include affective empathy. The evidence for affective empathy comes from how similar the states of the dyad are. For example, Stinson and Ickes (1992) investigated empathic accuracy in interactions between male friends and male strangers. They subtly videoed a dyad (two people) interacting while sitting in a waiting room, waiting for the experiment to begin. They then informed the men of the premise of the study and asked the men to rate how they were feeling while they were waiting. Each member of the dyad watched the video of their interaction twice, once coding how they felt and then coding how they thought the other person felt. They then found that the dyads of friends were more accurate at inferring the state of the other person than strangers. They also found significant correlations in the accuracy of describing what the other was feeling between the two peoples responses for friends but not for strangers. This indicates again that dyads who know each other well

had greater empathic accuracy than strangers. Roth and Altmann (2021) took empathic accuracy one step further and investigated whether a person's actual ability to recognise the emotions of others led them to be perceived as more empathic. They found that the closer the relationship to the target, the person whose emotional state was being rated, the more accurate the rater was at inferring how empathic the target was. The intimate partners were more accurate at judging how empathic a person was than family and friends.

1.3 Current Study

The current study looks to develop a statistical model of cognitive empathy that can effectively model the phenomena of cognitive empathy. Due to time constraints, affective empathy is planned to be modelled in the future. The model we build is intended to be a proof of concept and if it effectively models cognitive empathy, we have confidence that we can then model affective empathy. We build off the following definition of cognitive empathy for our model. Cognitive empathy is one person knowing the state of another person. An important difference between this definition and our model is that our model allows for the amount of cognitive empathy to change over time. Empathy is known to be a transitory phenomenon i.e. a phenomenon that changes across time (Coll et al., 2017). Then looking to the literature on empathic accuracy for more detail in how cognitive empathy has previously been modelled.

The report describes the intricacies of data preparation completed during the Summer Research Scholarship. The data preparation included building a pipeline that was used to process the data from the eMotion project. For example, this included restructuring, re-sampling, filling and cleaning a large amount of data. The project opted to create a single, reusable pipeline for data processing to radically decrease processing time, increase reliability and simplicity. The time required would have been increased dramatically if processing had been done by hand and likely led to numerous errors. The raw data from the eMotion study, the study from which our data was collected, was messy and un-analysable, necessitating the data processing. For example, the data used for the cognitive empathy model was in an unusable state with extra columns and extra output from the program it was recorded in. The scripts used for cleaning were created for a given data type and worked with all data of the type. Finally, creating a single script that processes all the data and allows for changes to be made to the processing from a single

point that trickle down to the other steps. The script was then deployed to a high-performance cluster of machines to clean all the data in parallel.

Chapter 2

Background

2.1 The eMotion Project

The eMotion study is an inter-disciplinary project being conducted by Hedwig Eisenbarth and Areito Echevarria at Victoria University of Wellington (VUW) alongside several research assistants and postgraduate students. The project is exploratory in nature with many outcomes ranging from the current report to an app being built to conversation analysis. Initially, the project was developed to investigate and construct a better understanding of empathy. The scope of the project has since increased to incorporate other questions and overlapped with other studies due to the breadth of data. It is important to briefly emphasise the diversity of perspectives that are incorporated in the project. The two lead researchers, Hedwig Eisenbarth and Areito Echevarria are from the School of Psychology and School of Design Innovation respectively. This report is sitting in the School of Mathematics and Statistics, with researchers from Computer Science and Linguistics also being involved in aspects of the project.

2.2 Study Methodology

A primary focus of the study is providing a data-driven naturalistic investigation of empathy. The study provides this by having dyads of participants discuss four topics while recording a large number of measures. The various measures allow for subtle and varied aspects of the interactions to be investigated.

2.2.1 Participants

Data has been collected from 50 dyads of two participants. Half of the dyads are made up of two participants that were either family, friends or partners, the other half are two strangers who have not met before. The ages range between 18 and 78 with a mean of 28 years. The sample is 68% individuals who identify as female, 28% who identify as male and 4% who identify as other genders.

2.2.2 Materials

Care Scale. Participants are asked to rate first the amount of care or attention they felt like they needed if the event shared had just occurred to them. They give a rating for how much care or attention they felt they needed at the beginning, middle and end of the sharing session. Needed care or attention is measured on a 0 to 10 Likert scale where 0 is "requiring no care and attention at all" and 10 is "requiring a lot of care and attention." Participants then rate the amount of care they think their partner would have needed if the event shared had happened to their partner on the same 0 to 10 Likert scale. The scale measures the intensity of the event that was shared.

Inclusion of the Other in the Self (IOS) Scale. Participants are presented with seven diagrams of two circles labelled "Self" and "Other" and asked to "Please select the diagram that best describes you and your experiment-partner" (Gachter et al., 2015). Each pair of circles has gradually more area overlapping from not at all overlapping to almost entirely overlapping.

Interpersonal Reactivity Index (IRI). Participants are presented with 28 statements relating to empathy (Davis, 1980). The statements are rated from "does not describe me very well" to "describes me very well" on a 1 to 5 Likert scale. For example, "I daydream and fantasize, with some, regularity, about things that might happen to me," or, "When I see someone get hurt, I tend to remain calm."

PANAS (Positive and Negative Affect Schedule) Scale. Participants are shown 20 adjectives and are asked to "Indicate to what extent you feel this way right now, that is, at the present moment" (Watson et al., 1988). The adjectives include "Interested," "Distressed" and are rated from "Very Slightly or Not at All" to "Extremely" on a 1 to 5 Likert scale.

Psychopathic Personality Inventory-Revised-40 (PPI-R-40). Participants are presented with 40 statements relating to psychopathic personality traits (Eisenbarth et al., 2014). The statements are rated as either "F," "MF," "MT" or "T" meaning "False," "Mostly False," "Mostly True" and "True." For example, "I have always seen myself as something of a rebel," or, "If I can't change the rules, I try to get others to bend them for me."

2.2.3 Procedure

First, the dyad is given a tour of the stage and equipment. The dyad then fills out informed consent sheets and demographic information. They are then fitted into the motion capture suits and put on the belts used for the physiological measures, then have the helmets with cameras mounted attached. The suits are black bodysuits. The belts have various sensors to measure heart rate, skin conductance etc. Next, reflecting dots were attached to the bodies and joints of the participants' motion capture suits. Then markers were drawn on the faces of the participants at landmark points such as their eyes, mouth etc. Then the dyad completed a closeness intervention task. The task has the participants hug and get to know each other. This is followed by a physiological baseline recording and range of motion exercises. The range of motion has the participants move their limbs around to calibrate the motion tracking software. The dyad then completed the IOS scale for the first time. The dyad then completed the four conversations, after the last of which the dyad completed the IOS scale again. The participants are given a topic to discuss for each conversation which they talk about. After each conversation, the dyad hugged and ended in a T-pose. The participants stand with their arms up and out to the sides in a T-like position to calibrate and re-calibrate the motion tracking software. They then completed the PANAS on their current state. The specific topics discussed by participants changed from the beginning of data collection, but the majority of participants discussed one positive topic, then two negative topics, then ending with a positive topic. A picture showing the stage on which the participants stand is included in the appendix figure 10.2.

The dyad took a break then watched the four videos of their conversations, first rating how they felt then re-watching the four videos rating how they think the other participant felt. A picture of the ratings computer set-up is included in the appendix figure 10.1. The participants are told to move the mouse attached to the computer moment-by-moment depending on how they are feeling. If they are feeling less emotionally intense they move the mouse

one way and more emotionally intense, the other. The scale the participants can rate their emotional intensity on ranges between 0 and 10. The dyad then completed the post-session ratings and the PPI-R-40 before being finally debriefed and given vouchers for participation.

Figure 2.1 shows the full procedure described above. It shows the approximate time spent on each part of the procedure.

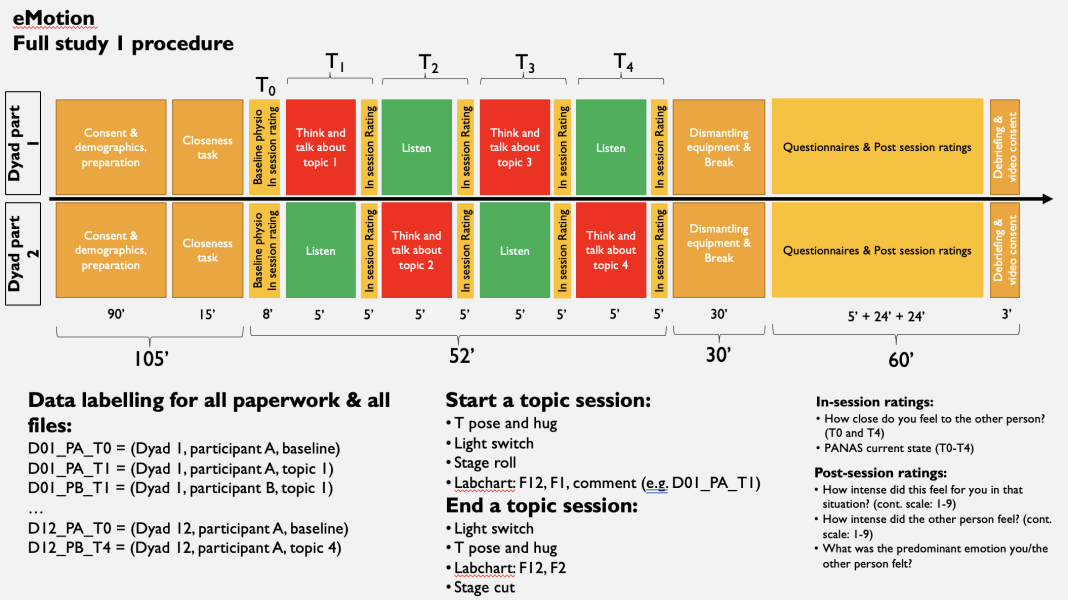
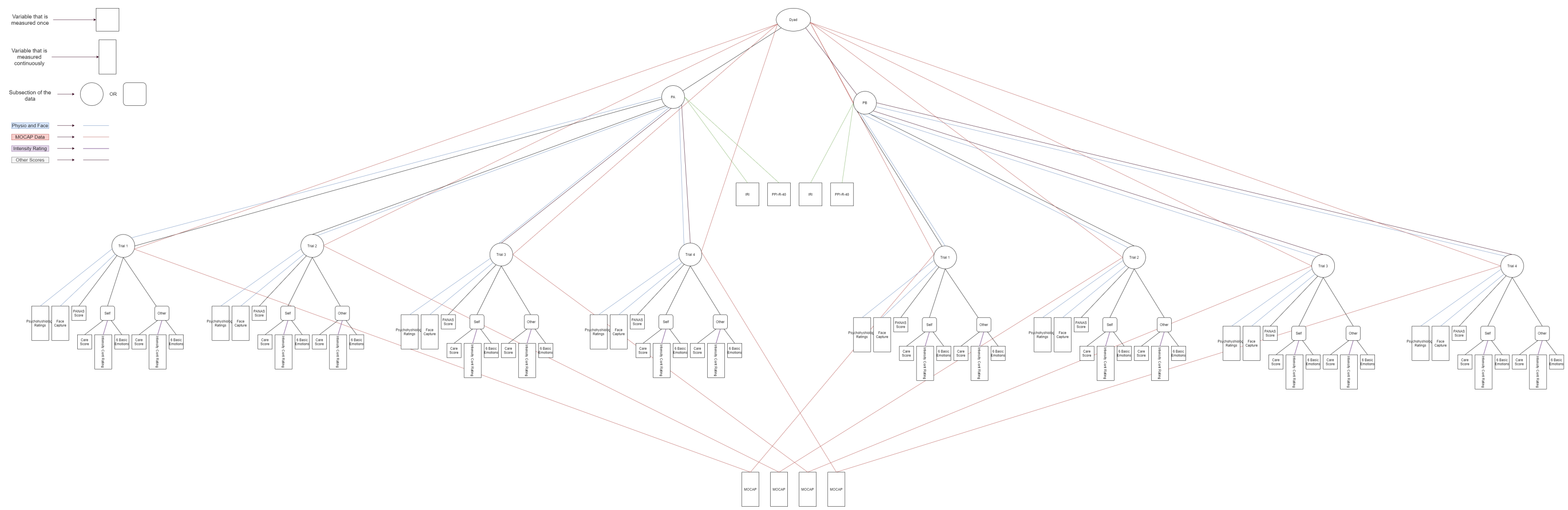


Figure 2.1: Study method and timings.

Chapter 3

Data

The following descriptions are with 50 dyads participating in the study. The data recording in a study as complex as this can lead to errors requiring entire dyads data to be excluded and requiring another dyad to replace their data.


Variable that is measured once
Variable that is measured continuously
Subsection of the data
OR
Physio and Face
MOCAP Data
Intensity Rating
Other Scores

3.1 Data Information

3.1.1 Example Measure

A brief description is given of the measure and how it is recorded.

Structure

The structure of the data is given here. For example, if a measure has the following structure:

Dyad × PX × Conversation

This indicates that the measure is recorded once per conversation, for each participant in the dyad separately and separately for each dyad.

Structure Size

The "number" of observations is given here. The actual number of observations depends on whether the measure is being continuously recorded over time within a conversation or if it is being recorded just once per conversation, for example. A structure size for the structure given above would be:

50 × 2 × 4

This shows that we have 50 dyads observations, these observations are split by participant for a total of 50 x 2 observations, then finally split again by conversation # for a total of 50 x 2 x 4 observations. Therefore, for this example measurement, we would have 400 total observations of this measure from across all our participants. This is a simple indicator of the volume of data for a specific measurement.

Frequency of Measurement

Continuing with the running example, this measurement would be taken once per conversation. If a measure is recorded continuously the frequency is given in Hz.

3.1.2 Continuous Emotional Intensity Ratings

These ratings form the basis of the model of cognitive empathy created later. They are the ratings of each participant's emotional intensity throughout the conversations, or what they think the other persons emotional intensity was.

Structure

Dyad × PX × Self/Other × Conversation × Continuous measurement

Structure Size

50 × 2 × 2 × 4 × Continuous measurement at variable rate

Frequency of Measurement

The data was recorded at a variable rate due to software used to record the measurements only recording observations when the mouse is moved and for a short time after that. Thus the data is up-sampled to 119.88Hz in processing to match the other continuous measurements.

3.1.3 Motion Capture Data

The motion capture data is taken from the motion capture suits that each of the participants wore. These suits have reflective dots attached to them which reflect infra-red (IR) light that is picked up by cameras that are all over the stage where the participants are standing and talking. The data from the cameras goes into the Motive software which builds a skeleton of each of the participants (NaturalPoint Inc., Corvallis, OR). This skeleton is made up of the positions and rotations of all of the bones and joints. The data is then exported from `.TAK` files to `.csv` files using Motive.

Structure

Dyad × PX (One rating for Participant A & Participant B) × Conversation × Many Rotation/Positions × Continuous Measurement

Structure Size

50 × 2 × 4 × Many Rotation/Positions × ~6 minutes of recording at 119.88Hz

Frequency of Measurement

119.88Hz

3.1.4 Psychophysiology

The psychophysiology measures consist of a number of physiological measures that are assumed to indicate the internal state of the individual. For example, if an individual has a heightened heart rate and increased sweat (higher skin conductance) they have increased autonomic nervous system activity. The autonomic nervous system activity is physiological arousal, and in turn, correlates with emotional arousal. The measures are therefore known as psychophysiology as the physiological measures are intended to shed some light on the individual's internal psychological state. The measures collected

are ECG (Electrocardiogram) & heart rate (calculated online from the ECG), chest expansion, GSR (Skin conductance/sweat), skin temperature and acceleration in the X, Y, Z directions.

Structure

Dyad × PX × Conversation × Different measurements × Continuous measurement

Structure Size

50 × 2 × 4 × Different measurements × ~6 minutes of recording at 250Hz

Frequency of Measurement

250Hz

3.1.5 Face Capture

Head-mounted cameras are put on participants heads and aimed at their faces with dots on them. The dots are landmark indicators and can be used to extract certain features and expressions from the participants in post-processing. The face capture data is used for exactly that, expression tracking across the conversations. These expressions then correlate with emotional arousal and specific emotional states.

Structure

Dyad × PX × Conversation × Different points on the face × Continuous measurement

Structure Size

50 × 2 × 4 × Different points on the face × ~6 minutes of recording

Frequency of Measurement

59.94Hz

3.1.6 Overall Emotion Ratings

Following the continuous rating of how the self or other were feeling in the videos of the dyad's interaction (as above) participants firstly rate how strongly they felt each of the six basic emotions, then how strongly they think the other person was feeling the six basic emotions. The six basic emotions are disgust, surprise, anger, sadness, happiness and fear.

Structure

Dyad × PX × Self/Other × Conversation × Six Emotions

Structure Size

50 × 2 × 2 × 4 × 6

Frequency of Measurement

Once per conversation for both the Self and Other

3.1.7 PANAS Score

The Positive and Negative Affect Scale score measures the self-reported amount of positive and negative emotions experienced by the participants at a given point in time i.e. their current state.

Structure

Dyad × PX × Conversation

Structure Size

50 × 2 × 4 × 1

Frequency of Measurement

Once per conversation

3.1.8 PPI-R-40 Score

The Psychopathic Personality Inventory-Revised measures self-reported psychopathic personality traits. The sub-factors are self-centred impulsivity, fearless dominance and coldheartedness traits.

Structure

Dyad × PX

Structure Size

50 × 2

Frequency of Measurement

Once per participant

3.1.9 Care Score

A measure of the amount of care needed if the event discussed in the conversation had just happened. The participants rate how much care they would have needed on a 0 to 10 scale, and how much care the other person would have needed.

Structure

Dyad × PX × Self/Other × Conversation

Structure Size

50 × 2 × 2 × 4

Frequency of Measurement

Once for the Self and once for Other per conversation

3.1.10 IOS Score

How close the participants feel like they are to the other person.

Structure

Dyad × PX × Before/After interactions

Structure Size

50 × 2 × 2

Frequency of Measurement

Once before and after each conversation

3.1.11 IRI (Interpersonal Reactivity Index) Score

A self-reported measure of how empathic a person states they are in general.

Structure

Dyad × PX

Structure Size

50 × 2

Frequency of Measurement

Once at the end of the experiment

Chapter 4

Outcomes

The work stemming from the eMotion project spans both the Summer Research Scholarship and this report, as well as other smaller projects for classes leading to the generation of various outcomes. The primary measures discussed, motion capture (mocap), emotional intensity ratings (called ratings, emotional state ratings, intensity ratings) and physiological measures (physio), are all "time series" variables. They are recorded and vary, across time. This leads to a very large number of data points thus increased processing time both in analyses and processing steps. The multiple outcomes required by both this report and Summer Research Scholarship are thus wary of the large amount of data that is being used. The outcomes I was part of or responsible for, with relevant programming language choices, are outlined below.

Firstly, the primary goal of the Summer Research Scholarship was to process the data and build a data-structure that could be used for subsequent machine learning analyses. The choice was thus made to complete data processing using `Python 3` as it is a primary language used for machine learning analyses. Processing the data in Python leads to some necessary consequences. Firstly, the aforementioned data structure created in `Python` is solely compatible with `Python` meaning analyses completed in `R` require a separate data structure. The data structure implemented in `Python` is a large dictionary object created for each dyad that contains all the measures, which itself is nested within another structure of all the dyads. The data structure used in R is a dataframe or similar due to its widespread use and ease of operation. However, it must be noted speed is a primary goal in analyses due

to another outcome down the line being a `Shiny` web-app requiring efficient data processing.

Secondly, several graphical displays were created to more easily visualise the large, complex data used in the current project. Both `Python` and `R` have good support for data visualisation with the `ggplot2` package in `R` and the `Matplotlib` package in `Python`. Due to my extensive experience working with `ggplot2` in `R`, the various plots are created using `ggplot2`. The plots are shown in the Descriptive Information section of this report. The plots show the ratings data and show how these data are changing over time both within conversations and across them. The graphs are intended to show basic patterns in the data and to provide evidence for how and why the primary analysis, the cognitive empathy model, is being performed.

Thirdly, a short video was produced detailing a number of the data-processing methods used in the Summer Research Scholarship. The video was aimed at explaining the abstract and little-understood methods to a general audience. It was submitted to the Summer Gold Competition which gives the Summer Research Scholarship students a chance to present their work. The video was also presented to the researchers involved in the eMotion project for feedback and to ensure the validity of the content. The video was then displayed for the general public to see on the VUW campus.

Finally, as part of work for this report and to help visualise the large amount of data produced by the aforementioned study, a `Shiny` application was produced. The application allowed the user to visualise and better understand data collected in the eMotion study. It has several interactive, user reactive elements that allow for visualisation of subtle relationships between the three main measures from the eMotion study (mocap, physio and ratings data). The application was created as part of a course taken alongside this report and is currently publicly accessible on the Shiny apps' website. The application was written using `RStudio` and the `Shiny` platform and libraries.

Chapter 5

Summer Work

5.1 Processing

The data collected in the eMotion project is from several different sources with varying degrees of pre-processing and processing required. The aim is to have a single data structure that contains all our continuous data at the same frequency of measurement and length of time, for correct analysis of the data. For example, correlating an individuals' emotional intensity ratings with their mocap data would require that these are of the same length of time. Our data structure must also include entries showing all our variables that are not measured continuously such as the six emotion ratings, or a participants trait empathy (IRI) score. To create this data structure we first process our measures as outlined below.

The data processing in the summer work focused on three main measures. The measures were the motion capture data, the physiological measures data and the ratings data. The data processing is completed as a series of steps executed by a single function. The function is in turn made up of a series of smaller functions that process small parts of the overall problem. For example, the motion capture data has a function that imports and processes the data's header in a table format, then sends the data into the next function that trims and exports the data. This style of building smaller functions to complete an overall processing step was used throughout the summer work. This style is known as functional programming.

The motion capture and ratings data are first processed together then physio data is processed next. The physio data requires that the mocap data has

been processed already. Following numerous discussions in the project team, the mocap data was decided on as the "source" dataset. The data collection often ran into issues with data being recorded slightly differently across measures or incorrectly and the motion capture data was simply the most reliable. Therefore, the length of time and structure of the mocap data was treated as the target for the two other data types. Towards the end of the summer work, a small change was added to slightly trim the end of the motion capture data which then trickles down the other measures.

The physiological measures were pre-processed in LabChart 8 (AD Instruments, Colorado Spring, CO). LabChart 8 was used for a preliminary clearing of the data and a brief visual inspection. It became clear that not all the data was not going to be ready for processing in `Python` close to the end of the project. The decision was made then to create a pipeline step for the physiological measures data processing right at the end of the summer work. There was not enough time to add this into the pipeline, thus it was implemented outside of the scope of the summer work.

The processed data are exported into a single folder per dyad. The data are then entered into a single master data structure. The data are all contained within a `Python` dictionary object which essentially allows for many different forms of data and information to be stored in a single object. The structure is exactly that outlined in the Data Information section. The data structure at the time of writing is still a work in progress with just the motion capture, physiological, ratings and six emotion ratings data being included. Several measures and demographic information still need to be processed and added to the data structure. The data structure serves as a single point where all the data is contained and utilised for later machine learning and other analyses. The structure is serialised and written to disk once fully developed because of the time-intensive nature of building the structure. Reading the data structure before utilisation is less time-intensive than an entire re-build. The data structure was considered for use in this report but due to technical limitations such as inter-language compatibility, it was decided against. Instead, the current report used the cleaned `.csv` files that have gone through the various processing steps imported into dataframes.

The three data types were aligned using timecode markers that print the local time when each data stream was recorded. The timecodes give the local time at and throughout the time of recording. These are industry standard in the film industry and as mocap data was being utilised it made sense to include these throughout the data types. The timecodes proved very helpful

in data processing and ensuring that measures began and ended at the right points.

Below are the processing steps and information that were needed to prepare each of the three main data types. All other measures were not processed as part of the summer work.

5.1.1 Emotional Intensity & Emotion Ratings

The Intensity and Emotion Ratings are recorded using the PsychoPy software (Peirce, 2007). PsychoPy exports three files, a `.log`, `.csv` and an irrelevant proprietary file. The `.csv` files are not used as they are very messy and difficult to use, leaving us with the `.log` files to extract relevant information. First, the `.log` files are imported to the `JupyterLab` integrated development environment (IDE) for `Python` and then exported as `.csv` files. The exported files contain a large amount of irrelevant information which is removed. Then the files just include the emotional intensity ratings, whether the participant is rating themselves or the other person, the conversation number, when each conversation ended and the six overall emotion ratings. The data is initially made up of a large number of strings that are first split into several columns. These columns were processed and the final set of columns with information is three time columns, emotional intensity rating column, a column for each of the six overall emotion ratings and two more columns indicating conversation and whether emotional intensity rating was for the self or other person. As a tabular data format was chosen, the primary method of finding relevant information was by row-based indexing. For example, if looking for conversation 1 for self, one selects "Self" in the "Self/Other" column and "1" in the "Conversation" column. A short extract of data is shown in the appendix in table 10.1.

At this point, we needed to up-sample and remove the head and tail of the files. PsychoPy only records the felt intensity when the intensity changes/only records the cursor position when the cursor moves. This means that the ratings data is recorded at a variable rate. To get the data to a constant rate, and the same sample rate as the mocap data (119.88Hz) we need to up-sample. Re-sampling changes the recording frequency of time-series data and up-sampling increases the frequency of observations. There are a few different ways this can be done but the method opted for here is to simply copy the most recent observation forward. The logic was that our best guess at what a person is feeling is simply the last thing they told us they were feeling. We have no evidence that their state has changed and therefore we

believe that their state has not changed since the last update of their state.

Once up-sampled we can deal with the next issue, the videos used to create the ratings data start before and finish after the motion capture data. The participants rate how they are feeling moment-by-moment throughout these videos. The videos and therefore ratings are slightly longer than the time the mocap data was recorded for. Thus we need to remove the head and tail of the up-sampled ratings data. To remove the head of the ratings data we need to calculate the difference in the time between when the video and the mocap data started recording. This value should always be positive i.e. the motion capture data should always start recording after videos. Regardless the processing works fine if the value is negative (motion capture data started recording first).

We first find $Time_{Video}$ which is the video start time, then we find $Time_{Mocap}$ which is the mocap start time. Once we know $Time_{Video}$ and $Time_{Mocap}$ we can calculate the difference between these

$$Time_{Diff} = Time_{Mocap} - Time_{Video}$$

Thus

$$Time_{Mocap} = Time_{Diff} - Time_{Video}$$

Then once we have $Time_{Diff}$ we find the corresponding point in the ratings data where the mocap started ($Time_{Diff} - Time_{Video}$) and remove all the data before that point. $Time_{Diff}$ is the amount of time for which we have data that we are not interested in. Removing the tail of the data is less complicated once we have removed the head. To do so one simply removes all the rows of data that are longer than the specific conversations mocap file. For example, if the mocap file for conversation 1 for some dyad is 44,000 rows long, and the ratings data is 44,500 rows long, we remove all the rows after the 44,000th row in the ratings data.

After recording their emotional intensity ratings for each conversation the participants then stated how happy, sad etc. they were feeling/the other person was feeling during these conversations. These overall emotion ratings were also recorded in PsychoPy and were extracted during the emotional intensity ratings data processing. Once all the other irrelevant information was removed from the data the emotional intensity ratings were split up into six separate columns. This was chosen to best and most clearly represent the data within the constraints of a tabular data format.

When recording the ratings in PsychoPy participants were intended to be able to move the mouse along the 0-10 scale, however, due to software issues, values outside of this range were recorded. Thus, a step was added in processing to manually restrict the range to 1-10. Any values < 1 were set to 1 and any values > 10 were set to 10. Finally, both the emotional intensity ratings and overall emotion ratings were entered into the master data structure.

5.1.2 Motion Capture Data

The motion capture data is recorded to `.TAK` files using the Motive software (NaturalPoint Inc., Corvallis, OR). The software can export to several formats and for the current project `.csv` was chosen. The reason for this is compatibility with `Python` and `.csv`'s accessibility. The other reason was so that all the data files are in the same format. Once the files are exported from Motive they are around 600MB, which leads to a significant computational cost for importing and exporting. A subset of the data is displayed in the appendix in table 10.2.

The exported files contained extra information in the beginning of each of the files which was changed. The extra information pertained to the exact position or rotation of a given point on the motion capture models body. For example, D21 participant A's shoulder bone rotation on the X axis would be shrunk to "D21.PA.Shoulder_bone.rotation.X." This includes all the necessary information but in a single column name rather than several rows for a single column/data stream. The mocap data also included several rows at the tail of each file for which the timecode was "00:00:00.00," which were removed as they did not indicate real data.

5.1.3 Physiological measures

The physio data is first recorded via the Equivital belts the participants are wearing, which are then fed into the LabChart 8 software (AD Instruments, Colorado Spring, CO). The files are processed in two steps. The first step is a brief cleaning in LabChart which converts one of the ECG channels to a heart rate signal then exporting the data as a `.txt` file. There is a small amount of extra information contained at the beginning of each of the files which is removed in `Python` after being imported. Then the data is down-sampled to 119.88Hz by taking the mean of a section of observations then just keeping the mean value of the section. The data is then at the same frequency as the motion capture and ratings data. The time at the beginning of the

study before the conservations for five minutes is then exported as a baseline reading for each participant and is then removed from the original object. The baseline is the physiological state of the participant when nothing is going on. The data between conversations is removed next leaving us with four conversations of data for each of the measures with columns indicating the time and conversation. Some of the dyads include X, Y, Z acceleration data which is not of interest to the study and was included initially in case of issues with the motion capture data. These measures are removed when adding the physio data to the master data structure. A subset of the data is displayed in the appendix in table 10.3.

Chapter 6

Descriptive Information

We now begin a visual analysis and description of the ratings data. The overall ratings are first characterised, followed by a closer look at a single dyad. Please note that Time presented here on the plots is the frame number at 119.88Hz. So Time 5000 is $\frac{5000}{119.88} = 41.7$ seconds into the trial. A single frame is $\frac{1}{119.88} = 0.008341675$ seconds long.

6.1 Across dyad ratings

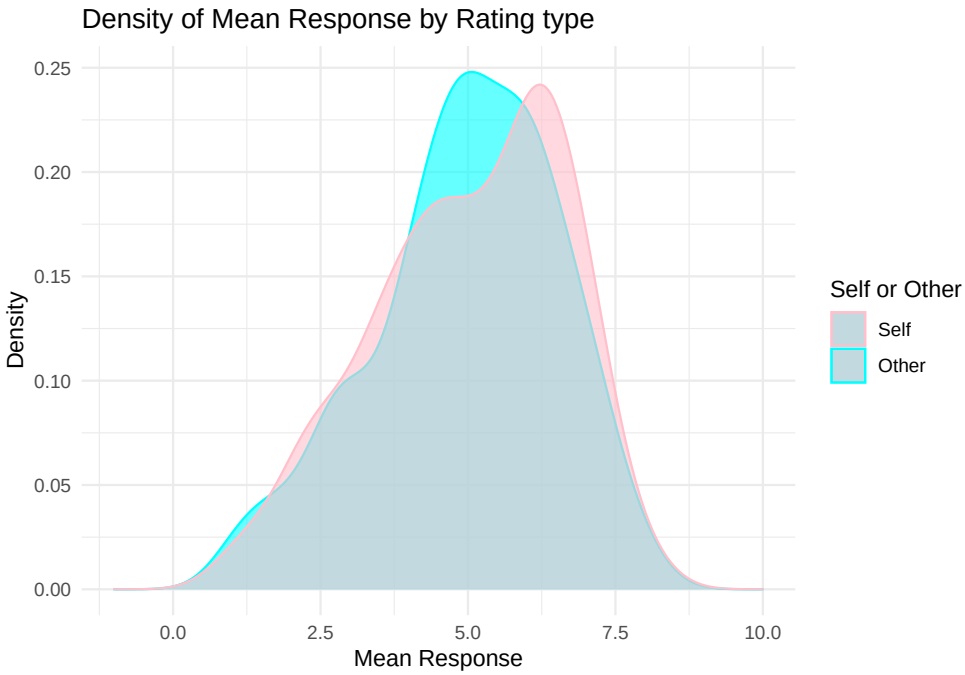


Figure 6.1: Density of mean responses across participants conversations by rating type.

Figure 6.1 shows the density of the mean ratings for Self and Other ratings over conversations across 81 participants. The mean values are taken over the sections of ratings for self and other, ignoring conversation. The plot is intended to show the typical emotional intensity that participants stated they felt and that they thought the other person felt. The Self rating means look approximately normally distributed with a median at around 5. The majority of the density is in the centre as would be expected with a normally distributed variable. The mean for Other ratings seem to be slightly higher with a median just above 6. The density has a small step around 4.5 which then falls off. Overall the distribution looks approximately normal and the two distributions overall look similar.

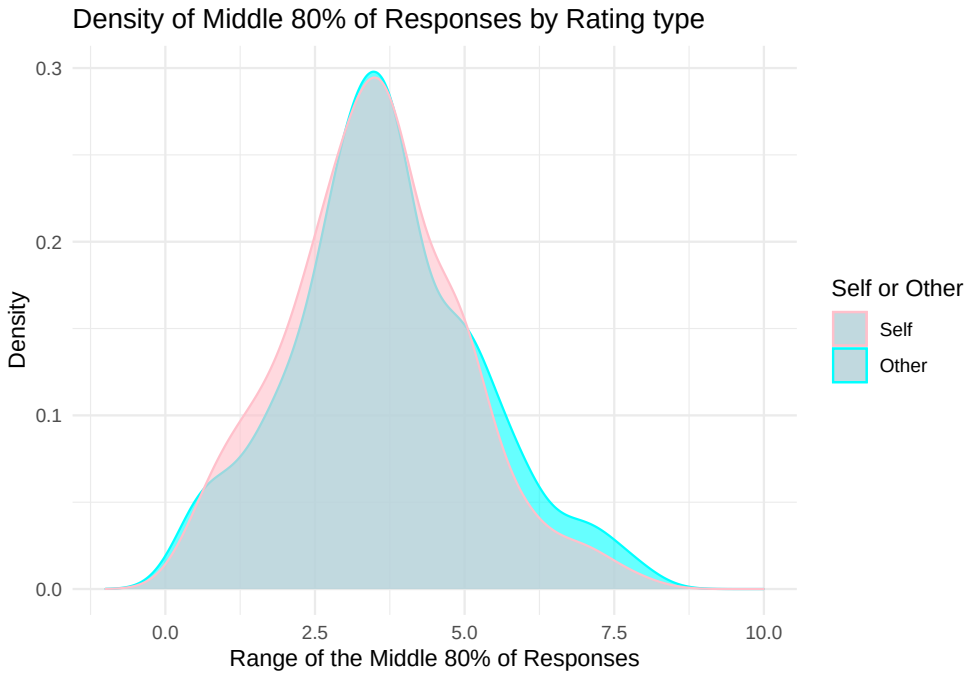


Figure 6.2: Density of the middle 80% of responses given by participants across conversations by rating type.

Figure 6.2 shows the density of the range of the middle 80% of responses for Self and Other ratings across participants. The middle 80% of values is calculated by taking either all the self or other person ratings from a participant and calculating the range between the 10th and 90th percentiles of the ratings. The plot is intended to show how much of the scale participants used depending on whether they were rating how they were feeling or how they think the other person was feeling. We can see that participants rated how they were feeling 80% of the time with a range of around 3.8. The density is concentrated between 1 and 6. This indicating that the range of responses is typically between 1 and 6 when rating how the participant was feeling.

The density of the range of the middle 80% of responses when rating how participants think the other person is feeling has a median or typical range of around 3.8 again. The distribution here has a bit more weight in the tails than the self ratings indicating participants may use slightly more of the

scale when rating how the other person is feeling compared to themselves. They also tend to avoid using none of, or all of the scale most of the time. Overall the plot shows that participants use approximately the same amount of the scale when they are rating how they are feeling as when they are rating how they think the other person is feeling.

6.2 Single dyad ratings

We are now going to investigate how the ratings change within a single dyad. Dyad 21 was arbitrarily selected. The participants in this dyad are strangers who have not met before. The four conversations that the dyad had were with the following prompts:

1. "Describe the best experience of lockdown."
2. "Describe a time you were bullied."
3. "Describe the worst experience of lockdown."
4. "Describe something kind that has been done for you."

These were ordered with "positive," "negative," "negative," "positive," valences for the conversations. The questions were made more relevant to the COVID-19 pandemic lockdown period for the people who participated in the study close to the time around the end of the first major lockdown in New Zealand.

The plots below are time-series plots that show how the ratings are changing over time. The reason these are shown is that the cognitive empathy model developed in the next section is a time-series model and thus the ratings are best shown in line plots. The plots are first shown over the four conversations then more closely examining single conversations. We expect an increase in cognitive empathy over time and across conversations as the two participants become more familiar with each other, with the effect being particularly noticeable as the dyad shown here were strangers. While the focus is primarily on the cognitive empathy plots, a short description of the affective empathy plots is given.

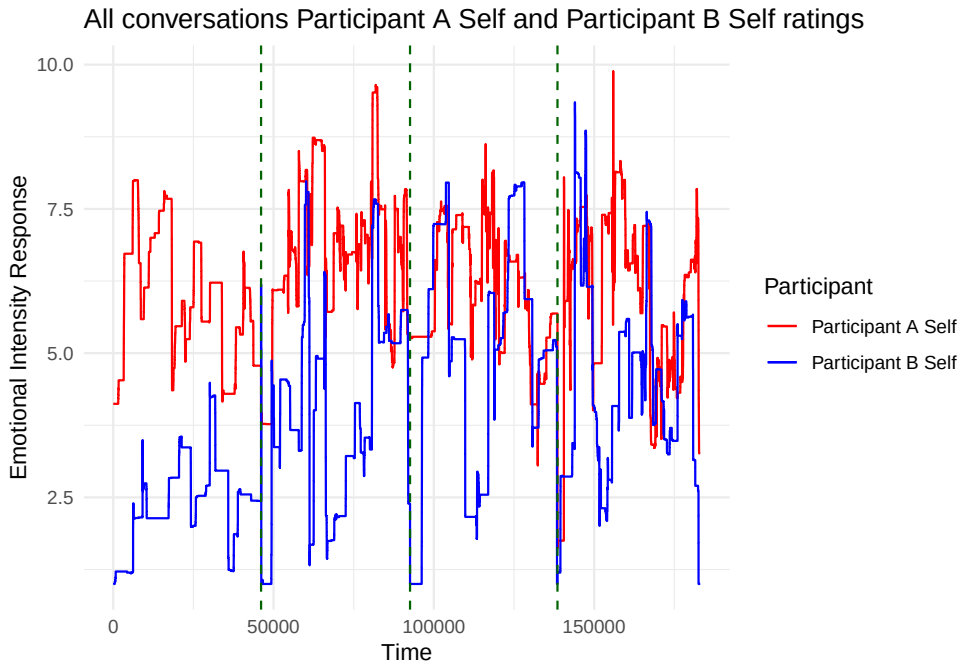


Figure 6.3: Affective Empathy over time across all conversations.

Figure 6.3 shows how Participant A and Participant B reported they were feeling over time across the four conversations for Dyad 21 with vertical dashed lines indicating when the conversations began and ended. We can see that Participant A tends to use values higher on the scale than Participant B. Participant A also seems to change state far more frequently than Participant B who's state seems to be more stable. Therefore, there are differences in how people rate their state. The ratings seem to be increasing/decreasing together, i.e. how one person is feeling is similar to how the other person is feeling. This gives some evidence that one person's state may be reliant on the other person's state which is affective empathy. Finally, the overlap and similarity of the ratings seems to be increasing over the trials. This would indicate that the amount of affective empathy is not stable and does seem to be increasing as the two participants get to know each other better.

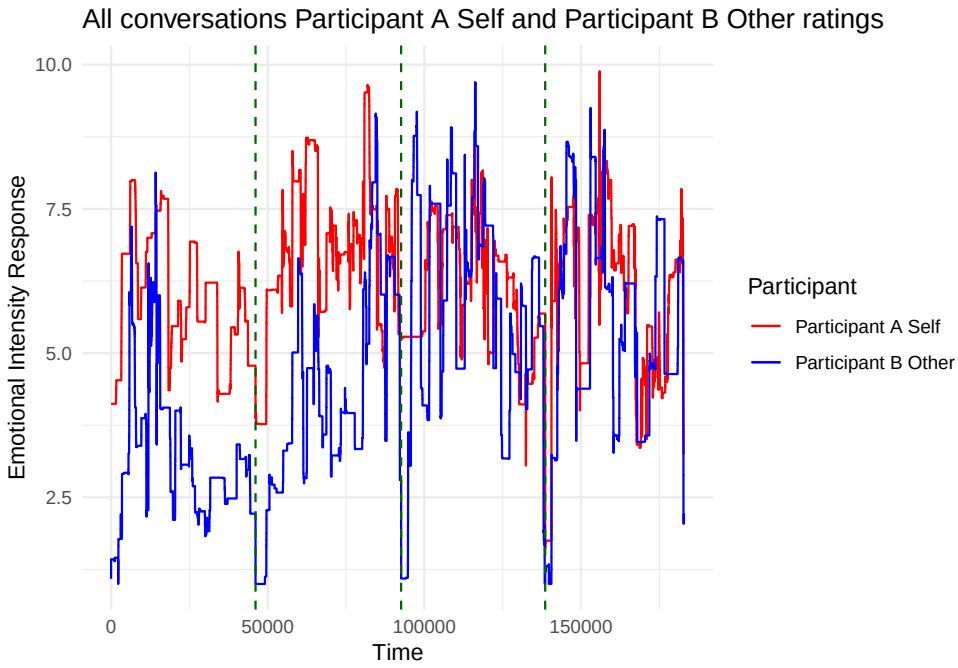


Figure 6.4: Cognitive Empathy over time across all conversations with Participant A as the target.

Figure 6.4 shows cognitive empathy. The plot shows how Participant A stated they felt and how Participant B thought Participant A was feeling across the four conversations. We consider Participant A to be the target here. The difference in ratings seems to be smaller than for the affective empathy plot. This may indicate that when rating another person, the rater takes into account that person's emotional intensity and rates in line with that average emotional intensity. We can next see that while a little lower, Participant B's rating of Participant A seems to be following the same general trend. For example, in the first conversation on the left side of the plot, we can see that when Participant A's rating increases, Participant B's rating of Participant A follows however with a small amount of delay. This provides clear visual evidence that cognitive empathy does seem to be present between the participants. If there was no evidence for cognitive empathy the Participant B's rating of Participant A would be moving around randomly with no clear relationship or change based on the others reported state. We can see as Participant A's emotional intensity increases, so does Participant B's rating of their emotional intensity. Finally, we can see that the overlap or

similarity in the two participants ratings seems to be growing closer together across the trials. The similarity in the ratings increasing is again likely due to the participants grower more familiar with each other and the amount of cognitive empathy between the two participants increasing.

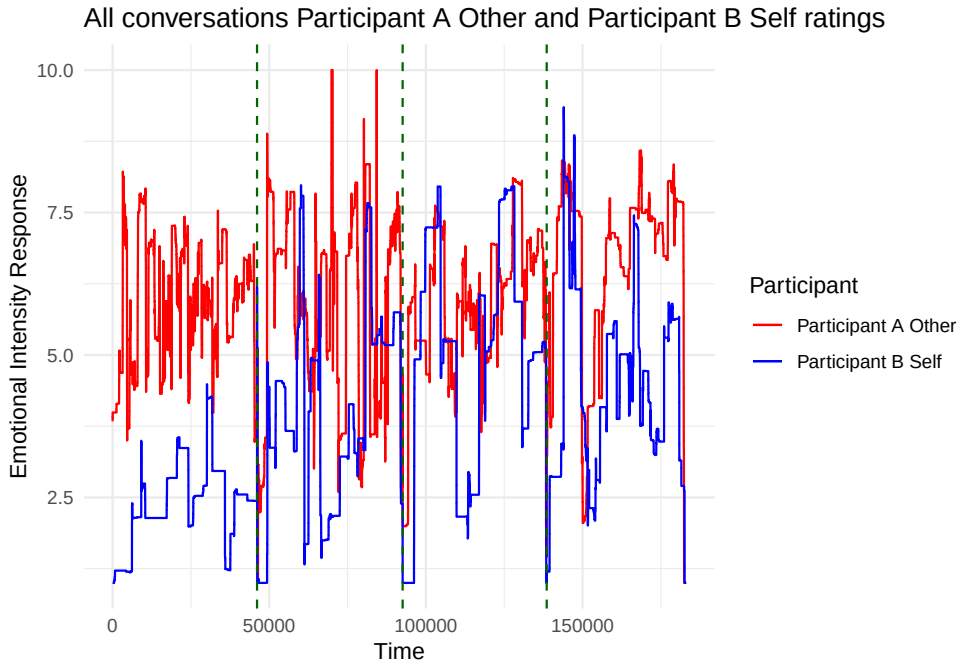


Figure 6.5: Cognitive Empathy over time across all conversations with Participant B as the target.

Figure 6.5 shows cognitive empathy again with Participant A rating how they think Participant B is feeling and Participant B rating how they were feeling across the conversations. Participant B is the target here. We can see much of the same pattern as with the previous plot; as Participant B's rating changes, Participant A's rating of how Participant B is feeling seems to follow. The relationship between the two seems to be a bit weaker than when Participant A is being rated as Participant A's rating of Participant B seems to be changing randomly and rapidly. Overall however towards the end of the conversations, the ratings do seem to be moving together, again with a small amount of delay.

Figure 6.6: Cognitive emapthy over time within the first conversation with Participant A as target.

Figure 6.6 shows how Participant A stated they were feeling and how Participant B thought Participant A was feeling for the first conversation. We can see at time point around 8000 Participant A's emotional intensity increases and after a small delay what Participant B thinks Participant A is feeling follows. This pattern is seen throughout this first conversation. Participant B's rating follows that of Participant A but the standard level seems to be a lower overall emotional intensity. We need to account for this difference in standard level in our model.

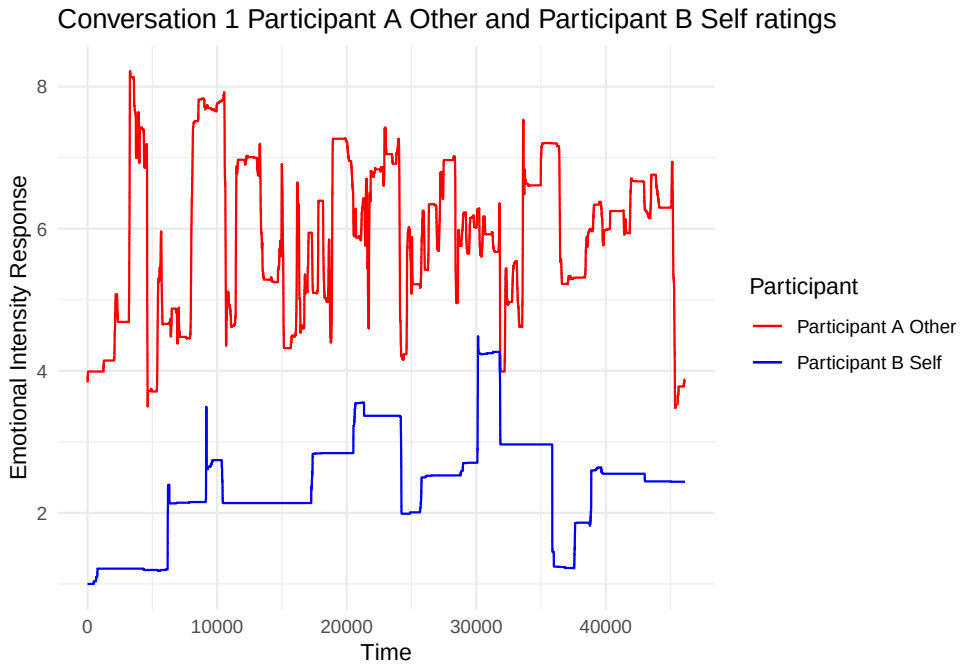


Figure 6.7: Cognitive emapthy over time within the first conversation with Participant B as target.

Figure 6.7 shows how Participant B stated they were feeling and how Participant A thought Participant B was feeling for the first conversation. We can see that Participant A's rating of Participant B follows somewhat closely to Participant B's rating of how they were feeling. Participant B's reported state seems to be very stable compared to Participant A's rating of their state. The amount of cognitive empathy occurring here may be low as the two ratings don't seem to be following each other very closely.

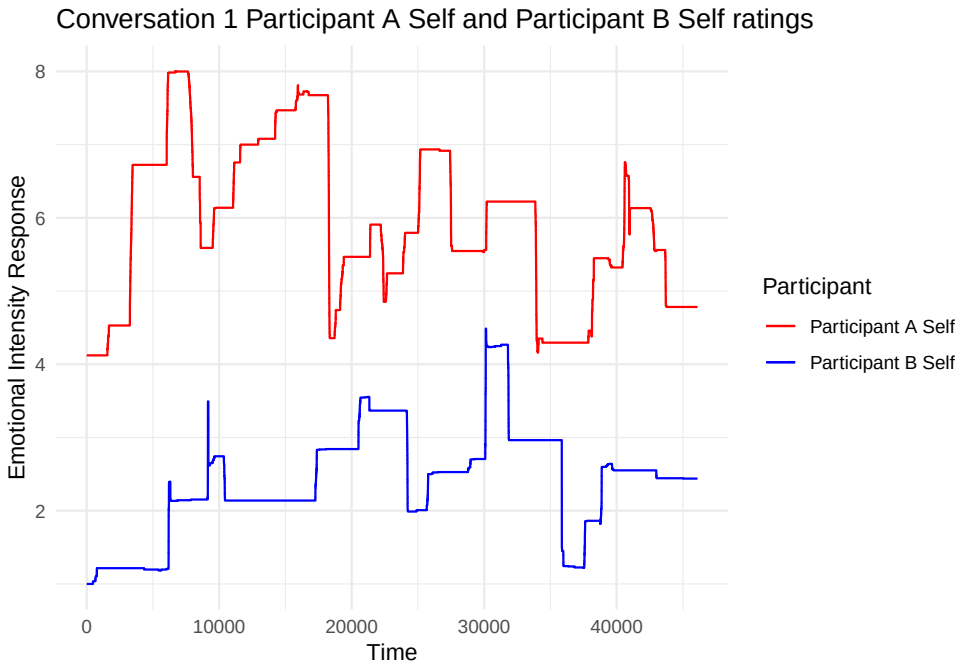


Figure 6.8: Affective emapthy over time within the first conversation.

Figure 6.8 shows how Participant A and Participant B each stated they were feeling for the first conversation. We can see that both participants ratings of their states follow each other reasonably closely. For example, at time point's 9,000 and 30,000 the states are increasing at the same time. There seems to be some evidence for affective empathy here.

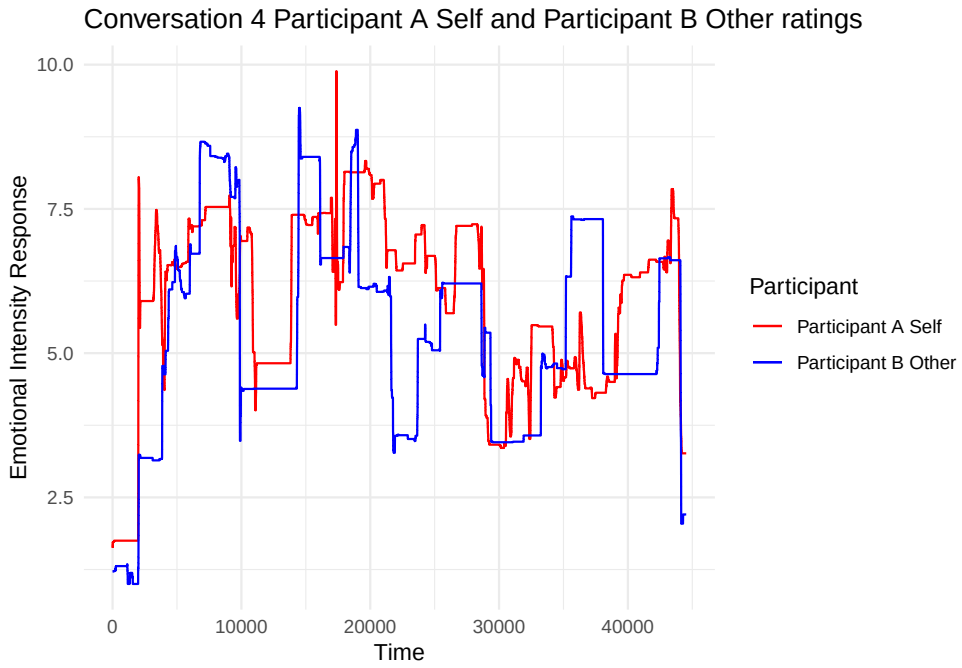


Figure 6.9: Cognitive emapthy over time within the fourth conversation with Participant A as target.

Figure 6.9 shows how Participant A stated they were feeling and how Participant B thought Participant A was feeling for the fourth conversation. We can see the same pattern as the previous plots. Participant B's rating of Participant A changes slightly after Participant A's rating but follows the same general trend. There does seem to be more overlap of values at the fourth conservation with Participant B being closer in how they think Participant A is feeling to how they are actually feeling. Next, we formally test whether our visual analysis holds up in our models.

Chapter 7

Models

The current section outlines the models for affective and cognitive empathy. While a final model was not developed for affective empathy a large amount of work was done before this point which is included. This section is split into the various aspects of modelling the final cognitive empathy model. First, the random walk part of the model is described. Next, the models for affective and cognitive empathy which allow empathy to change over time. Next, we deal with missing data.

Within each participants ratings there are large patches of time in which there was no mouse movement with repeating values. These can be treated one of two ways, firstly keeping the values and treating them as valid responses. This would imply that the participant's state is exactly as the mouse is suggesting. The validity of this is questionable as the participant may have forgotten to move the mouse, therefore, their state stays the same until they remember to move the mouse. Or, some error in the software used for recording led to an unchanging state response. During the processing, the data was also up-sampled which introduced a large amount of these repeating values. However, we may choose to keep the data in because we have no other information about their true state. Removing what information we do have introduces an assumption that their true state is different to their reported state. While this may or may not be the case, there is no way to know for sure. Therefore, removing the data is best for modelling as if they are kept in, we only reduce statistical power and have a poorer fit. Secondly, we can remove the unchanging values and treat them as non-response. This leads to large patches of sequential `NA` values in the data that we need to

model. The algorithm used to remove points starts at the end of the data and checks if $rating_t$ and $rating_{t-1}$ the same, if they are then $rating_t$ is set to NA and $rating_{t-1}$ becomes $rating_t$ and so on (where t is the time point).

The missing points need to be accounted for in our model. In the spaces where we do not have data, we add a linearly increasing amount of error to our previous estimate. The error adding method is known as random-walk. The missing data is expected to be a random-walk from the last known point. The random-walk method works by adding a random amount of error to the last point within the number of time steps where data is missing. The random walk section details further what the method looks like.

Next, we need to be able to get parameter estimates from our models. With distributions such as the Poisson distribution and the Normal distribution, this is simple as they have closed-form estimators of values such as the mean and variance. The models we are building do not however have easily derivable solutions for the mean and variance and other estimators. Therefore, we move from deriving these by hand, as with Poisson and Normal distributions, to computational methods of deriving these parameters. Once we have our model we are wanting to get parameter estimates of, we optimise the log-likelihood of the distribution using functions in `R` such as `optim()` and `optimise()`. Before we can optimise the log-likelihood, we first have to derive it by hand. Then we pass the log-likelihood function to `optim()` and give it data generated from the specified distribution and let it optimise the function. We use the L-BFGS-B method to optimise the functions (Byrd et al., 1995). Once the function has been optimised we can get the standard errors using the Hessian matrix and build confidence intervals associated with our estimates.

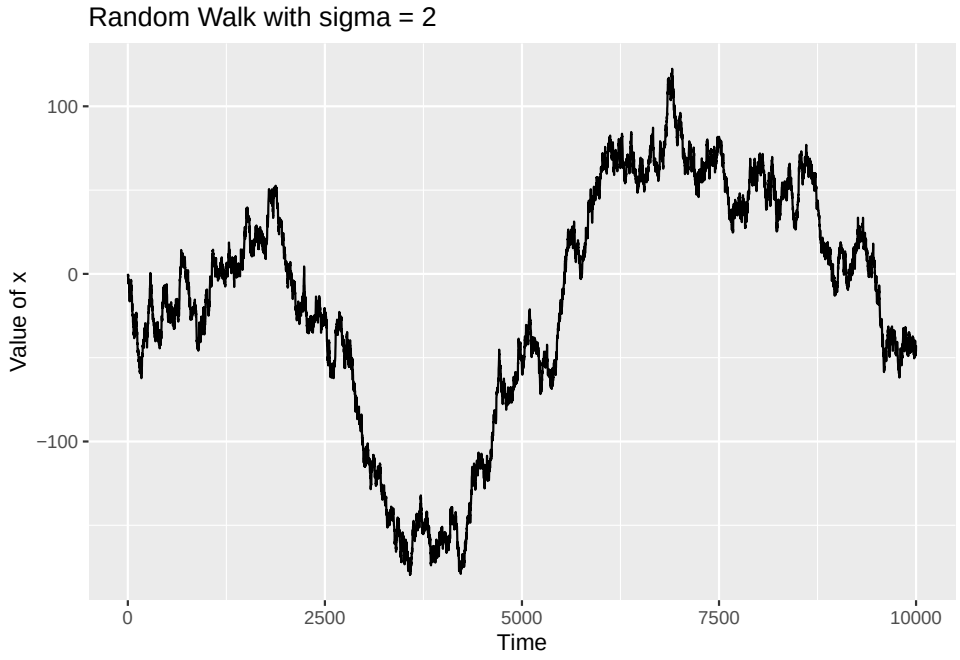
We finally test the model with real data to verify that the model works both with randomly generated data and also data representing what we believe to be empathy, the emotional intensity ratings data. The data is split into Self and Other sets and is then entered into `optim()` which then provides our parameter estimates. The estimates are then sanity checked to ensure that are approximately what we would expect using data generated from the specific model.

7.1 Modelling Random Walk

7.1.1 Varying σ

We begin the theoretical models by simply modelling a standard random walk model where each observation at time $_t$ is simply the previous observation with some random error added on. A plot of this is given below where the random error is normally distributed with $\mu = 0$ and $\sigma = 2$. We can see that the observations are moving up for the first ~3000 observations then down again, then back up in a non-linear seeming fashion. These movements are the result of the "walk" that the observations are doing. The models below show first the basic model where the current observation is the last observation plus some error. The second model then allows for missing data where the current observation is the last known observation plus some error multiplied by the k missing observations before the current observation.

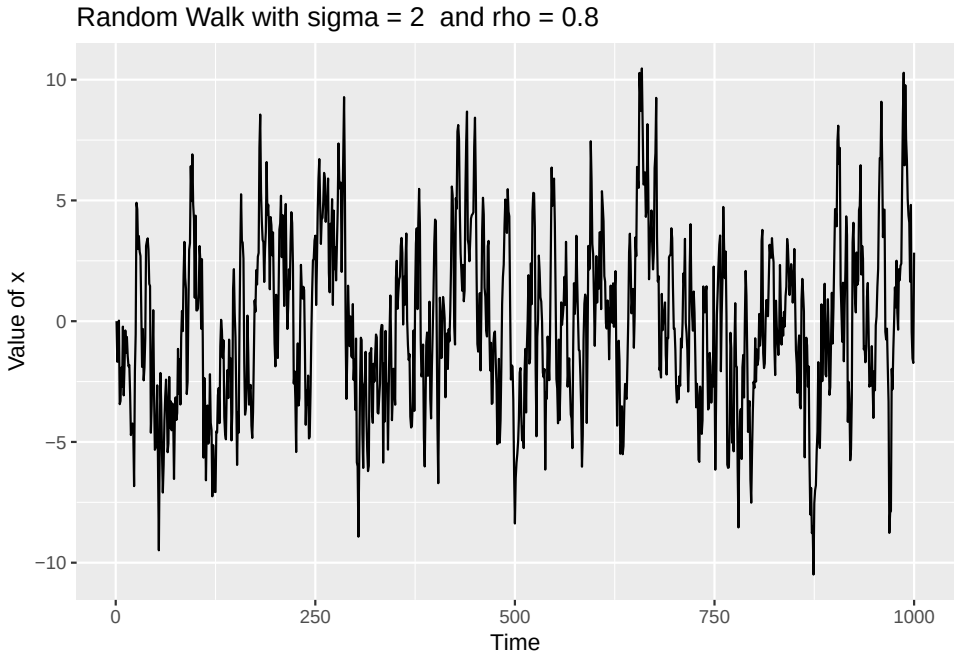
$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_t, & \varepsilon_t &\sim N(0, \sigma^2) \\ x_t &= x_{t-k} + \upsilon_{t|k}, & \upsilon_{t|k} &\sim N(0, k\sigma^2) \end{aligned}$$



7.1.2 Varying σ and ρ

We next added ρ to the model to serve as a measure of autocorrelation, or how strongly the current observation relies on the previous. The function of ρ here essentially indicates how strongly the observations are being brought back towards the mean value which is 0 in this case. The plot below shows a random walk where $\sigma = 2$ and $\rho = 0.8$, $\mu = 0$ here. We can see that the values are moving around far less than in the plot above and seem to be mostly around the mean of 0. There is far less walk than was seen in the previous plot.

$$x_t = \rho x_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \sigma^2)$$

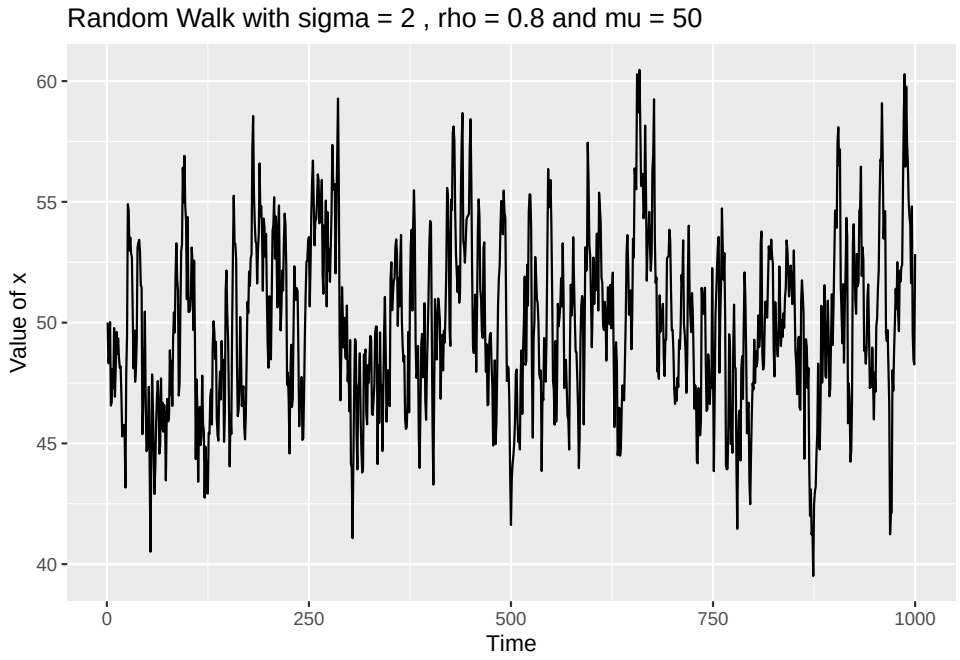


7.1.3 Varying σ, ρ and μ

$$x_t = \mu + \rho\,(x_{t-1} - \mu) + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

We next add μ to the model to change where the data is centred based on a value that is not necessarily 0. Previously we had $\mu = 0$ as is the default, now $\mu = 50$ to test starting at a different value. We keep $\sigma = 2$ and $\rho = 0.8$. We can see the plot indicates that the observations falling around the μ of 50

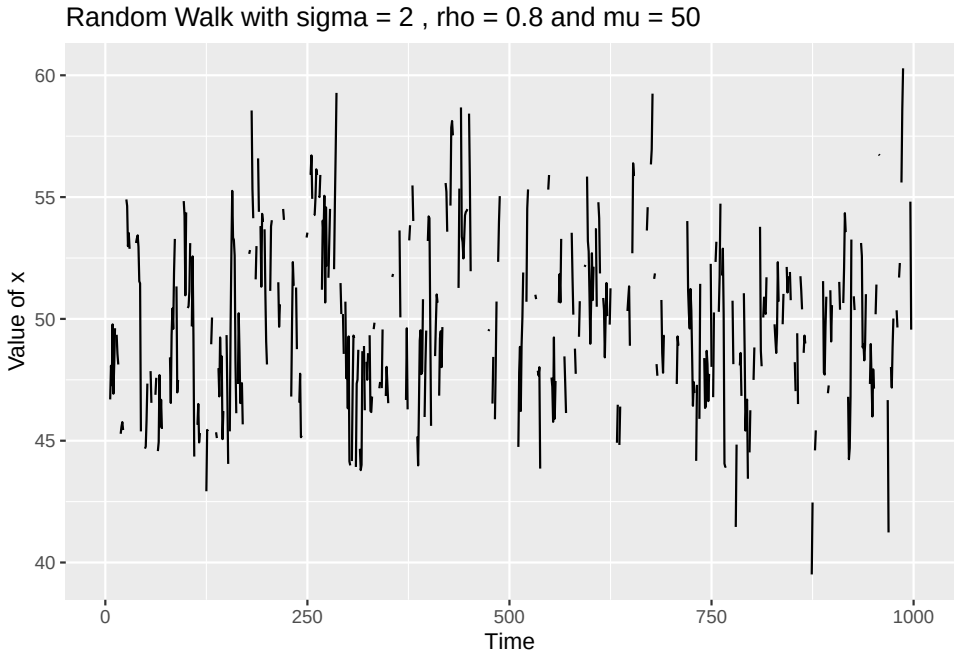
and one can imagine if ρ was decreased to 0.6 for example, the observations would fall far more closely around 50 with little variation.



7.1.4 Varying σ, ρ and μ with missing data

$$x_t = \mu + \rho\,(x_{t-k} - \mu) + \varepsilon_{t|k}, \quad \varepsilon_{t|k} \overset{iid}{\sim} N(0, k\sigma^2)$$

The model below is with data removed which can be seen in the sporadic gaps in the data. We kept $\sigma = 2$, $\rho = 0.8$ and $\mu = 50$ here with 40% of the observations removed. We can see that the data is mostly around the μ of 50 and randomly removing the data doesn't affect the overall trend as it is still moving up and down as it was previously.



7.2 Modelling Affective Empathy

The model below is that developed for affective empathy. This model is still a work in progress as it is out of the scope of the current project. If affective empathy is occurring it would follow the form $z_{At} = z_{Bt}$ meaning that Participant A's true state is exactly the same as Participant B's true state, at some point in time. However, we cannot know Participant A and Participant B's true states, z_{At} & z_{Bt}, and therefore need to use our participant's ratings y_{AAt} and y_{BBt} to model these.

$$
\begin{aligned}
z_{At} &= \mu_A + \rho_{AA} z_{At-1} + \rho_{AB} z_{Bt-1} + \varepsilon_{At} \\
&\text{then,} \\
y_{AAt} &= \mu_{AA} + \beta_{AA} z_{At} + \varepsilon_{AAt} \\
where \;\; t &= 1 \ldots T, \; \varepsilon_{AAt} \sim N\left(0, \; \sigma^2_{AAt}\right)
\end{aligned}
$$

The second model above can be expressed for Participant B by simply changing the A's to B's. We can see then the second model states that participant A's rating of how they are feeling, y_{AAt}, is made up of several components. We have μ_{AA} which is the standard emotional intensity rating

when nothing is going on. Then $\beta_{AA} z_{At}$ is the how accurate a participant is at rating how they are truly feeling as if β_{AAt} is 0, their rating is just random error and if it is larger their rating is based off their true state. The model has a number more parts that need to be added before it is ready to be applied to real data.

7.3 Modelling Cognitive Empathy

Cognitive empathy refers to one individual being able to correctly know the state of another individual. If cognitive empathy was occurring with Participant B correctly knowing Participant A's state it would follow the form $y_{BA_t} = \beta_{BA} z_{A_t}$ where $\beta_{BA} = 1$. Where y_{BA_t} is Participant B's rating of Participant A's true state at time $_t$. Participant A's true state is expressed as z_{A_t} and β_{BA} is the accuracy of Participant B's rating of Participant A's true state. We cannot measure z_A or z_B, Participant A and B's true states, instead, we use each participant's ratings of how they thought themselves or the other person was feeling. These ratings have an amount of uncertainly and measurement error associated with them.

We consider Participant A to be the target here (the person who's state we are guessing/is reporting their state). Our model is built using a series of steps. First we state that Participant A's rating of their state is y_{AAt} and Participant B's rating of Participant A is y_{BAt}. Next, we next model how Participant A's rated internal state is changing over time by using an autoregressive time series model which has the form

$$\begin{aligned} y_{AAt} &= \mu_A + \rho_A(y_{AAt-1} - \mu_A) + \varepsilon_{At} \\ \varepsilon_{At} &\sim N(0, \sigma_k^2), \\ t &= 1, \ldots, T \end{aligned} \tag{1}$$

Then we know Participant A's reported state is their standard rating μ_A plus some amount of autocorrelation back to the difference between the standard state and the previous observation $\rho_A(y_{AAt-1} - \mu_A)$. Now given that we have a time series set of data $\{y_{AAt}\}_{t=1}^{T}$ we can estimate μ_A, ρ_A and σ_A. This is complicated by missing data but we are still able to estimate the parameters by noting that (1) implies

$$y_{AAt} = \mu_A + \rho_A^k(y_{AAt-1} - \mu_A) + \sum_{\ell=0}^{k-1} \rho^\ell \varepsilon_{At-\ell}$$

From this the cognitive empathy model has Participant B observing Participant A's state

$$y_{BAt} = \mu_B + \beta_{BA}\left(y_{AAt-1} - \mu_A\right) + \varepsilon_{Bt}$$
$$\varepsilon_{Bt} \overset{iid}{\sim} N(0, \sigma_B^2)$$

Where μ_B is Participant B's standard rating and β_{BA} how accurate Participant B is at rating Participant A's state. We can then estimate μ_B, β_{BA} and σ_B directly assuming we have no missing data. However, if we are to do this we need to know μ_A, ρ_A and σ_A. We could estimate these all simultaneously but for this investigation, we take our estimates from the autoregressive model as known parameters.

7.4 Building Components for Empathy Models

Given that we can now vary σ, ρ and μ with and without missing data, we can now move onto estimating the values of these parameters via randomly generated data. This process is necessary as previously we randomly generated the data knowing the values of these parameters. However, moving towards the cognitive empathy model now, we need to derive these as they are unknown and cannot be derived by hand.

7.4.1 Normal Distribution

We set our values for μ and σ to 7 and 2 respectively then verify can return the correct value. We see the likelihood function for the Normal Distribution below

$$f\left(y_i|, \mu, \sigma^2\right) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{\frac{(y_i-\mu)}{2\sigma^2}}$$
$$\ell\left(y_i|, \mu, \sigma^2\right) = -\frac{n}{2}\log(2\pi\sigma^2) \; - \frac{1}{2\sigma^2} \; \sum_{i=1}^{n} (y_i - \mu)^2$$

We are able to correctly return μ (6.99) and σ (1.98) with standard errors of 0.02 and 0.01, respectively. We have now verified that our estimation method works for both the Normal Distribution but also multi-parameter models. We next move onto our simplest model in which we are estimating ρ and σ.

7.4.2 Random Walk

The likelihood function for the model including ρ and σ has the form shown below. We are setting ρ to 0.9 and σ to 2. We randomly generate data using the model below

$$x_t = \rho x_{t-1} + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

We then optimise the log-likelihood below to get our parameter estimates

$$f\left(y_i|y_{i-1}, \rho, \sigma^2\right) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(y_i - \rho y_{i-1})^2}{2\sigma^2}}$$
$$\ell\left(y_i|y_{i-1}, \rho, \sigma^2\right) = -\frac{n-1}{2}\log(2\pi\sigma^2)\ - \frac{1}{2\sigma^2}\ \sum_{i=2}^{n}(y_i - \rho y_{i-1})^2$$

Both ρ and σ are correctly estimated as 0.9 and 1.99 respectively. These parameters have standard errors of 0.05 and 0.01, respectively. We now move on and add μ into the model so that can be varied as well.

7.4.3 Centred Autoregressive Process

We now add μ into the model setting it to 10 , and keeping ρ as 0.9 and σ to 2. We randomly generate data from the model including μ, ρ and σ as is shown below

$$x_t - \mu = \rho_1(x_{t-1} - \mu) + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

The model has the log-likelihood shown below which we then optimise

$$f\left(y_i|y_{i-1}, \mu, \rho, \sigma^2\right) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{\left((y_i - \mu) - \rho(y_{i-1} - \mu)\right)^2}{2\sigma^2}}$$
$$\ell\left(y_i|y_{i-1}, \mu, \rho, \sigma^2,\right) = -\frac{n-1}{2}\log(2\pi\sigma^2)\ - \frac{1}{2\sigma^2}\ \sum_{i=2}^{n}\left((y_i - \mu) - \rho(y_{i-1} - \mu)\right)^2$$

Once the log-likelihood shown above has been optimised we can verify that we get the correct estimates of our parameters. We found μ is close to 10 at 10.1 with a standard error of 0.19. The same can be said for σ which we specified as 2 and returned a value of 1.99 with a standard error of 0.01.

Finally, we verify our estimates of ρ are close to 0.9 which we specified earlier as the value returned is 0.9 with a standard error of 0.05.

7.4.4 Centred Autoregressive Process with Missing Data

The final step before moving onto the models that are ready for real data is to ensure that we can deal with missing data. We are therefore randomly removing selected parts of the data and checking that we still return the correct values. The likelihood and log-likelihood now include a new term, $\frac{1-\rho^{2k_i}}{1-\rho^2}$, which accounts for the k_i steps that are taken with missing data. Where k_i is the kth step for the ith observation. If there is no missing data the term is ignored, but otherwise, it linearly increases the amount of variance and uncertainty about where the next step should be. The model that is used to randomly generate the data is the same as above except μ is constrained to 0 and is therefore dropped from the likelihood. We have set ρ to 0.7 and σ to 2 once again.

$$y_t - \mu = \rho^k(y_{t-k} - \mu) + \varepsilon_{t|k}, \quad \varepsilon_{t|k} \sim N\left(0, \frac{1-\rho^{2k_t}}{1-\rho^2}\sigma^2\right)$$

The log-likelihood is given below and is optimised to find our parameters

$$\begin{aligned}\ell(y_i|y_{i-1}, \rho, \sigma^2, \mu) = &-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log\left(\frac{1-\rho^{2k_i}}{1-\rho^2}\right) \\ &- \frac{1}{2\sigma^2}\left(\frac{1-\rho^2}{1-\rho^{2k_i}}\right)\left(y_i - \mu - \rho^{k_i}(y_{i-1}-\mu)\right)^2\end{aligned}$$

Then

$$\begin{aligned}\ell(Y|\rho, \sigma^2, \mu) = &-\frac{n-1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{t=2}^{n}\log\left(\frac{1-\rho^{2k_i}}{1-\rho^2}\right) \\ &- \frac{1}{2\sigma^2}\sum_{i=2}^{n}\left(\frac{1-\rho^2}{1-\rho^{2k_i}}\right)\left(y_i - \mu - \rho^{k_i}(y_{i-1}-\mu)\right)^2\end{aligned}$$

Note:

y_i is at time t_i

y_{i-1} is at time t_{i-1}

and $k_i = t_i - t_{i-1}$

Our returned value of σ is just the value of 2 we specified initially at 2.02 with a standard error of 0.04. Our value of ρ is also close to the value of 0.7 we specified initially at 0.69 with a standard error of 0.04.

7.5 Empathy Models

We have now completed building and testing all the separate parts used to build our model of cognitive empathy. We move on to test our models with real data. These models are first tested with randomly generated data, with real data with repeated measurements included, finally with real data with repeated measurements removed. We expect throughout that the generated data with and without data removed will perform best with the lowest standard errors as the data is generated using the exact expected model. We then expect that the real data with repeating data removed will perform next best with the next largest standard errors. The reason for this is that the repeating data impairs our ability to view the underlying phenomena which we expect to follow our specified model. Therefore, if we remove this repeating data we will return better estimates. The contrary is therefore true for a model with the repeating data kept in. We now see whether these hypotheses are supported.

7.5.1 Autoregressive Self-Rating Model

The first model we discuss here is the autoregressive model from the emotional intensity ratings. It is how each person states they themselves were feeling. The model below is needed for the following model which is the cognitive empathy model. The autoregressive model is run on three different datasets. The first data randomly generated using the model below which has subscripts "$_1$" to indicate it is person 1/the target's rating

$$x_t - \mu_1 = \rho_1(x_{t-1} - \mu_1) + \varepsilon_t^1, \quad \varepsilon_t^1 \overset{iid}{\sim} N(0, \sigma_1^2)$$

With the log-likelihood given below

$$\ell(y_i|\mu_1,\rho_1,\sigma_1) = -\frac{1}{2}\log(2\pi\sigma_1^2) - \frac{1}{2}\log\left(\frac{1-\rho_1^{2d_i}}{1-\rho_1^2}\right) - \frac{1}{2\sigma_1^2}\left(\frac{1-\rho_1^2}{1-\rho_1^{2d_i}}\right)(y_i - \mu_1 - \rho_1^{d_i}(y_{i-1}-\mu_1))^2$$

$$\ell(Y|\mu_1,\rho_1,\sigma_1) = -\frac{(n-1)}{2}\log(2\pi\sigma_1^2) - \frac{1}{2}\sum_{i=2}^{n}\log\left(\frac{1-\rho_1^{2d_i}}{1-\rho_1^2}\right) - \frac{1}{2\sigma_1^2}\sum_{i=2}^{n}\left(\frac{1-\rho_1^2}{1-\rho_1^{2d_i}}\right)(y_i - \mu_1 - \rho_1^{d_i}(y_{i-1}-\mu_1))^2$$

Note:

y_i is at time t_i

y_{i-1} is at time t_{i-1}

and $d_i = t_i - t_{i-1}$

$t = 1, \ldots, T$

The next dataset is the emotional intensity ratings stated by each person of how they were feeling with the repeating data included. These sections of data are then removed in the last dataset that we look at.

7.5.1.1 Generated Data

Table 7.1: Estimates and Standard Errors for Generated Data

	μ	ρ	σ
Parameters	50.0000	0.5000	2.0000
Estimates	50.0413	0.5039	1.9923
Standard Errors	0.0402	0.0086	0.0141

Table 7.1 shows the estimates for each of the parameters from the randomly generated data. We would expect this data set to have the lowest amount of error and the most accurate estimates of the three data types. We can see that the value for μ is close to 50, the value we got was 50.04. The standard error for μ is also quite small at 0.04. We specified σ to be 2 and that exactly what we returned at 1.99 with a standard error of 0.01. The same can be said for ρ which was specified to be 0.5 and we returned 0.5 with a standard error of 0.01.

Table 7.2: Estimates and Standard Errors for Generated Data with Data Removed

	μ	ρ	σ
Parameters	50.0000	0.5000	2.0000
Estimates	50.0317	0.5019	1.5925
Standard Errors	0.0384	0.0089	0.0113

Table 7.2 shows that we able to return our specified parameters even when we have a large amount of data removed. We had 35.93% of the values removed as NA here.

7.5.1.2 Real Data with Repeated Ratings

Table 7.3: Estimates and Standard Errors including repeating data for Participant A Rating Self

	μ	ρ	σ
Estimates			
Conversation 1	6.3671	1.0000	0.0053
Conversation 2	7.1662	0.9999	0.0096
Conversation 3	6.1599	1.0000	0.0084
Conversation 4	6.3275	0.9999	0.0146
Standard Errors			
Conversation 1	0.8757	0.0000	0.0000
Conversation 2	0.6942	0.0000	0.0000
Conversation 3	1.0137	0.0000	0.0000
Conversation 4	0.7326	0.0000	0.0000

Table 7.3 shows the estimates and standard errors for μ_1, ρ_1, and σ_1 for Participant A including the data we consider to be missing. Overall we can see that μ_1 is reasonably stable except for Conversation 2 which seems to have a higher resting emotion intensity than the other conversations. The values for ρ_1 are almost all exactly 1 which means that the amount of pullback to the resting emotional state is low and the persons emotional state is driven strongly by how they state they are feeling. The amount of error or change in the ratings, σ_1, seems to increase across the conversations. This may suggest

that people are using more of the scale or possibly that they are less certain about how they were feeling, both increasing the amount of variability in the ratings. The standard errors for both ρ_1 and σ_1 here and for Participant B with and without repeating data are both very small because of the needed adjustment to account for the log and logit transforms.

The standard errors for σ_1 are calculated as

$$\text{SE}(\sigma_1) = \sigma_1 \times \text{SE}(\log\sigma_1)$$

The standard errors for ρ_1 are calculated as

$$\text{SE}(\rho_1) = \rho_1(1-\rho_1) \times \text{SE}(\text{logit}(\rho_1))$$

Therefore, we can see that the standard errors will be very low for both parameters which explains why they are showing up as 0's in the table.

Table 7.4: Estimates and Standard Errors including repeating data for Participant B Rating Self

	μ	ρ	σ
Estimates			
Conversation 1	2.7754	0.9999	0.0065
Conversation 2	2.0553	1.0000	0.0137
Conversation 3	4.9483	1.0000	0.0105
Conversation 4	4.4485	1.0000	0.0132
Standard Errors			
Conversation 1	0.4357	0.0000	0.0000
Conversation 2	2.2156	0.0000	0.0000
Conversation 3	3.6706	0.0000	0.0000
Conversation 4	2.3203	0.0000	0.0000

Table 7.4 shows the estimates and standard errors for μ_1, ρ_1, and σ_1 for Participant B including the data we consider to be missing. The estimates for μ_1 seem to be increasing over the conversations, meaning the resting emotional intensity for Participant B seems to be increasing. The μ_1's here seem to be a bit lower for Participant B than Participant A. This indicates Participant B's resting emotional intensity is lower than that for Participant

A. Individual differences like this are expected as people rate how they are feeling differently. The values for ρ_1 are about the same as with Participant A, all around 1. The estimates for σ_1 do not seem to be changing hugely from Conversation 2 onwards, but there is a large jump from Conversation 1 to Conversation 2. This may be due to Participant B getting used to the scale. The standard errors for μ_1 are much larger for Conversations 2-4 which fits with the σ_1 values also being quite large. Participant B may use a larger amount of the scale which would lead their resting emotional intensity to be more variable and have a larger standard error and larger σ_1, amount of change.

7.5.1.3 Real Data without Repeated Ratings

Table 7.5: Estimates and Standard Errors removing repeating data for Participant A Rating Self

	μ	ρ	σ
Estimates			
Conversation 1	6.1093	1.0000	0.0035
Conversation 2	6.7817	1.0000	0.0066
Conversation 3	6.1205	1.0000	0.0056
Conversation 4	5.8566	0.9999	0.0115
Standard Errors			
Conversation 1	1.3886	0.0000	0.0000
Conversation 2	1.2671	0.0000	0.0000
Conversation 3	0.9414	0.0000	0.0000
Conversation 4	1.0775	0.0000	0.0000

Table 7.5 shows the estimates and standard errors for μ_1, ρ_1, and σ_1 for Participant A after removing the data we consider to be missing. We would expect these estimates to be somewhat closer to the unknown, but true values of μ_1, ρ_1, and σ_1 with lower standard errors as we have removed the missing, repeating data. We can see that all our estimates are approximately the same as those found by using the repeating data. We can't comment on the standard errors for σ_1 and ρ_1 as they are 0 but are they are overall smaller for μ_1 as expected.

Table 7.6 shows the estimates and standard errors for μ_1, ρ_1, and σ_1 for Participant B after removing the data we consider to be missing. We see the

Table 7.6: Estimates and Standard Errors including repeating data for Participant B Rating Self

	μ	ρ	σ
Estimates			
Conversation 1	2.3511	1.0000	0.0043
Conversation 2	3.6449	0.9999	0.0131
Conversation 3	5.8943	1.0000	0.0074
Conversation 4	4.5481	1.0000	0.0092
Standard Errors			
Conversation 1	0.5391	0.0000	0.0000
Conversation 2	0.8440	0.0000	0.0000
Conversation 3	6.7282	0.0000	0.0000
Conversation 4	1.3446	0.0000	0.0000

same pattern here as with Participant A where the estimates are about the same as when including the repeating data with the standard errors being smaller for μ_1. Therefore, we can see that when removing the repeating data we can have more certainty about our estimates due to lower standard errors for μ_1. The data with repeating values removed has standard errors slightly more in-line with the randomly generated data however they are still not quite as small overall for μ_1.

7.5.2 Model of Cognitive Empathy

We next move onto our model of Cognitive Empathy. Cognitive empathy is one persons ability to know what another person is feeling. This gives rise to our model below that is based on y_t which is how the target person states they are actually feeling which comes from x_t which is the rater's ability to state what the other person is feeling. The model for the targets rating of themselves is essentially the same as the autoregressive model but with extra indicators, then the second model is how the other person rates the target. The subscripts "$_2$" indicate it is the raters rating of the target

$$x_t - \mu_1 = \rho_1(x_{t-1} - \mu_1) + \varepsilon_t^{(1)}, \quad \varepsilon_t^{(1)} \overset{iid}{\sim} N(0, \sigma_1^2)$$
$$y_t - \mu_2 = \beta_2(x_{t-1} - \mu_1) + \varepsilon_t^{(2)}, \quad \varepsilon_t^{(2)} \overset{iid}{\sim} N(0, \sigma_2^2)$$

The models have the log-likelihood given by

$$
\begin{aligned}
&\text{Given data } \{(t_{x,i}, x_i)\}_{i=1}^{n_x} \text{ and } \{(t_{y,i}, y_i)\}_{i=1}^{n_x} \text{ we define} \\
&\quad d_{x,i} = t_{x,i} - t_{x,i-1} \\
&\quad p_i = max\{j \in \{1, \ldots, n_x\} : t_{x,j} < t_{y,i}\} \\
&\quad d_{y,i} = t_{y,i} - t_{x,p_i} \\
&\quad \upsilon_i = \sigma_2^2 + \mathrm{I}(d_{y,i} > 1)\beta_2^2\sigma_1^2 \left(\frac{1 - \rho_1^{d_{y,i}-1}}{1 - \rho_1^2} \right)
\end{aligned}
$$

Then the log likelihood is

$$
\begin{aligned}
\ell(\mu_1, \rho_1, \sigma_1, \mu_2, \beta_2, \sigma_2) = &-\frac{1}{2}(n_x - 1)\log(2\pi\sigma_1^2) - \frac{1}{2}\sum_{i=2}^{n_x} \log\left(\frac{1 - \rho_1^{2d_{x,i}}}{1 - \rho_1^2} \right) \\
&-\frac{1}{2\sigma_1^2}\sum_{i=2}^{n_x} \left(\frac{1 - \rho_1^2}{1 - \rho_1^{2d_{x,i}}} \right) \left(x_i - \mu_1 - \rho_1^{d_{x,i}}(x_{i-1} - \mu_1) \right)^2 \\
&-\frac{1}{2}\sum_{i=2}^{n_y} \left[\log(2\pi\upsilon_i) + \frac{(y_i - \mu_2 - \beta_2\rho_1^{d_{y,i}-1}(x_{\rho_i} - \mu_1))^2}{\upsilon_i} \right] \\
&t = 1, \ldots, T
\end{aligned}
$$

The indicator function $\mathrm{I}(d_{y,i} > 1)$ returns 1 if the given time point has a missing value and 0 otherwise. It is used for measuring missingness and adds error scaling by the number of missing values. So the more missing values we have the more uncertainty or error we have.

7.5.2.1 Generated Data

Table 7.7: Estimates and Standard Errors for Generated Data without Data Removed

	σ_1	σ_2	ρ_1	μ_1	μ_2	β_2
Parameters	2.0000	4.0000	0.7000	10.0000	50.0000	0.5000
Estimates	1.9922	3.9863	0.7044	10.0686	50.0163	0.4784
Standard Errors	0.0141	0.0282	0.0071	0.0674	0.0513	0.0142

Table **??** shows our cognitive empathy model with μ_1, μ_2, ρ_1, σ_1, σ_2 and β_2. We can also see the true parameter values that the data was generated

with and see we can return estimates that are very close to the original parameters. We used 10000 randomly generated observations here and use the same amount below with the seed set to the same value for both runs. We did not remove any data here, this is done next to verify our model also works with missing data. An example dataset is given in the appendix for closer inspection in 10.4.

Table 7.8: Estimates and Standard Errors for Generated Data with Data Removed

	σ_1	σ_2	ρ_1	μ_1	μ_2	β_2
Parameters	2.0000	4.0000	0.7000	10.0000	50.0000	0.5000
Estimates	1.9872	3.9621	0.7073	10.0372	49.9916	0.4752
Standard Errors	0.0149	0.0299	0.0074	0.0689	0.0518	0.0155

Table **??** shows when running the model on randomly generated data we can find these estimates exactly with reasonably small standard errors even when randomly removing the majority of the data. We had 10.56% of the values removed as NA here for the "target" rating and 10.39% of the values removed for the "rater" rating.

7.5.2.2 Real Data with Repeated Ratings

Table 7.9: Estimates and Standard Errors with repeating data for Participant A Self, Participant B Other

	σ_2	μ_2	β_2
Estimates			
Conversation 1	0.9164	3.5363	0.7221
Conversation 2	1.5217	4.5538	0.6705
Conversation 3	1.8471	5.8005	0.5625
Conversation 4	1.4168	5.6416	0.7404
Standard Errors			
Conversation 1	0.0030	0.0046	0.0039
Conversation 2	0.0050	0.0074	0.0054
Conversation 3	0.0061	0.0087	0.0084
Conversation 4	0.0047	0.0069	0.0043

We now move to our estimates using real data. Table 7.9 shows where Participant A is rating how they are feeling and Participant B is rating how they think Participant A was feeling. We are including the repeating data for these estimates. The table is only showing μ_2, Participant B's standard rating of Participant A's state, σ_2, the random error due to Participant B's rating of Participant A and finally β_2 which is Participant B's accuracy at rating Participant A's state. β_2 is also considered to be the strength of the cognitive empathy going on between Participant A (in this case the target) and Participant B (in this case the rater). If β_2 is 0 then Participant B's rating of Participant A is not dependent on how Participant A is feeling and the rating is just random error. However, if β_2 is larger than 0 then the size indicates how strongly dependent Participant B's rating of Participant A is on Participant A's state. Therefore, larger values indicate more cognitive empathy being detected.

We can see here that σ_2 seems to be increasing somewhat as the conversations go on with the standard errors staying almost the same. The estimates for μ_2 seem to be increasing somewhat, with the standard errors increasing a little then dropping again. The estimates indicate that Participant B's standard rating of Participant A is increasing, therefore, Participant A is being rated as feeling more emotionally intense as the conversations go on. Finally, the β_2's drop off somewhat after the first conversation but then increase again in the last conversation, with the standard errors staying reasonably stable but increasing a bit in the third conversation. This indicates that Participant B best understood what Participant A was feeling in the first and last conversations. This may be as the first and last conversations were discussing good experiences and as the dyad were strangers, it was easier to discuss and empathise for happy experiences rather than sad ones.

NB. the standard errors for σ_2 are calculated as

$$\mathrm{SE}(\sigma_2) = \sigma_2 \times \mathrm{SE}(\log\sigma_2)$$

Table 7.10 shows Participant A rating how they think Participant B was feeling and Participant B rating how they were feeling. The repeating data is included here again. We see a similar pattern as before for σ_2's estimates and standard errors. The emotional intensity rating of the target, Participant B, is increasing until the third conversation 3 where drops a bit. The standard errors are staying reasonably stable here but are a bit larger than when Participant A was the target as above. Participant A's accuracy in rating

Table 7.10: Estimates and Standard Errors with repeating data for Participant A Other, Participant B Self

	σ_2	μ_2	β_2
Estimates			
Conversation 1	0.6901	1.7107	0.2161
Conversation 2	1.6476	3.0955	0.2311
Conversation 3	1.8832	4.3466	0.6012
Conversation 4	1.3459	3.2461	0.5618
Standard Errors			
Conversation 1	0.0023	0.0096	0.0029
Conversation 2	0.0054	0.0180	0.0044
Conversation 3	0.0062	0.0111	0.0068
Conversation 4	0.0045	0.0106	0.0039

Participant B's state jumps up greatly from conversation 3. The reason for this is unknown. It is possible Participant A simply rated Participant B poorly the first couple of conversations or they were still becoming more aware of Participant B's state.

7.5.2.3 Real Data without Repeated Ratings

Table 7.11: Estimates and Standard Errors removing repeating data for Participant A Self, Participant B Other

	σ_2	μ_2	β_2
Estimates			
Conversation 1	0.5620	3.8476	0.6844
Conversation 2	1.6134	5.0783	0.2865
Conversation 3	1.0548	5.9718	0.8590
Conversation 4	1.6605	5.8575	0.4962
Standard Errors			
Conversation 1	0.1352	0.0390	0.0264
Conversation 2	0.0612	0.0445	0.0910
Conversation 3	0.0841	0.0503	0.0342
Conversation 4	0.0692	0.0538	0.0583

Table 7.11 shows the estimates and standard errors of our parameters of Participant A's self rating and Participant B's other person rating with the repeating data removed. The σ_2's are mostly similar to the estimates with the repeating data except for conversation 3 σ_2 drops from ~1.8 to ~1.1. We also notice that β_2 was highest for conversation 3. Therefore, the decreased amount of error and increasing accuracy indicates that the amount of cognitive empathy was highest here. We see a different pattern here in the β_2's where conversation's 1 and 3 result in the highest values where previously conversation 3 had the lowest β_2. The standard errors for σ_2 and β_2 are increased a lot with the repeating data removed however these values are also subject to a much smaller sample size which inflates the amount of error somewhat. The standard rated emotional intensity for the target Participant A approximately the same as it was when the repeated data was included. The standard errors are also increased quite from when the repeating data was included.

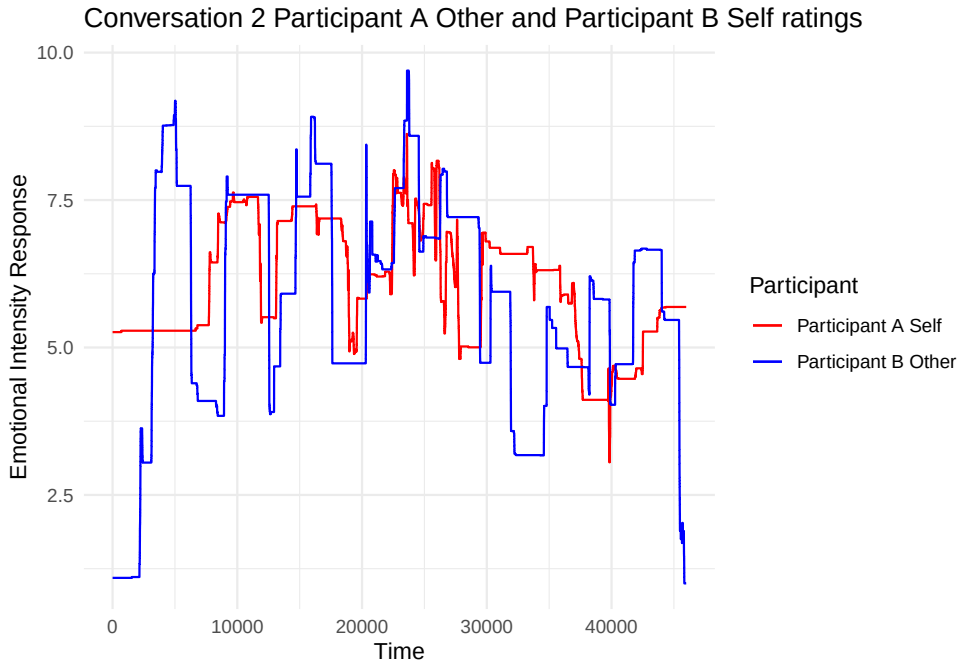


Figure 7.1: Cognitive emapthy over time within the third conversation with Participant A as target.

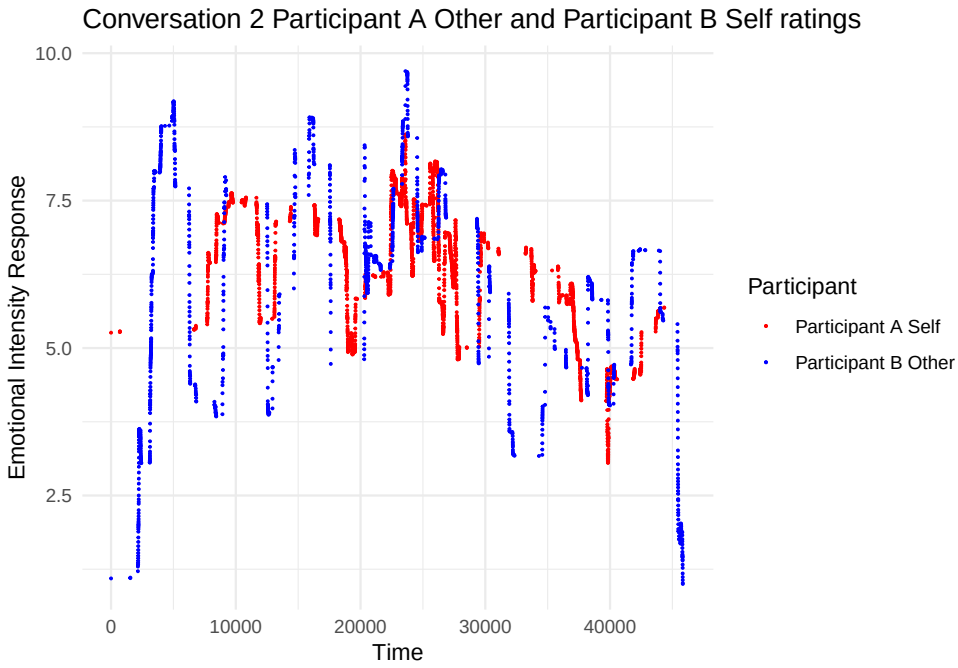


Figure 7.2: Cognitive emapthy over time within the third conversation with Participant A as target with data removed.

We are next going to take a brief look at figures 7.1 and 7.2 which shows how Participant A said they were feeling and how Participant B thought Participant A was feeling for conversation 3. We can see clearly that the two sets of ratings are moving together following the same pattern. This pattern is obvious with and without the repeating data included.

Finally, table 7.12 shows Participant A's rating of Participant B and how Participant B stated they were feeling with the repeating data being removed. The σ_2's are reasonably similar to how they were the repeating data included but with slightly larger standard errors. The standard rated emotional intensity for the target, Participant B, is a bit higher than it was previously with the repeating data included across the board. The standard errors are also increased a bit from the previous estimates. Finally, the values of β_2 are a little confusing. The values are very low for conversations 1 to 3 and then jumping up again in conversation 4. The negative estimate for conversation 2 with the large standard error indicates the estimate is not different to 0. This would mean that Participant A's rating of Participant B was not reliant on how Participant B stated they were feeling. The reason for this is unclear,

Table 7.12: Estimates and Standard Errors removing repeating data for Participant A Other, Participant B Self

	σ_2	μ_2	β_2
Estimates			
Conversation 1	0.7087	2.4631	0.1302
Conversation 2	1.7228	4.4423	-0.0542
Conversation 3	1.7000	4.8781	0.0720
Conversation 4	1.1935	4.8989	0.4594
Standard Errors			
Conversation 1	0.0388	0.0387	0.0281
Conversation 2	0.0424	0.0532	0.0874
Conversation 3	0.0649	0.0561	0.0869
Conversation 4	0.0807	0.0431	0.0294

especially when considering that we found a β_2 of ~0.6 for conversation 3 when the repeating data was included.

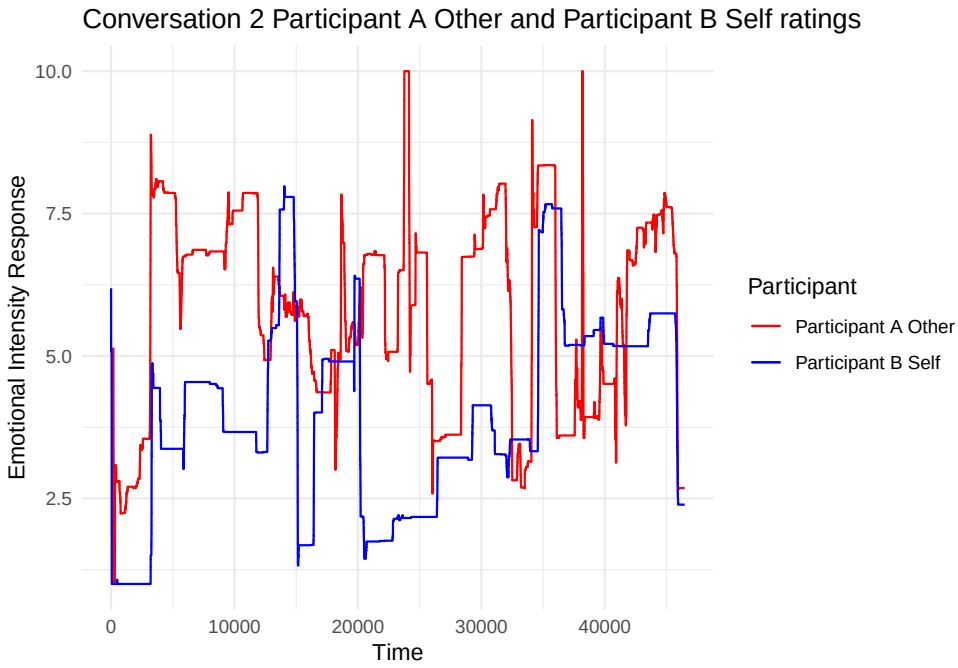


Figure 7.3: Cognitive emapthy over time within the second conversation with Participant B as target.

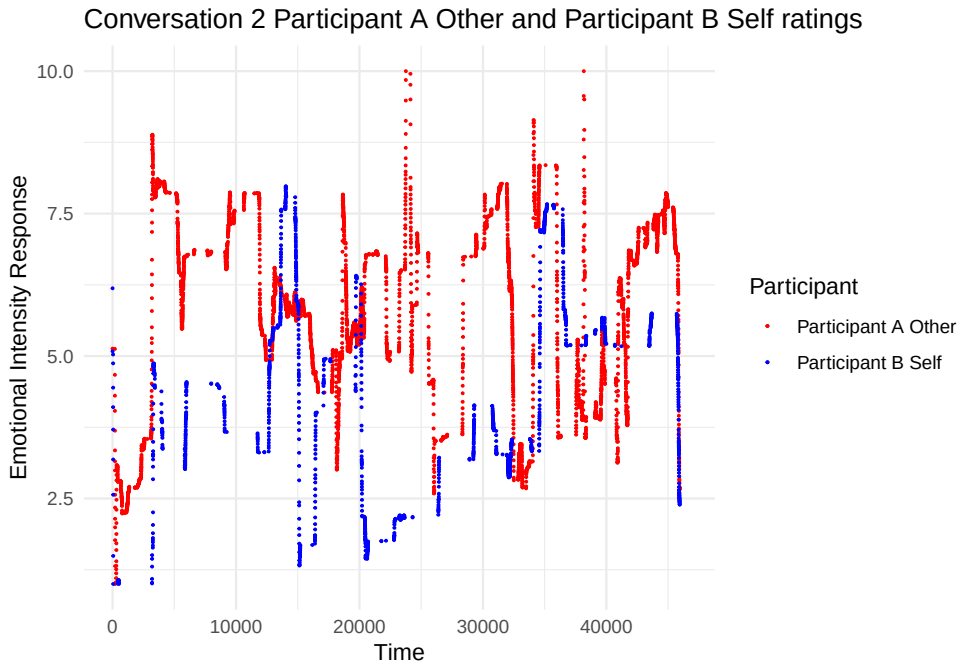


Figure 7.4: Cognitive emapthy over time within the second conversation with Participant B as target with data removed.

Figure 7.3 and figure 7.4 show the second conversation with the very low β_2 value when removing data. We can see in figure 7.3 that Participant B's rating of how they were feeling is not at all similar to Participant A's rating of Participant B's state. However, this is still with the repeating data included. The following graph shows us the data without the repeating data which the model is using to find β_2. We can see that there is no perceptible relationship between the two ratings. This is likely where the low value is coming from. It gives us more confidence in our estimates as we know they are correctly interpreting when there seems to be no relationship between the rated and target's state.

Chapter 8

Discussion

8.0.0.1 Model Summaries

We first validated our autoregressive model with generated data and found that we were able to return the correct estimates with and without removing data. We then moved onto Participant A and B's rating of themselves for the four conversations with and without repeating data included. We found that the estimates of μ_1 are reasonably stable over time and between conversations within participant's A & B. Participants A and B tended to rate how they were feeling differently, with Participant A rating themselves as feeling more emotionally intense in general using values higher on the scale. The standard errors for μ_1 were not very stable, varying greatly across conversations and with and without the repeating data. The values for ρ_1 are quite stable, staying around 1 for all conversations with standard errors varying somewhat but staying reasonably similar across conversations. The reason for this primarily is the high sampling rate. The values of ρ_1 were always less than 1 (however some were rounded up) and if one computed $\rho_1^{119.88}$ to get the persistence over 1 second, the actual amount of pull will be much lower. The standard errors would be lower here as well with the form

$$\begin{aligned} \text{SE}(\rho_1^{119.88}) &= 119.88 \times \rho_1^{118.88} \times \text{SE}(\rho_1) \\ &= 119.88 \times \rho_1^{119.88}(1-\rho_1) \times \text{logit}(\rho_1) \end{aligned}$$

The standard errors for ρ_1 and σ_1 both stayed very small ≈ 0 for all conversations with and without the repeating data. The value for σ_1 stayed reasonably similar across conversations. Overall, our estimates were reason-

ably consistent within participants, with the removal of the repeating data only substantially decreasing the standard errors for μ_1.

We now move on to discuss our model of cognitive empathy. We first verified that our model was able to return the parameters used to randomly generate a dataset with and without data removed. Given we found no issues we then moved onto the model with real data. We found a few main things, firstly, overall the standard errors increased in size when removing the repeating data. This is likely due to the greatly increased sample sizes. All conversations were ~44,000 rows with the repeating data and drop to around ~5,000-~10,000 rows without. While still large samples, they are much smaller than previously and therefore the standard errors are artificially inflated as a result. Next, μ_2 was reasonably stable before and after removing the repeating data, only changing a large amount for a couple of conversations across the two participants. The overall emotional intensity rating of the target by the rater somewhat increased from the first conversation to the later ones for both participants, regardless of whether the repeating data was included. There were differences in the size of μ depending on who the rater was and who the target was, i.e. there were individual differences in how the emotional intensity ratings were created. The values for σ_2, the amount of random error due to the rater's rating of the target was reasonably stable after increasing between conversations 1 and 2. Finally, β_2, the amount of cognitive empathy or raters accuracy in rating the target varied greatly both with and without the repeating data and also across conversations. The reasons for the high variability in the magnitude of β_2 are unclear and would require further investigation. The vast majority of conversations contained a β_2 greater than 0 which allows us to be confident our model can detect cognitive empathy.

8.0.0.2 Implications

Our first set of findings from the autoregressive model imply that there are individual differences in how people rate their emotional intensity. This is important to take note of because while two people may be having the same conversation, they are rating how they were feeling differently. This verifies intuition that two people can feel differently about a given shared event. Next, further investigation and the development of an affective empathy model is required. This would allow us to understand whether the two individuals were feeling differently across time or if the standard emotional intensity ratings of themselves are simply different. Next, we found that removing

repeating data decreased the standard errors from the autoregressive model for μ_1, but made no noticeable change for ρ_1 and σ_1. This was not found to be the case for the cognitive empathy model where the standard errors increased for μ_2, σ_2 and β_2 in removing repeating data. The reason for this may be as previously observed, simply decreasing the sample size impacts our standard errors, either decreasing or increasing their magnitude with the change in model complexity. However, it is also possible that we do not require as much information to model how one person is feeling as we do with modelling cognitive empathy. If the conversations were longer, perhaps ten instead of six minutes, then after removing the repeating data we may then lower our standard errors with the increased information. The lack of substantial differences as a result of the specific conversation topic indicates that our measure of emotional intensity may not be sufficiently specific. The participants are merely asked to rate their "emotional intensity" which ignores the differences in valence and topic between conversations. We would have expected to see some differences in rating between the first, fourth and second and third conversations. The first and last conversations are discussing happy/good topics and the second and third conversations are discussing sad/bad topics. However, this is not what we found. The lack of difference in how the conversations were rated i.e. there was no difference in rated emotional intensity implies that our participants felt equally emotionally intense when discussing happy or sad topics. This similar intensity is seen in the μ values being similar within individuals and the two conversation valences (positive and negative topics).

8.0.0.3 Limitations

Our investigation was limited by a few factors. Firstly, the principal measure was a one-dimensional report of how the participants were feeling that solely looked at "emotional intensity." The measure itself is limited in its scope by only looking at how strongly participants were feeling emotionally rather than looking at how "happy" or "sad" etc. they were feeling during the conversation. Next, we are generally limited by our measure being self-reported or reported by the other participant. It is well known in psychology that humans are prone to bias and can be poor at judging aspects of themselves. We believe the participant's ratings of their emotional intensity is our best guess at their actual underlying emotional state. This limits our ability to truly know how the participants were feeling, introducing bias and decreasing certainty of our estimates. The parameters μ_1 and μ_2 are included to attempt to account somewhat for these differences in rating

style between participants but are still unable to tap into the participants underlying state. Finally, our investigation only looked at one dyad who were strangers. We expected the increase in cognitive empathy to be most salient for these strangers as they get to know each other. The two participants are in the study for 2-3 hours in total and spending an hour in the setup before the conversations take place. If there was an effect over the conversations for strangers we would have likely expected to see it here. However, this is not what was found. The model, therefore, needs to be tested on a dyad who knew each other to see if there is a difference in the amount of cognitive empathy. While we would expect strangers to have the largest increase in cognitive empathy, we would expect a dyad who knew each other well to have a higher amount of cognitive empathy in general. People who know each other well would also be expected to be better at knowing what and how the other person is feeling. Therefore, our results are limited by only having investigated one dyad, of strangers.

8.0.0.4 Applications

The findings here and work completed within the Summer Research Scholarship can be applied in several places. Firstly the models we built allow us to better understand how cognitive empathy works. This knowledge can then be applied to help people improve their ability to understand how others are feeling. Next, the pipeline created in the Summer Research Scholarship allows for further detailed analyses to be completed in the future using the eMotion study data.

8.0.0.5 Future Directions

The first future step is to build and test the affective empathy model. The model fell out of the scope of our current investigation but is equally interesting. There is a decent amount of literature detailing empathic accuracy/cognitive empathy models and testing them in a similar way to that presented here. There is far less research looking at time series investigations of empathy on the same scale as we looked at. The completion of the cognitive empathy model thus brings into question whether we will see similar effects for affective empathy which is a distinct phenomenon in and of itself. The novelty of building an affective empathy model for our high-frequency data is enough to warrant its construction. Next, the model needs to be run on a larger number of dyads or the whole dataset including both strangers and people who knew each other. This would first allow us to investigate how

well the model works for dyads of people who knew each other. From this, we would be able to aggregate results over a larger number of dyads and see if the results found here generalise to the rest of the dataset. There would also be the possibility for between-group comparisons. For example, are there differences in the amount of cognitive empathy between strangers and people who knew each other? We could also bring in measures such as the IRI (a measure of self-reported trait empathy) or the PPI-R-40 (trait psychopathy). The cognitive empathy model can be further expanded and refined in future too. The model was intended as a proof of concept to verify we are able to return correct-seeming results. The model however is not perfect as while it follows the general concept of cognitive empathy, more recent literature has more clear ideas of what the concept is and how it should be applied (Fernandez & Zahavi, 2020). The model could be refined by incorporating the care scores which get participants to rate how much care they would have needed in the situation discussed or vice-versa for how much care the other participant would have needed. These would give a one-number metric to solidify the ratings and could act as a covariate of sorts, for example.

Chapter 9

Conclusion

The report here detailed the work completed within and alongside the Summer Research Scholarship and following the completion of this work, the construction of autoregressive self-rating and cognitive empathy models. The pipeline created within the Summer Research Scholarship and the data preparation completed therein will allow researchers in the future to better understand the concept of empathy. The setting of the eMotion study and the large number of measures only serves to increase better the amount of detail this investigation can be completed in. The two models built and tested were both novel and useful additions to the empathy and empathic accuracy literature that can be further expanded on later. I found the building of the cognitive empathy and autoregressive models to be challenging and required me to draw on the large body of learning I have done up during my studies. In sum, the work reported here allows for future research to be completed while extending the empathy literature.

Chapter 10

Appendix

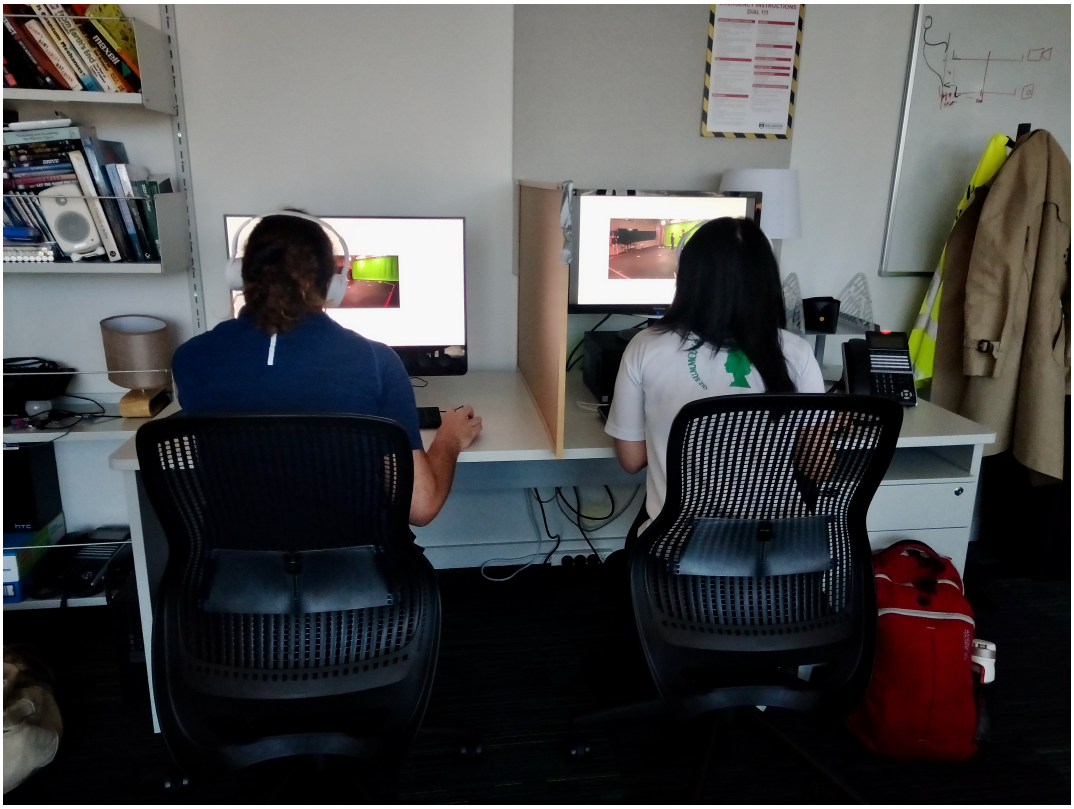


Figure 10.1: Image showing two participants completing the ratings of how they and the other person were feeling during their previous conversations at the ratings computers.

Figure 10.2: Image showing the stage setup with two participants in mocap suits with an experimenter standing across from them.

Table 10.1: First six rows of a ratings dataframe. Data for the six emotion columns is further down the dataframe and is therefore not shown here.

Time (ms)	Response	Conversation	Self/Other
0.0000	4.1219	1	Self
0.0083	4.1219	1	Self
0.0167	4.1219	1	Self
0.0250	4.1219	1	Self
0.0334	4.1219	1	Self
0.0417	4.1219	1	Self

Table 10.2: First six rows of a mocap dataframe with a subset of columns showing X, Y, Z positions for a single point.

Time (ms)	PA Hip Bone Rotation.X	PA Hip Bone Rotation.Y	PA Hip Bone Rotation.Z
0.000000	0.000968	0.024553	-0.003660
0.008342	0.000993	0.024532	-0.003662
0.016683	0.000990	0.024544	-0.003655
0.025025	0.000974	0.024557	-0.003643
0.033367	0.000982	0.024561	-0.003632
0.041708	0.000989	0.024586	-0.003632

Table 10.3: First six rows of a physio dataframe with only Participant A's data.

Time (ms)	BPM	ECG	ChestExpansion	SkinTemperature	SkinConductance
0.0000	89.9411	-0.0904	481.0977	31.8986	12.1125
0.0083	89.9411	-0.0498	480.5702	31.8986	12.1125
0.0167	89.9411	-0.1398	480.5702	31.8986	12.1125
0.0250	89.9411	-0.1314	480.5702	31.8986	12.1125
0.0334	89.9411	-0.1181	480.5702	31.8986	12.1125
0.0417	89.9411	-0.0976	480.5382	31.8986	12.1125

Where the the columns follow from the formula's below as outlined in the cognitive empathy model section

$$\begin{aligned}\text{Given data } \{(t_{x,i}, x_i)\}_{i=1}^{n_x} \text{ and } \{(t_{y,i}, y_i)\}_{i=1}^{n_x} \text{ we define}\\ d_{x,i} &= t_{x,i} - t_{x,i-1}\\ p_i &= max\{j \in \{1, \ldots, n_x\} : t_{x,j} < t_{y,i}\}\\ d_{y,i} &= t_{y,i} - t_{x,p_i}\end{aligned}$$

Table 10.4: Example generated dataset with 20 observations.

Time	x	y	$d_{x,i}$	$d_{y,i}$	x_{p_i}
1	10.0000	50.0000	0	NA	NA
2	9.3475	51.5096	1	1	10.0000
3	12.2029	50.2071	1	1	9.3475
4	14.0869	NA	1	NA	NA
5	13.6901	NA	1	NA	NA
6	NA	NA	NA	NA	NA
7	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA
12	NA	NA	NA	NA	NA
13	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA
15	NA	52.1355	NA	10	13.6901
16	NA	53.7684	NA	11	13.6901
17	NA	52.9696	NA	12	13.6901
18	NA	47.8349	NA	13	13.6901
19	9.1845	53.7483	14	14	13.6901
20	6.9541	48.4749	1	1	9.1845

Chapter 11

References

Bloom, P. (2017). Empathy and Its Discontents. *Trends in Cognitive Sciences*, *21*(1), 24–31. https://doi.org/10.1016/j.tics.2016.11.004

Byrd, R. J., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208. https://doi.org/10.1137/0916069

Coll, M.-P., Viding, E., Rütgen, M., Silani, G., Lamm, C., Catmur, C., & Bird, G. (2017). Are we really measuring empathy? Proposal for a new measurement framework. *Neuroscience & Biobehavioral Reviews*, *83*, 132–139. https://doi.org/10.1016/j.neubiorev.2017.10.009

Davis, M. H. (1980). A Multidimensional Approach to Individual Differences in Empathy. *Catalog of Selected Documents in Psychology*, *10*, 85–104.

Decety, J. (2011). Dissecting the Neural Mechanisms Mediating Empathy. *Emotion Review*, *3*(1), 92–108. https://doi.org/10.1177/1754073910374662

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2014). Using a genetic algorithm to abbreviate the Psychopathic Personality InventoryRevised (PPI-R). *Psychological Assessment*, *27*(1), 194–202. https://doi.org/10.1037/pas0000032

Eklund, J. H., & Meranius, M. S. (2020). Toward a consensus on the nature of empathy: A review of reviews. *Patient Education and Counseling*, *104*(2), 300–307. https://doi.org/10.1016/j.pec.2020.08.022

Fernandez, A. V., & Zahavi, D. (2020). Basic empathy: Developing the concept of empathy from the ground up. *International Journal of Nursing Studies*, *110*, 103695. https://doi.org/10.1016/j.ijnurstu.2020.103695

Gachter, S., Starmer, C., & Tufano, F. (2015). Measuring the Closeness of Relationships: A Comprehensive Evaluation of the 'Inclusion of the Other in the Self' Scale. *PLOS ONE*, *10*(6), e0129478. https://doi.org/10.1371/journal.pone.0129478

Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, *159*(3), 225–243. https://doi.org/10.1080/00224545.2018.1477442

Hein, G., & Singer, T. (2010). Neuroscience meets social psychology: An integrative approach to human empathy and prosocial behavior. In M. Mikulincer & P. R. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature.* (pp. 109–125). Washington: American Psychological Association. https://doi.org/10.1037/12061-006

Ickes, W. (1993). Empathic Accuracy. *Journal of Personality*, *61*, 587–610. https://doi.org/10.1111/j.1467-6494.1993.tb00783.x

Li, S., Cui, L., Zhu, C., Li, B., Zhao, N., & Zhu, T. (2016). Emotion recognition using Kinect motion capture data of human gaits. *PeerJ*, *4*, e2364. https://doi.org/10.7717/peerj.2364

Neumann, D. L., & Westbury, H. R. (2011). The Psychophysiological Measurement of Empathy. In D. J. Scapaletti (Ed.), *Psychology of Empathy* (p. 24). Nova Science Publishers.

Peirce, J. W. (2007). PsychoPyPsychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Roth, M., & Altmann, T. (2021). The self-other agreement of multiple informants on empathy measures and its relation to empathic accuracy. *Personality and Individual Differences*, *171*, 110499. https://doi.org/10.1016/j.paid.2020.110499

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(4), 393–414. https://doi.org/10.1017/S0140525X12000660

Stinson, L., & Ickes, W. (1992). Empathic Accuracy in the Interactions of Male Friends Versus Male Strangers. *Journal of Personality and Social Psychology*, *62*(5), 787–797.

Watson, D., Anna, L., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.