Thesis Proposal

Multimodal and Dyadic Modeling of Client-Therapist Interaction

An Interpretable and Causal Strategy

Alexandria K. Vail

November 16, 2022

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Louis-Philippe Morency, Chair
Dr. Jeffrey F. Cohn
Dr. Robert Kraut
Dr. Adam Perer

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Productive interaction between client and therapist is central to successful therapy, but is often hindered by substantial challenges along the way. During each therapy session, the therapist is constantly assessing the client's symptoms through their behavior. These behaviors may be expressed through multiple channels: spoken language and "body language". Therefore, the first construct we focus on is the *multimodal* aspect of behavior. Another fundamental challenge during therapy is the development and maintenance of a collaborative relationship between the client and the therapist. This relationship develops over the course of several weeks, requiring longitudinal study within and across multiple sessions. Thus, the second construct we focus on is the *social* aspect of behavior. The ambition of this thesis is to address these challenges of long-term psychotherapeutic interaction by approaching behavioral analysis through the lens of both *multimodal* and *social* behavior dynamics.

We pursue the challenge of *multimodal* behavior dynamics through three perspectives: verbal behavior, nonverbal behavior, and cross-modal behavior. This work addresses the difficulty of evaluating client symptoms across multiple modalities. The *verbal* component of behavior conveys information not only through high-level message intent, but also through more detailed aspects of speech, such as word choice and sentence structure. We present a multifaceted analysis of the client's language use as it relates to their psychological health, including a detailed consideration of lexical, structural, and disfluency components of their speech. The *nonverbal* component of behavior includes behaviors such as facial expressions, gestures, or eye gaze patterns. In particular, we study the ever-prevalent nonverbal signal of gaze aversion patterns and how they provide considerable information about the severity of the client's symptoms. Building upon this work, we then propose the consideration of *cross-modal* behavior: we seek to identify what knowledge can be gained from multiple modalities in unison that we cannot gain from single modalities in isolation.

We pursue the challenge of *social* behavior dynamics in three dimensions: facilitative behavior, convergent behavior, and divergent behavior. This work investigates the growth and decline of the collaborative relationship between the client and therapist over the course of multiple dyadic interactions. Through *facilitative* behavior, interaction participants attempt to maintain the flow of conversation, such as through turn-taking patterns. We recount a detailed analysis of turn-taking behaviors and mirroring of head gestures as they signal the quality of the collaboration between client and therapist. Through *convergent* behavior, participants (consciously or subconsciously) coordinate their behavior, such as through linguistic entrainment. We present a modeling of stylistic and content entrainment over multiple sessions as it relates to the client-therapist relationship. The final remaining component that we propose to address is *divergent* behavior, which occurs with an increase in contrast between the behavior of the participants.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Through psychotherapy, psychologists aim to help people from all walks of life achieve happier, healthier, and more productive lives. Although there are several research-based approaches to psychotherapy, the goal remains the same: to help people work through their problems and develop more effective habits. Psychotherapy has been shown to be highly effective in treating common behavioral health issues, such as depression and anxiety [11, 89]. Individuals with heart problems or other chronic illnesses have been shown to live longer when their physical treatment is augmented with psychotherapeutic treatment [4, 63, 152]. In the United States, approximately one in ten people seek mental health treatment in a given year, with the majority treatment plan including psychotherapy [151].

A core principle of psychotherapy is that it is a collaborative treatment plan based on the *client-therapist relationship*. The objective of a psychologist is to provide a healthy and supportive environment that allows a client to speak openly with someone who is objective, neutral, and nonjudgmental. This relationship is by necessity asymmetric: the client opens up to discuss their thoughts and concerns, and the therapist generally does not. This contrasts with the most common friendship and acquaintance social interactions, inherently two-sided relationships in which one side opens up gradually in parallel to the other. As a result, developing trust and mutual respect in a therapeutic relationship has the potential to be a considerable challenge, but the many benefits of doing so are well-established; e.g., reduction in client dropout [57, 94] and improvement in treatment outcomes [23, 91].

A second core principle of psychotherapy is that most psychological symptoms and concerns have a significant effect on a client's behavior. Abnormal behavior patterns and events have long been linked to the identification of psychiatric symptoms and concerns [18, 144, 162] and are often included as a critical component of the diagnostic criteria themselves [5]. These indicator behaviors span a wide range of modalities on the verbal, vocal, and visual spectra. Averted gaze [143, 162], increased fidgeting [44], heightened [90, 113] or reduced [22, 124] emotional expressivity, and disfluent language [13] are some of the many examples of behavioral markers identified by clinicians as significant indicators of psychological concerns. Although most work in psychology and psychiatry has historically relied on manual annotation of behavior, computational modeling of these and similar behaviors has been met with reasonable success in other domains, such as social rapport [29, 58] or educational tutoring [37, 167]. However, the intersection between symptomatic behavior and social behavior is a challenge unique to the medical domain.

The challenge of studying both symptomatic behavior and the client-therapist relationship produces a unique opportunity for us as human-centered computer scientists. As computation becomes increasingly pervasive in our everyday lives, it also facilitates the discovery and development of new opportunities for communication and interaction. The challenge of supporting complex tasks and mediating difficult interactions has been the focus of considerable study in the fields of human-computer interaction and computer-supported collaborative work. Although the use of computational behavior modeling in mental healthcare has been approached in the recent past, there is a relative paucity of work that is not constrained to unimodal analysis through monadic perspectives [34, 49, 150], which we argue is insufficient for the purposes of real-world application.

The central theme of this thesis is the critical examination of computational behavior analysis as an enhancement to the therapeutic process, with a focus on symptomatic behaviors and the development of the client-therapist relationship. Human behavior is a complex phenomenon: consequently, we approach this thesis in two dimensions. First, we recognize *multimodal* behavior dynamics: alongside the overt modality of spoken language, “body language” is also expressed through gesture, facial expression, pose, gaze, and many more aspects. Second, we consider *social* behavior dynamics: we know that humans do not behave in an individual vacuum — we are always behaving in relation and in reaction to the behaviors of those around us. These two dimensions of behavior characterize the essential underlying structure of this work: client symptom severity is principally studied in the context of multimodal behavior, while the client-therapist relationship is principally studied in the context of social behavior.

In the following Section 1.1, we review the key challenges to the analysis of each of these dimensions of behavior, and the core research topics we aim to address. In Section 1.2, we describe the primary contributions of this work to address these challenges, and in Section 1.3 we propose the next steps for this line of research.

1.1 Challenges

The primary aim of this thesis is to explore and assess the role of computational behavior analysis in the future of psychotherapy. Given the complexity of human behavior, this thesis evaluates computational behavior analysis through two fundamental components: *multimodal* behavior dynamics and *social* behavior dynamics.

Multimodal Behavior Dynamics

Human behavior is inherently *multimodal* in nature. Although we often associate ‘communication’ with ‘conversation’, communication consists of far more than simply words — a fact well-established through several decades of study by psychologists who dedicate entire careers to the research of nonverbal behavior [93, 135]. Early research on the various components of communication implied that nonverbal behavioral patterns are considered with significantly more weight than explicitly spoken messages, especially when recognizing emotion or interpreting incongruent modalities [112]. According to modern research, the true proportional significance varies widely depending on the social context and condition [159, 160].

The *verbal* component of communication conveys information through both explicit (linguistic) and implicit (paralinguistic) channels: we consider not only what words are spoken, but also how they are spoken. Beyond the high-level information relayed through spoken messages, much research has established that a moderate amount of information about the speaker's affective state can be inferred directly from surface-level lexical features [6, 62, 147]. On the other hand, paralinguistic features include aspects of spoken language that surround the explicit message and influence its meaning without altering its content: prosodic elements such as pitch or tempo, non-linguistic vocalizations such as disfluencies (e.g., "umm...", "oh!") or laughter, and turn-taking patterns such as extended silence or repeated interruptions. The unspoken component of communication is commonly referred to as the *nonverbal* component: this component includes behaviors such as facial expressions, gestures, or eye gaze patterns.

A unique challenge of psychotherapy is the reality that many, if not most, psychological symptoms and concerns have a marked impact on a person's behavior, often by definition [5]. In the context of mental healthcare, we must acknowledge that all behavior change is possibly — if not expectedly — influenced by both affective state and psychological health. A significant body of work has focused particularly on the behavior of depressed individuals, finding abnormal patterns in facial expression, voice features, and body movement [26, 34]. However, although depression is the most prevalent mental health diagnosis, psychological concerns can span a much broader spectrum, ranging from psychosis to anxiety to obsessive-compulsion. Each of these diagnoses consists of a range of psychiatric symptoms, most of which have a notable impact on an individual's behavior [5], introducing considerable complexity to the analysis of these interactions.

In summary, it appears undeniable that we must consider human behavior through a variety of modalities — that is, we must consider the *multimodal* dynamics of human behavior. Another important aspect particularly applicable to the domain of mental healthcare is the impact of various psychological symptoms and concerns on an individual's behavior. We know that a great deal of study has been dedicated to the modeling of multiple modalities in the technological fields, and that the influence of psychological health on behavior is a major area of interest within the fields of psychology and psychiatry. However, the synthesis of these two elements is a topic of study in its relative infancy. This intersection is one of the major challenges we aim to address in this work.

To study client-therapist behavior across multiple modalities, we begin with an investigation of the *verbal* (R1.1) and *nonverbal* (R1.2) components of behavior; we then combine these modalities in a study of *cross-modal* behavior (R1.3, proposed).

Social Behavior Dynamics

No person exists in a solitary vacuum: we all exist within relationships that we are constantly regulating through *social* behaviors. In every single interaction we have, participants not only convey information about the task or topic at hand, but also indirectly develop a connection between themselves and their conversational partners. Humans leverage a wide variety of strategies to establish and maintain social relationships: e.g., developing rapport through small talk, intimacy through personal disclosure, and respect through politeness. These social behaviors are important not only in casual conversation, but they are also particularly key to the development of any collaborative relationship.

The existing work investigating the importance of the client-therapist relationship has firmly established its significance in ensuring positive treatment outcomes [74, 110]. In particular, much of the psychological literature on this relationship focuses on what is commonly known as the *working alliance* [73]. The working alliance aims to capture the collaborative aspect of the client-therapist relationship, divided into three components: agreement on the overall goal of the treatment, agreement on the tasks required to reach that goal, and the feeling of emotional bond between the participants. The quality of this working alliance between client and therapist plays a crucial role in ensuring many positive therapeutic outcomes, including reduction of the client's symptoms and concerns [51, 73, 74], reduced drug abuse and recidivism [106] improved medication compliance [47], and decreased rates of client dropout [47, 95, 141]. A thorough understanding of the developing relationship between client and therapist is critical to the success of any therapeutic treatment.

Most social interactions pose their own domain-specific challenges, but there are many challenges particular to the domain of psychotherapy. The predominant challenge is the fact that therapeutic conversations innately involve highly sensitive or personal topics [45]. The process of developing a supportive relationship and 'opening up' to the therapist may therefore at times elicit strong negative emotions such as shame, apprehension, or even fright [109]. The evocation of these challenging emotions is not only common, but often an explicit goal of the treatment: several lines of research have established that high emotional arousal during therapeutic sessions is positively correlated with treatment outcomes across therapeutic approaches and psychiatric symptoms [9, 101, 122].

These high-arousal emotions will frequently have a significant impact on the therapist-client relationship, for better or worse [155]. If the client experiences these emotions, but distances themselves from the therapist and refuses to discuss their emotions, the therapist-client relationship suffers as a result [9, 101], and these cases are considerably more likely to result in client dropout [45]. However, instances in which clients actively approach and explore these emotions with the therapist have been shown to significantly strengthen the therapeutic relationship [53, 155]. As a result, these high-arousal moments of emotional experience have the potential to change the overall trajectory of the treatment.

Through these considerations, an important theme emerges: we cannot consider the behavior of the therapist and client as individuals, but as two interacting members of a *social* dyad. The long-term development of a relationship between participants is key to the success of treatment, but certain short-term events within the interaction also have the potential to drastically impact treatment outcomes. This complex component of behavior is the second major challenge we aim to address in this work.

To investigate the dynamics of the client-therapist relationship, we explore three aspects of social behavior: *facilitative* behavior through which participants maintain conversation (R2.1), *convergent* behavior that brings the participants together (R2.2), and *divergent* behavior that drives the participants apart (R2.3, proposed).

1.2 Contributions

R1.1. Multimodal Challenge: Verbal Behavior Dynamics (Chapter 2)

- We present an analysis of three forms of spoken language markers as indicators of symptom severity:
  - lexical markers, through a study of the function of words;
  - structural markers, through a study of grammatical fluency; and
  - disfluency markers, through a study of dialogue self-repair.
- We identify multiple language markers indicative of the type and severity of symptoms the client is experiencing.
  - A general lack of using relative language (i.e., ‘yesterday’, ‘lately’) is highly indicative of more severe symptoms, regardless of the type of symptom.
  - Words of power, such as ‘superiority’ and ‘important’, are significantly associated with the severity of “positive” distorted symptoms, such as hallucinations or delu-

sions.

    - Linguistic difficulty during cognitive processing, reflected through an increased use of disfluencies — such as repeating oneself or restarting an utterance — can be related to the severity of “negative” reduction symptoms, such as blunted affect and emotional withdrawal.

- We also develop a predictive model to estimate the severity of different forms of symptoms based on the client’s use of language.

R1.2. Multimodal Challenge: Nonverbal Behavior Dynamics (Chapter 3)

- We present a computational analysis of gaze aversion during clinical interviews.
- We identify multiple gaze markers indicative of the types of symptoms the client is experiencing.
    - Clients tend to avert their gaze more often during introspective questions when experiencing “negative” reduction symptoms, such as blunted affect and emotional withdrawal.
    - Clients also tend to avert their gaze more often (and especially downward) when experiencing negative symptoms.
    - Clients tend to avert their gaze laterally more frequently when experiencing “positive” distorted symptoms, such as hallucinations or delusions.
- We also develop a predictive model capable of distinguishing between symptom-based subtypes of schizophrenia based on the gaze aversion behaviors of the client.

R2.1. Social Challenge: Facilitative Behavior Dynamics (Chapter 5)

- We present an analysis of head gesture and speaking turn patterns as indicators of the strength of the working alliance between the client and therapist.

- We develop a predictive model capable of predicting participant-reported ratings of working alliance based on behavioral markers of head gestures and speaking turn patterns.
- We present an ablation study comparing the contribution of head gestures and speaking turn patterns on the prediction of working alliance ratings.
    - Head gestures tend to be more indicative of the task-oriented components of the working alliance, while turn-taking behaviors tend to be more related to the emotional component.
- We also present an ablation study comparing the contribution of self and partner behaviors on participants' ratings of working alliance.
    - Participant ratings of the working alliance are largely uninformed by the behavior of the other participant.
    - However, beyond simply being uninformed by the partner's behavior, in certain cases, working alliance ratings are misinformed by the partner's behavior.

R2.2. Social Challenge: Convergent Behavior Dynamics (Chapter 6)

- We present an analysis of stylistic and content entrainment as it reflects participants' self-reported ratings of the working alliance between the client and therapist.
- We identify several markers of working alliance ratings based on the entrainment behaviors of the participants.
    - Stylistic entrainment tends to be associated with the emotional components of the working alliance, while content is more related to the task-oriented components.
    - The linguistic entrainment patterns of the client are significantly indicators of their perception of the working alliance.
    - Therapist linguistic entrainment behaviors have a marked impact on the client's perception of bond.

- We also establish evidence of the importance of considering causality in studying and modeling these relationships.

1.3 Proposed Contributions

R1.3. Multimodal Challenge: Cross-Modal Behavior Dynamics (Chapter 4)

After having studied the verbal (Chapter 2) and nonverbal (Chapter 3) components of behavior, we next aim to study *cross-modal* behavior. The aim of cross-modal analysis is to identify what information can be discovered through the analysis of multiple modalities in unison that cannot be discovered from the analysis of single modalities in isolation. We study two forms of cross-modal behavior: *monadic* cross-modal behavior (e.g., client verbal $\times$ client nonverbal), and *dyadic* cross-modal behavior (e.g., therapist verbal $\times$ client nonverbal).

Through our *monadic* analysis, we focus on moments of heightened emotion, which have been consistently linked to positive therapy outcomes, such as reduction in client symptoms [101, 122]. These moments are particularly interesting from a cross-modal perspective, as these moments of heightened emotion are also moments where the verbally conveyed information may conflict with the nonverbally conveyed information [9]. Through our *dyadic* analysis, we focus on backchanneling behaviors. Backchanneling responses are non-intrusive interjections that signal the listener's attention, interest, understanding, or attitude towards the speaker's message. Existing research has shown that psychological health often affects unimodal backchanneling behaviors [66, 166], but we extend this research into backchanneling across modalities.

R2.3. Social Challenge: Divergent Behavior Dynamics (Chapter 7)

In the social challenge, we study collaborative dyadic interaction in three parts: facilitative behavior that maintains conversation, such as turn-taking (Chapter 5); convergent behavior that brings the participants together, such as linguistic entrainment (Chapter 6); and *divergent* behavior that drives participants apart. Divergent behavior occurs with an increase in contrast between participants' behavior. These behaviors could appear as overt as open conflict, or could manifest more subtly, such as through withdrawal behaviors.

Due to the highly personal nature of therapeutic conversations, a large portion of divergent behavior during these sessions occurs when sensitive topics arise [45]. Since these topics are generally relevant to the therapeutic treatment, the therapist is often inclined to press their discussion despite client discomfort [9], which may result in the client feeling a sense of emotional aggression [101]. Given this possibility of conflict, these events have the potential to damage the client-therapist relationship, but if addressed appropriately, these events also have the potential to strengthen the relationship as well [155]. We focus our analysis on both the acute moment of divergence and the continuing interaction that immediately follows.

Part I

Multimodal Behavior

"When the eyes say one thing, and the tongue another, a practiced man relies on the language of the first... How many furtive inclinations avowed by the eye, though dissembled by the lips!"

– Ralph Waldo Emerson, 1860 [42]

Chapter 2

Verbal Behavior Dynamics

The evaluation of psychotic disorders is often complex, as their multifaceted nature is often difficult to quantify. While written language has been previously studied, the analysis presented in this chapter takes the novel approach of examining the rarely studied modality of *spoken* language of individuals with psychosis as naturally used in social, face-to-face interactions. Our analyses expose a series of language markers associated with psychotic symptom severity, as well as interesting interactions between them. In particular, we examine three facets of spoken language: (1) lexical markers, through a study of the function of words; (2) structural markers, through a study of grammatical fluency; and (3) disfluency markers, through a study of dialogue self-repair.

The work described in this chapter first appeared in the following publication:

> Alexandria K. Vail, Elizabeth Liebson, Justin T. Baker, Louis-Philippe Morency. Toward Objective, Multifaceted Characterization of Psychotic Disorders: Lexical, Structural, and Disfluency Markers of Spoken Language. *Proceedings of the Twentieth International Conference on Multimodal Interaction (ICMI 2018)*, Boulder, Colorado, 2018.
>
> https://doi.org/10.1145/3242969.3243020

2.1 Overview

Psychotic disorders are forms of severe mental illness that cause significant functional impairment and can result in profound lifetime disability and loss of productivity [85]. Assessment of psychotic disorders often relies upon clinical interviews and observation of an individual's day-to-day behaviors, but unfortunately, clinicians put in this role are often bounded by constraints such as time availability, clinician fatigue, or the simple human inability to study all channels of behavior at once. These difficulties necessitate the development of tools for the computational phenotyping of mental illness, which can offer objective support and data analysis to clinicians to aid in assessment and treatment.

When assessing the psychiatric condition of an individual, clinicians rely upon subjective analysis of atypicality in the individual's behavior, such as nonverbal cues, social behaviors, and language use. Critically, these behaviors can also be evaluated through multimodal behavior analysis systems. Although a moderate amount of work has focused on nonverbal behaviors through audiovisual information [158, 168], little work has focused on the language use of these individuals with psychotic disorders. Further, to date, almost all work on language use in psychotic disorders has focused on written texts, such as autobiographical narratives and social media interactions [72, 115]. The present work is one of the first studies to examine *spoken* language use in individuals with psychotic disorders from a computational perspective in clinical settings.

Furthermore, most prior work has examined differences between individuals diagnosed with psychotic disorders and those who are not [20, 31, 83, 115], but few studies have examined behaviors within psychotic disorder groups. The primary line of work to date on symptom-specific written language use focuses on anhedonia, a negative symptom of schizophrenia characterized by a reduction in expression of positive affect [15, 20]. This prior work studies only one specific symptom of schizophrenia and does not yet cover the full range of symptoms expressed by psychotic disorders. The present analysis takes the novel approach of examining language use as it pertains to a broad range of psychotic symptoms and more fully characterizes an individual's manifestation of the disorder.

In this chapter, we analyze three facets of spoken language use in individuals with psychotic disorders: (1) lexical markers, through a study of the function of words; (2) structural markers, through a study of grammatical fluency; and (3) disfluency markers, through a study of dialogue self-repair. For each of these three facets, we perform single-facet analyses, which will inform our multi-faceted fusion approach. Our multi-faceted interaction analysis is conducted in two parts: a moderation analysis and predictive model building. Our moderation analysis examines how the relationship between an individual's symptom severity and two facets at a time. Our multi-facet predictive models consider the set of features emerging as significant in single-facet analyses as predictors of psychotic symptom severity. We perform our analyses and experiments on a dataset consisting of semi-structured clinical interviews between clinicians and adult individuals with schizophrenia or bipolar disorder recently admitted to an inpatient unit at a major psychiatric facility.

2.2 Psychosis and Language

Extended analysis of language use has the potential to influence the understanding of language dysfunction in psychosis, as well as the potential further development of clinical assessment tools. We examine features at three levels of a participant's language use: *lexical markers*, *structural markers*, and *disfluency markers*. The following subsections detail previous work in these areas. This prior work will inform our single-faceted research described in Section 2.4.

Lexicon

As previously mentioned, any previous studies implementing lexical analysis have (1) focused on written language and (2) compared between psychotic disorder and control groups [20, 31, 83].

However, few studies have examined research within psychotic disorder groups to investigate whether word use is linked with psychotic symptoms themselves. For our lexical marker analyses, we focus on five lexical categories, which we introduce as part of three main groups: affect, power, and reality monitoring.

Affect. The foremost line of study of this topic is focused on anhedonia [15, 20], a negative symptom of schizophrenia characterized by a reduction in expression of positive affect. Cohen et al. observed that participants exhibiting high levels of anhedonia used more negative affect words when discussing pleasant topics than those exhibiting low levels of anhedonia [31]. In our analysis, we follow this line of work by investigating *affect* words as they relate to the broader spectrum of psychotic symptoms.

Power. The most characteristic symptoms of psychotic disorders revolve around delusions and grandiosity [87]. Individuals that express high levels of delusion tend to hold beliefs which are unfounded, unrealistic or idiosyncratic [87]. Grandiosity, on the other hand, involves an exaggerated self-opinion and unrealistic convictions of superiority, which can include delusions of extraordinary abilities, wealth, knowledge, fame, power, or moral righteousness [87]. Our analysis examines the impact of delusions and grandiosity on the language of individuals with psychotic disorders via words of *power*: words relating to the drive for influence and dominance.

Reality monitoring. Another significant segment of the lexicon examined in the present analysis involves words related to *reality monitoring* [81]. The concept of reality monitoring extends from the idea that people recall information from two primary sources: external sources (such as perceptual processing and contextual information) and internal sources (such as reasoning). Reality monitoring refers to the processes people use to decide whether information was generated from an external source or an internal source.

Numerous studies have observed reality monitoring impairments in individuals with psychotic disorders compared to healthy controls [46, 50, 88], but most work focuses on the neurocognitive aspects of the phenomenon, rather than detection in the field. The present analysis takes the novel approach of investigating reality monitoring as it manifests in conversational settings (i.e., spoken language). In particular, it features a focus on the use of words that reflect each of the two potential sources of information: external sources through *perceptual processing* and *relative* (contextual) words, and internal sources through *cognitive processing* words.

Language Structure

Individuals with speaking disorders or cognitive impairment tend to express themselves atypically compared to control groups [48]. Prior work on written language has used language models to study this phenomenon by estimating the probability of a given utterance being produced, e.g., in studies of language impairment in children [48] and language dominance prediction in multilingual individuals [148]. Hong et al. conducted a study of autobiographical narratives written by individuals with and without schizophrenia; this work suggested that different language models optimally explain part-of-speech tag sequences within the two groups [72].

Few previous studies have examined perplexity itself as a measure of grammatical integrity in schizophrenia and psychosis. A study by Mitchell et al. compared posts by social media users voluntarily self-labeled as experiencing schizophrenia against posts from a control group; a marginal difference between these sets of users suggested that those with schizophrenia generated higher-perplexity posts than the control group [115]. The present study takes the novel approach of investigating perplexity as an indirect measure of psychotic symptom severity, rather than as a distinguishing characteristic between individuals with psychotic disorders and those without.

Disfluency

Disfluencies, such as self-repairs, pauses, and fillers (such as *er* and *umm*) are pervasive in day-to-day dialogue [142]. These disfluencies are generally regarded as symptomatic of problems in communication, whether caused by production or self-monitoring issues [104]. Disfluencies can also highlight the interactive nature of dialogue — some disfluencies occur as a result of tailoring dialogue to a specific listener, or in response to feedback from interlocutors [56].

Individuals with psychotic disorders tend to have difficulties with language and social cognitive skills, and especially with self-monitoring [80] and turn-taking [117], but little research has examined how these problems affect interaction. Work by Leudar et al. found that the less self-repair that an individual with schizophrenia employs, the more verbal hallucinations they tend to experience [103]. Further work by McCabe et al. discovered that other-initiated repairs (clarification of a clinician's dialogue, in particular) are associated with improved adherence to treatment [111]. The present work, therefore, examines the disfluencies and self-repairs present in the dialogue of individuals with psychotic disorders as they relate to symptom severity.

2.3 Dyadic Psychosis Interview Dataset

The dataset examined in the present analysis consists of a series of clinical interviews with adult individuals recently admitted to an inpatient psychotic disorder unit at a major psychiatric facility. Video and audio recordings, as well as transcripts, were collected from 53 sessions (28 unique participants). Each session consisted of a semi-structured clinical interview between the admitted individual and a clinician, lasting approximately 10–15 minutes each. The interview script was modeled upon existing everyday clinical interactions designed to elicit reactions that may be illustrative of the psychiatric condition of the individual[1]. A list of interview questions is presented in Table 2.1.

Following the conclusion of each interview, each participant was administered a series of clinical scales, including the Positive and Negative Syndrome Scale (PANSS) [87], a scale used for measuring psychotic symptom severity. PANSS involves seven-point ratings of 30 symptoms across three dimensions: *positive symptoms*, involving behaviors in excess or distortion of nor-

[1]Although participants varied regarding previous exposure to interactions of this type, this diversity is reflective of the larger population, and we believe that this strengthens the applicability of this analysis.

TABLE 2.1
List of interview questions administered during the session.

What brought you into the hospital?
Has anything in particular been on your mind?
What has the team here been helping you with?
Would you say that they are doing a good job?
What are your goals for the hospitalization?
How are people treating you?
How is the food?
How is your mood? / How are your spirits?
How is your thinking/focus?
How is your energy?
How have you been sleeping?
How is your self-confidence compared to how it usually is?
What changes do you observe since you were hospitalized?

mal function; *negative symptoms*, involving behaviors diminished or suppressed below normal function; and *general psychiatric symptoms*, involving items that cannot be linked decisively to either syndrome. In this paper, we focus on the symptoms from the positive and negative scales (see descriptions in Table 2.2). The average positive scale score in the present sample is $\mu = 14.88$ ($\sigma^2 = 7.82$), and negative scale score $\mu = 12.14$ ($\sigma^2 = 4.71$), both in a possible range of 7 to 49 (see Figure 2.1 for the distribution of the present sample).

For the following analyses, the dataset was separated into a training set (43 sessions) and a held-out test set (10 sessions). The single-facet analyses were performed upon the training set, and only the multi-faceted predictive models were tested upon the held-out test set after the analysis.

2.4 Single-Facet Language Analysis

Our first set of analyses examines spoken language use at three levels of a participant's dialogue: lexical markers, structural markers, and disfluency markers. The following subsections detail the

TABLE 2.2
Enumeration and brief description of a selection of symptoms contained in the PANSS positive and negative scales [87].

Scale Item	Brief Description of Behavior
Positive Scale	
Delusions	Beliefs which are unfounded, unrealistic, and idiosyncratic.
Conceptual Disorganization	Disorganized process of thinking characterized by disruption of goal-directed sequencing, e.g., circumstantiality, tangentiality, loose associations, non-sequiturs, gross illogicality, or thought block.
Hallucinatory Behavior	Verbal report or behavior indicating perceptions which are not generated by external stimuli. These may occur in the auditory, visual, olfactory, or somatic realms.
Grandiosity	Exaggerated self-opinion and unrealistic convictions of superiority, including delusions of extraordinary abilities, wealth, knowledge, fame, power, and moral righteousness.
Hostility	Verbal and nonverbal expressions of anger and resentment, including sarcasm, passive-aggressive behavior, verbal abuse, and assaultiveness.
Negative Scale	
Blunted Affect	Diminished emotional responsiveness as characterized by a reduction in facial expression, modulation of feelings, and communicative gestures.
Emotional Withdrawal	Lack of interest in, involvement with, and affective commitment to life's events.
Poor Rapport	Lack of interpersonal empathy, openness in conversation, and sense of closeness, interest, or involvement with the interviewer. This is evidenced by interpersonal distancing and reduced verbal and nonverbal communication.
Difficulty in Abstract Thinking	Impairment in the use of the abstract-symbolic mode of thinking, as evidenced by difficulty in classification, forming generalizations, and proceeding beyond concrete or egocentric thinking in problem-solving tasks.
Lack of Spontaneity and Flow of Conversation	Reduction in the normal flow of communication associated with apathy, avolition, defensiveness, or cognitive deficit. This is manifested by diminished fluidity and productivity of the verbal-interactional process.

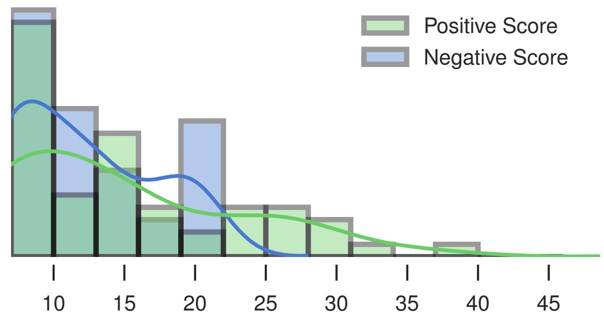


FIGURE 2.1
Distribution of PANSS positive and negative scores in the examined sample.

computational analyses of these three facets of spoken language. The results of these single-facet analyses will be used during the multi-faceted prediction task.

Lexicon Analysis

In this study, we focus on five categories of lexical markers: cognitive processing words, affect words, power words, relative words, and perceptual processing words (see Section 2.2 for details). Lexical features of participant speech were extracted using the Linguistic Inquiry and Word Count (LIWC) tool [123], a computerized measure that assesses speech and language content using a dictionary of over 4500 words across over 60 categories. LIWC has demonstrated validity in measuring expression in verbal dialogue [84] and has been used previously to assess word use in schizophrenia for written text [20, 31]. We computed a Spearman's rank correlation coefficient to assess the relationship between each of these categories and two PANSS scales (positive and negative). To account for multiple hypothesis testing, results were filtered within each scale using the Benjamini-Hochberg procedure, with a family-wise error rate of $\alpha = 0.05$. All analyses were performed upon the training set only. Results are reported in Table 2.3; significant correlations are discussed below and illustrated in Figure 2.2.

Affect. Affect words relate to the emotions: for example, *happiness*, *gloomy*, and *sadly*. Previous work has suggested that greater levels of emotion are significantly associated with lower

TABLE 2.3
Reported Spearman's rank correlation coefficient between selected LIWC features and PANSS scores. Boldface indicates significant correlations holding under a Benjamini-Hochberg procedure for multiple hypothesis testing, where $\alpha = 0.05$.

	Positive Score		Negative Score	
	corr(ρ)	p-value	corr(ρ)	p-value
Cognitive Processing	$+0.048$	0.736	$+0.018$	0.898
Affect	-0.063	0.655	$\mathbf{+0.287}$	$\mathbf{0.037}$
Power	$\mathbf{+0.374}$	$\mathbf{0.006}$	$+0.091$	0.516
Relative	$\mathbf{-0.302}$	$\mathbf{0.028}$	$\mathbf{-0.352}$	$\mathbf{0.010}$
Perceptual Processing	$\mathbf{+0.351}$	$\mathbf{0.010}$	$+0.111$	0.429

functioning in psychotic disorders [15], and expression of negative affect, in particular, has been linked to anhedonia, a major negative symptom, in the past [31]. There was a significant positive correlation between affect words and negative PANSS score ($\rho(53) = +0.287$, $p = 0.037$). The more negative symptoms expressed by a participant, the more affect words they used.

Power. Power words relate to the drive for dominance: for example, *superiority*, *important*, and *exploit*. Individuals with psychotic disorders often exhibit symptoms of grandiosity and delusions, which are associated with a perception of greater self-power [87]. There was a significant positive correlation between power words and positive PANSS score ($\rho(53) = +0.417$, $p = 0.002$). Overall, the more positive symptoms expressed by a participant, the more power words they used.

Reality monitoring. Relative words relate to situations regarding time and space: for example, *yesterday*, *lately*, and *nearby*. These words relate to the phenomenon of reality monitoring, and particularly to the attachment of information to external stimuli [81]. There was a significant negative correlation between relative words and negative PANSS score ($\rho(53) = -0.381$, $p = 0.005$), as well as a significant negative correlation between positive PANSS score ($\rho(53) = -0.302$, $p = 0.028$). We can infer from this result that the more positive or negative symptoms expressed by a participant, the fewer relative words they used.

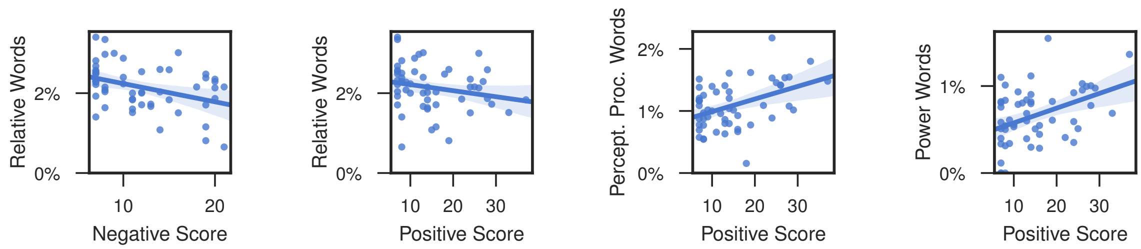


FIGURE 2.2
Regression plots of four of the significant correlations between LIWC features and PANSS scores.

Perceptual processing words relate to the senses: for example, *feeling*, *see*, and *listened*. Like relative words, these words also tend to relate to reality monitoring, and these words are also linked to the perception of external stimuli [81]. There was a significant positive correlation between perceptual processing words and positive PANSS score ($\rho(53) = +0.434$, $p = 0.001$). Overall, the more positive symptoms expressed by a participant, the more perceptual processing words they used.

Language Structure Analysis

The structure of the language — including vocabulary and syntactic constructions — expressed by a participant can be measured via *perplexity*, a measurement based on entropy, and can be interpreted to roughly estimate how predictable is a sequence of words. The present work trains a trigram backoff language model on the Switchboard corpus [54], a sizable multispeaker corpus of conversational speech and text through telephone conversations about varying topics. This corpus can be viewed as an approximation of non-psychotic disorder spoken dialogue. The model is then tested on the transcript of each session, and the overall perplexity is calculated. A Spearman's rank correlation coefficient is computed to assess the relationship between perplexity and each of the PANSS scales. All analyses were performed upon the training set only.

Results. The results suggest no significant correlation between negative PANSS score and perplexity ($\rho(53) = -0.046$, $p = 0.746$), but a significant positive correlation between positive

PANSS score and perplexity ($\rho(53) = +0.313$, $p = 0.022$). The more positive symptoms an individual expresses, the higher the perplexity of their utterances. Individuals high in positive scale symptoms tend to express symptoms such as excitement and conceptual disorganization, which may interfere with sentential construction [87].

Disfluency Analysis

Disfluencies in the form of speech repair are typically assumed to have a tripartite *reparandum-interregnum-repair* structure [145], as illustrated in the following example.

$$\text{“John} \quad [\quad \underbrace{\text{likes}}_{\text{reparandum}} \quad \underbrace{\text{uh}}_{\text{interregnum}} \quad \underbrace{\text{loves}}_{\text{repair}} \quad] \quad \text{Mary”}$$

A *reparandum* is an error in speech that is subsequently corrected by the speaker; a *repair* term is the corrected speech. An *interregnum* term is a filler token or a cue phrase between the reparandum and repair terms, often a stalling measure while the speaker generates the repair term.

We examine three forms of disfluencies: edits, repeats, and restarts. If the reparandum and the repair terms are absent, the disfluency is considered to be reduced to an isolated *edit* term. In this canonical example, the interregnum is a pause filler token (“uh”), but more phrasal terms such as “I mean” and “you know” are also often used.

The other two forms of repair we examine in the present analysis are *repeat* terms and *restart* terms. The occurrence of a *repeat* term is reasonably straightforward — this is when an individual repeats a word or a short phrase. A *restart* term occurs when an individual changes a partially complete spoken utterance, as in the example above.

Self-repairs were annotated automatically using a deep-learning-driven incremental disfluency detection model developed by Hough et al. [75]. This model consists of deep learning sequence models that consume incoming words and use word embeddings, part-of-speech tags,

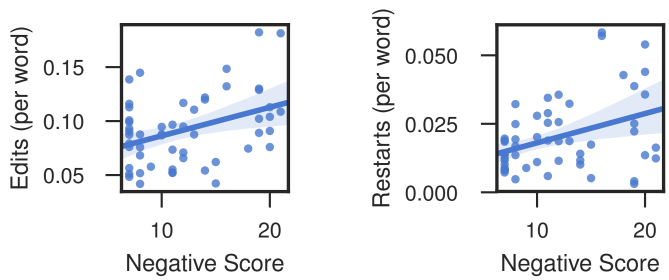


FIGURE 2.3
Regression plots of the significant correlations between self-repair features and PANSS scores.

and other features to predict disfluency labels for each word in a strictly left-to-right, word-by-word fashion.

Similar to the lexicon analysis, a Spearman's rank correlation coefficient was computed to assess the relationship between each type of self-repair and each PANSS scale (positive and negative). To control for multiple hypothesis testing, results were filtered within each scale using the Benjamini-Hochberg procedure, with a family-wise error rate of $\alpha = 0.05$. All analyses were performed upon the training set only.

Results. Results are reported in Table 2.4; significant correlations are discussed below and illustrated in Figure 2.3. Both significant correlation results are related to negative PANSS score. The negative PANSS score is characterized by symptoms such as poor rapport, difficulty in abstract thinking, and lack of spontaneity and awkward flow of conversation [87]. There was a significant positive correlation between the negative PANSS score and edit terms ($\rho(53) = +0.309$, $p = 0.024$) as well as a significant positive correlation between the negative PANSS score and restarts ($\rho(53) = +0.334$, $p = 0.014$). The more negative symptoms expressed by an individual, the more edit terms and restarts they express.

TABLE 2.4
Reported Spearman's rank correlation coefficient between selected self-repair features and PANSS scores. Boldface indicates significant correlations holding under a Benjamini-Hochberg procedure for multiple hypothesis testing, where $\alpha = 0.05$.

	Positive Score		Negative Score	
	corr(ρ)	p-value	corr(ρ)	p-value
Edits	-0.089	0.525	$\mathbf{+0.309}$	$\mathbf{0.024}$
Restarts	$+0.173$	0.217	$\mathbf{+0.334}$	$\mathbf{0.014}$
Repeats	$+0.028$	0.844	$+0.215$	0.123

Discussion

In this section, we summarize our observations for all three facets of spoken language: lexical markers, structural markers, and disfluency markers. For lexical markers, we group our observations following the three lexical category groups introduced in Section 2.2.

Affect. Our analyses investigated a series of lexicon categories as used by individuals with psychotic disorders (Section 2.4). There existed a positive correlation between affect words and negative symptoms: the more affect words an individual used, the more severe their negative symptoms. Interestingly, this counters the intuition regarding the negative symptom of emotional withdrawal and blunted affect [87]; one might believe that an individual with severe negative symptoms may not be very forthcoming about their emotions. This result relates to prior work on anhedonia, which suggested that individuals with this negative symptom do not use significantly fewer affect words than those without, but instead use affect words with a more negative valence [15].

Power. Another result involves power words: the more power words an individual expresses, the higher the severity of their positive symptoms. Some characteristic positive symptoms include delusions and grandiosity, which involve holding beliefs that are unfounded, unrealistic, or idiosyncratic, exaggerated self-opinion, and unrealistic conventions of superiority [87]. Considering that these symptoms are central to the positive symptom scale, this finding represents a useful contribution toward computational phenotyping of psychotic disorders.

Reality monitoring. Two lexicon categories emerged that are related to reality monitoring: relative words and perceptual processing words, both of which are related to information recall from external sources [81]. Relative word use is negatively associated with both negative and positive symptoms: that is, the more severe the psychotic symptoms an individual expresses, the less they speak in relative terms. It is interesting to see that this correlation holds for both symptom scales; this may be an indication of a general difficulty in psychotic disorders, rather than dependent on its manifestation. This result reinforces the findings from previous studies that suggested that reality monitoring impairments are generally characteristic of psychotic disorders [46, 88]. There was also a positive association between positive symptoms and perceptual processing: the more perceptual processing words an individual used, the more severe their positive symptoms. Unlike relative word use, perceptual processing word use appears to be dependent upon the particular manifestation of the disorder: one of the characteristic positive symptoms is hallucinatory experiences, which may lead to an individual being more aware of their surroundings, real or imagined, which in turn leads to more discussion about what they feel, see, and hear.

Structure. A correlation was discovered between positive symptom severity and language perplexity (Section 2.4). Positive symptoms entail higher-activity behaviors exceeding typical function, so individuals expressing these symptoms acutely may experience difficulty in constructing sentences; this follows from previous work suggesting that individuals with cognitive impairment may express themselves atypically compared to control groups [48].

Disfluency. There were two results regarding self-repairs during dialogue (Section 2.4). In particular, negative symptom severity was positively correlated with both edit terms and restarts. Disfluencies are generally regarded as symptomatic of problems in communication [104]. Individuals with high negative psychotic symptom severity characteristically experience problems in communication through poor rapport and flow of conversation [87]; it follows logically that this may be expressed linguistically through dialogue disfluencies.

2.5 Multi-Faceted Language Analysis

Building from the results of the single-facet computational analyses, we are interested in examining the interactions between the different facets of spoken language. In this section, we leverage these results in two multi-facet analyses: an analysis of moderation and predictive modeling. The moderation analysis will focus on two facets at a time, while the predictive modeling will integrate all three facets.

Moderation Analysis

Each of the two PANSS scales (positive and negative) were examined as a moderator of the relation between each of the lexicon features and each form of self-repair. In other terms, the analysis focused on how individuals expressing high positive or negative symptoms might self-repair more frequently when speaking on particular topics (see Figure 2.4 for an illustration). This work is conducted as a form of regression analysis [32]. Given a PANSS score X_S and a lexicon feature X_L, we predict a given dependent variable (i.e., a self-repair feature) Y_R with the model

$$Y_R = \beta_S X_S + \beta_L X_L + \beta_{SL} X_S X_L, \tag{2.1}$$

such that β_S, β_L, and β_{SL} are learned parameters via ordinary least squares on the training set [129]. For example, Y_R could indicate self-repair repeats, while X_S and X_L indicate positive PANSS score and affect words, respectively. We describe below three moderation models with significant interactions.

Negative symptoms, affect, restarts. The first model involves negative PANSS score, affect words, and restarts (see Figure 2.4a). In the first step of the regression analysis, negative PANSS score and affect words are entered as predictors of restarts; this model significantly predicted restarts ($F(50, 2) = 4.797$, $p = 0.012$, $r = +0.401$). In the second step of the analysis, the

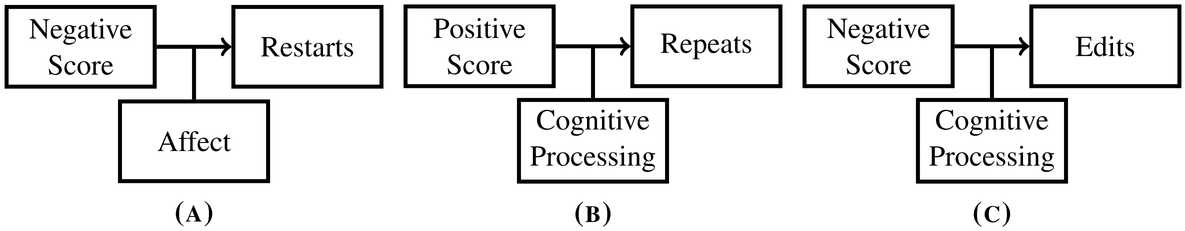


FIGURE 2.4
Illustration of the structure of the moderation analyses with significant interaction effects described in Section 2.5.

interaction term (the product of the negative PANSS score and affect word use) was introduced; this model also significantly predicted restarts ($F(49, 3) = 4.733$, $p = 0.006$, $r = +0.474$). This difference was statistically significant ($\Delta r = +0.073$, $p = 0.050$). See Table 2.5a for the final interaction model; β is the coefficient for each term, and t and p refer to a t-test value and p-value indicating its significance. From these results, we can observe that the higher an individual's negative PANSS score and the more affect words they used, the more they restarted their sentences, but when high-negative-score individuals spoke about affective utterances, they expressed *fewer* restarts than in general.

Positive symptoms, cognitive processing, repeats. The second model involves positive PANSS score, cognitive processing words, and repeats (see Figure 2.4b). In the first step of the regression analysis, positive PANSS score and cognitive processing words are entered as predictors of repeats; this model marginally predicted repeats ($F(50, 2) = 1.952$, $p = 0.153$, $r = +0.269$). In the second step of the analysis, the interaction term (the product of the positive PANSS score and cognitive processing word use) was introduced; this model did significantly predict repeats ($F(49, 3) = 2.754$, $p = 0.052$, $r = +0.380$). This difference was statistically significant ($\Delta r = +0.111$, $p = 0.048$). See Table 2.5b for the final interaction model; β is the coefficient for each term, and t and p refer to a t-test value and p-value indicating its significance. From these results, we can observe that the higher an individual's positive PANSS score, and the more cognitive processing words they used, the more repeats in their dialogue, but when high-positive-score individuals spoke about cognitive processing terms, they expressed *fewer* repeats

TABLE 2.5
Regression models examining the moderation between PANSS scores and lexical categories as predictors of self-repairs.

(A)

Restarts =	β	t	p
Affect Words	$+0.468$	$+1.443$	0.155
Negative PANSS Score	$+1.107$	$+2.947$	0.005
Interaction Term	-1.020	-2.006	0.050

(B)

Repeats =	β	t	p
Cognitive Processing Words	$+0.335$	$+1.116$	0.270
Positive PANSS Score	$+2.255$	$+2.168$	0.035
Interaction Term	-2.171	-2.028	0.048

(C)

Edits =	β	t	p
Cognitive Processing Words	-1.278	-1.568	0.123
Negative PANSS Score	-0.572	-1.716	0.092
Interaction Term	$+1.788$	$+2.070$	0.044

than in general.

Negative symptoms, cognitive processing, edits. The third model involves negative PANSS score, cognitive processing words, and edits (see Figure 2.4c). In the first step of the regression analysis, negative PANSS score and cognitive processing words are entered as predictors of edits; this model significantly predicted edits ($F(50, 2) = 4.559$, $p = 0.015$, $r = +0.393$). In the second step of the analysis, the interaction term (the product of negative PANSS score and cognitive processing word use) was introduced; this model also significantly predicted edits ($F(49, 3) = 4.667$, $p = 0.006$, $r = +0.471$). This difference was statistically significant ($\Delta r = +0.078, p = 0.044$). See Table 2.5c for the final interaction model; β is the coefficient for each term, and t and p refer to a t-test value and p-value indicating its significance. From these results, we can observe that the higher an individual's negative PANSS score, and the more cognitive processing words they used, the fewer edits in their dialogue, but when high-negative-score individuals spoke about cognitive processing terms, they expressed *more* edits than in general.

Discussion. There were three significant results observed during our moderation analysis. In particular, as individuals speak of specific topics, individuals with more severe symptoms tend to repair their language more or less often than in general. For example, individuals with high levels of negative symptoms were much less likely to restart their sentences when speaking about affective topics than in general, which may be explained by the blunted affect symptoms; it may be more straightforward for these individuals to speak about their emotions if they are not experiencing many of them. In another case, individuals with more severe positive symptoms were less likely to repeat themselves when speaking with cognitive processing terms, and individuals with more severe negative symptoms were more likely to edit themselves when speaking with cognitive processing terms. These three results are hinting to the fact that there are multi-faceted interactions in spoken language of individuals with psychotic disorders. Following these intuitions, we next learn multi-faceted prediction models.

Predictive Modeling

The final multi-faceted analysis consisted of the development of two sets of predictive models, one for each of the PANSS scales: positive and negative. Each model includes features that appeared as significant in the single-faceted analyses (see Section 2.4). For the positive PANSS scale, the features are the lexicon categories of power words and perceptual processing words, as well as perplexity. For the negative PANSS scale, the features are lexicon category of time words and the self-repair features of edits and restarts. As previously mentioned, all the single-facet analyses were performed on the training set, allowing for a fair evaluation of the prediction models on the test set (with new participants not in the training set).

Prediction experiments. We compare both ϵ-support vector machines [39] and multi-layer

TABLE 2.6
Mean Pearson's r correlation coefficient achieved over ten-fold cross-validation, hold-out testing on prediction of positive and negative PANSS scores.

PANSS Scale	SVM	MLP
Positive Scale	$+0.570$	$+0.879$
Negative Scale	$+0.566$	$+0.710$

perceptron models [64] for prediction of PANSS scales. These models were trained using ten-fold cross-validation for hyperparameter tuning on the training set, optimizing upon the Pearson's r correlation coefficient. Hyperparameters included the kernel (linear or radial basis function), $C = \{10^{-5}, 10^{-4}, \ldots, 10^{4}\}$, $\epsilon = \{10^{-5}, 10^{-4}, \ldots, 10^{-1}\}$, and $\gamma = \{0.00, 0.05, \ldots, 1.00\}$ (in the case of the RBF kernel) for the support vector machines, and the number of hidden units ($\{1, 5, 10, 50, 100, 500\}$) and activation function (logistic, hyperbolic tangent, or rectified linear unit) in the multi-layer perceptron. Test set results are summarized in Table 2.6. The multilayer perceptron significantly outperformed the SVM in both cases ($p < 0.01$ in both cases according to a one-way ANOVA).

Feature analysis. To examine the relative importance of the included features in the multi-layer perceptron model, a greedy step-wise feature selection process was performed, using a ten-fold cross-validation procedure over the entire set[2]. At each iteration, candidate features were evaluated, and the single best feature to be added was selected via the highest average change in Pearson's r (Δr). Results are summarized in Table 2.7.

Discussion. In our predictive modeling analysis, we compared the performance of support vector machines (SVMs) and multi-layer perceptrons on a prediction task for positive and negative symptom severity. Although SVMs performed reasonably on both tasks, they were outperformed by multi-layer perceptrons in both cases. A higher performance was observed in predicting positive symptom severity, which may suggest that an individual's language use is more reflective of positive symptoms than negative symptoms in general. While positive scores

[2] The full dataset was used in this step as a post-hoc analysis for feature importance.

TABLE 2.7
Tabulation of the most significant features in each of the multi-faceted predictive models.

Positive Scale

	Top Predictive Features	Δr
1	power words	+0.406
2	perceptual processing words	+0.336
3	perplexity	+0.046

Negative Scale

	Top Predictive Features	Δr
1	self-repair edits	+0.330
2	time words	+0.262
3	self-repair restarts	+0.239

were significantly predicted by lexical categories, negative scores were more significantly predicted by self-repairs. This may suggest that individuals with high negative scores have more difficulty in communication, while individuals with high positive scores are more characterized by what they speak about.

2.6 Discussion and Conclusions

Most psychiatric disorders are diagnosed with significant clinical evaluation of an individual's abnormalities in behavior patterns, but the complexity of the many ways these disorders can manifest can limit this evaluation. Multimodal behavior analysis systems have the potential to fill this gap, but limited work has focused on the computational analysis of spoken language, despite psychological evidence for its pertinence. The present analysis approached language in three facets — through lexical, structural, and disfluency perspectives — and exposed a series of exciting results within each category as well as within interactions between them.

Words of power are heavily associated with positive symptom severity. Power words, such as *superiority*, *important*, and *exploit*, emerged as significantly predictive of positive symptom severity. The most characteristic symptoms of the positive scale involve delusions and grandiosity, which are defined by unfounded and exaggerated self-opinion and convictions of superiority, so the capability to detect these symptoms through language use is critical. Furthermore, the proportion of words of power used by an individual was the feature providing the most influence in a predictive model for positive symptom severity, above all other features.

Lack of relative language is highly indicative of more severe psychotic symptoms. Although much work has identified reality monitoring as a particular difficulty for individuals with psychotic disorders, little to no work has examined how this difficulty might be reflected in language use. Our analyses revealed that a lack of contextual language — relative words such as *yesterday*, *lately*, and *nearby* — is highly predictive of both positive and negative symptom severity. The fewer of these words an individual uses, the more severe their psychotic symptoms in general.

Linguistic difficulty during cognitive processing can be related to negative symptom severity. Although speaking in cognitive processing terms does not strictly indicate negative symptom severity, the higher an individual's negative symptom score, the more they will self-repair (and specifically edit their language) while speaking in cognitive processing terms. This behavior is often indicative of hesitation while constructing the sentences, so it may be representative of the cognitive difficulties characteristic of the negative psychotic symptom scale.

Future work will delve into more symptom-specific analyses, as each of the positive and negative scales are subdivided into measures of seven different symptom items. Augmenting these analyses with those of audiovisual modalities also holds great promise for improving the explanatory power of these models. Through these analyses, we can achieve an even more nuanced characterization of psychotic disorders, which will constitute a significant step toward the design of future multimodal clinical decision support tools for computational phenotyping of mental illness.

Acknowledgments

This material is based upon work partially supported by National Science Foundation Award #1722822 and the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, and no official endorsement should be inferred.

Chapter 3

Nonverbal Behavior Dynamics

Many of the essential clues to the psychiatric condition of an individual lie within the nonverbal and communicative behavior patterns they express during social interactions. The analysis described in this chapter examines quantified patterns of gaze aversion across a set of individuals recently admitted to an inpatient psychotic disorder unit at a major psychiatric hospital. These patterns are used to inform the development of discriminative models with the task of predicting schizophrenic symptom severity from both a typological and a dimensional assessment perspective. The results expose a novel set of gaze aversion behaviors distinguishing between positive subtype schizophrenia, characterized by excessive behaviors such as hallucinations and grandiosity, and negative subtype schizophrenia, characterized by diminished behaviors such as blunted affect and emotional withdrawal.

The work described in this chapter first appeared in the following publication:

> Alexandria K. Vail, Elizabeth Liebson, Justin T. Baker, Louis-Philippe Morency. Visual Attention in Schizophrenia: Eye Contact and Gaze Aversion during Clinical Interactions. *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, Texas, 2017.
>
> https://doi.org/10.1109/ACII.2017.8273644

3.1 Overview

When assessing the psychiatric condition of an individual, medical professionals often rely on a subjective assessment of abnormality in nonverbal and communicative behaviors during clinical interviews and day-to-day interactions. Although expert clinicians have a lifetime of experience and knowledge from which to draw a diagnosis, accurate judgment of individual cases is often inhibited by time constraints, clinician fatigue, or merely the human inability to judge every dimension of a person's condition at once. These limitations can interfere with determining the most accurate and timely diagnosis, and by extension the most effective plan of treatment.

One approach to addressing this challenge is to augment the assessment of these medical professionals with tools that can provide objective, automated analysis of a person's behaviors during these focused interactions. These systems would be capable of evaluating behavior patterns regarding previously collected data of the same individual (perhaps despite changing clinicians), in addition to the information gained from a wider demographic set of individuals carrying similar diagnoses. Such a tool could offer more detailed insight into a person's psychiatric condition, allowing the attending clinician to reach a better-informed diagnosis.

In everyday interaction, eye contact is widely considered to be an important signifier of aggression, social rapport, confidence, or interest; on the other hand, the lack of eye contact is often considered an indicator of respect, submissiveness, or even anxiety [92]. As a result, abnormal patterns in eye contact and gaze aversion behaviors are often adopted as significant indicators of psychiatric disorders [149]. Unusual behavior in this space is often a critical indicator of psychiatric illness, most notably in anxiety, depression, and cases of high suicidality [13, 162].

In this paper, we present a detailed investigation of eye gaze behaviors for patients with schizophrenic symptoms. Our analysis focuses on identifying behavior markers differentiating two subtypes of schizophrenia: positive subtype and negative subtype [87]. These subtypes of schizophrenia have been shown to respond differently to a variety of treatment plans [146] and exhibit different predispositions to comorbid conditions [121]. These findings motivate our analysis, since they suggest that correct identification of schizophrenic subtype is critical to determining the appropriate course of treatment for a given individual. We analyze eye gaze patterns in the context of the patient's facial expressions, as well as the dialogue cues from the clinician. In the later part of this paper, our detailed analysis will inform the development of predictive models for schizophrenic subtypes (i.e., typological assessment) and for continuous symptom severity (i.e., dimensional assessment).

3.2 Related Work

Many psychiatric disorders cause disruption in the normal function of nonverbal or communicative behaviors of an individual [13, 70]. In particular, multiple studies have suggested the importance of identifying gaze aversion in depression and cases of high suicidality; individuals with depression are suggested to fixate more frequently [70] and maintain significantly less eye contact when speaking with an interviewer [162] than those without. An avoidance of eye contact has also been seen in individuals diagnosed with other adverse clinical states, such as attention deficit disorder or autism [163].

Some studies have suggested particular differences in gaze behavior in individuals diagnosed with schizophrenia. Rutter suggested that many of these individuals are behaviorally indistinguishable from the general population during conversations of no personal importance, but display markedly abnormal gaze aversion patterns when asked to speak about personal matters [138]. Bergman et al. supported this finding, and suggested that in these afflicted individuals, much of the nonverbal behavior expressed does not synchronize with the verbal utterances [13]. Interestingly, in this study a lack of eye contact was not only observed in the case of the diagnosed person, but in the interviewing clinician as well. Laing suggests that persons diagnosed with schizophrenia may feel particularly vulnerable or exposed under the gaze of others, and may actively avoid eye contact as a result [100]. The present analysis uses this to inform 'categories' of interview questions (see Section 3.3).

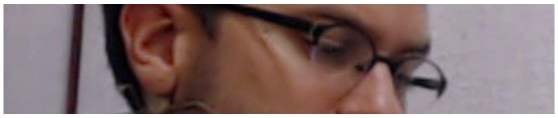
Our work examines a variety of gaze aversion behaviors regarding an individual's results on a clinical inventory of schizophrenic symptoms. Section 3.3 continues with a detailed description of the interview dataset and the various feature extractions performed upon it. Section 3.4 describes a set of hypothesis-driven experiments, which informed a predictive analysis described in Section 3.5. We interpret some significant features identified in Section 3.6. The report concludes with a brief overview and some thoughts toward future directions in Section 3.7.

3.3 Clinical Interview Dataset

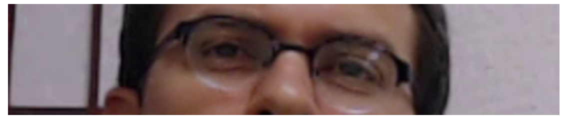
The dataset examined consists of a series of clinical interviews with adult individuals recently admitted to an inpatient psychotic disorder unit at McLean Hospital, a major psychiatric facility. Video and audio recordings were collected from 21 unique participants (six of whom were female). Each session involved a semi-structured clinical interview between the admitted individual and a clinician, lasting approximately 10–15 minutes each. The interview script was modeled upon existing everyday clinical interactions designed to elicit reactions that may be illustrative of the psychiatric condition of the individual.[1] A list of interview questions is presented in Table 3.1.

Following the conclusion of each interview, the participant was administered a series of clinical scales, including the Positive and Negative Syndrome Scale (PANSS) [87], a scale used for measuring schizophrenic symptom severity. PANSS involves seven-point ratings of 30 symptoms across three dimensions: *positive symptoms*, involving behaviors in excess or distortion of normal function, *negative symptoms*, involving behaviors diminished or suppressed below normal function, and *general psychiatric symptoms*, involving items that cannot be linked decisively to either syndrome. Items from the Positive and Negative scales are listed and described in Table 3.2.

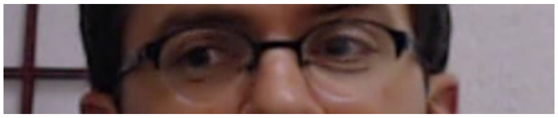
[1] Although participants varied in previous exposure to similar interactions, this diversity is reflective of the larger population, and we believe that this strengthens the applicability of this analysis.



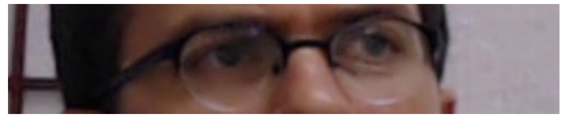
(*right*, *down*)



(*center*, *center*)



(*left*, *down*)



(*right*, *up*)

FIGURE 3.1
Example set of annotated gaze direction labels for sample video frames.

Participants are grouped by their PANSS *composite score*, defined as the difference between the positive and negative symptom scores [87]. Participants with a composite score above zero are classified as having *positive subtype* schizophrenia, whereas those below or equal to zero were classified as having *negative subtype* schizophrenia. Twelve of the participants are classified as expressing positive subtype schizophrenic symptoms, and nine are classified as expressing negative subtype. The average Positive Scale score in the present sample is $M = 17.48$ ($SD = 8.09$) and Negative Scale $M = 13.95$ ($SD = 3.92$), both in a possible range of 7 to 49; the average composite score is $M = 3.52$ ($SD = 9.35$), in a possible range of -42 to 42.

Gaze Aversion Annotation

Each session video was manually annotated for gaze behavior. This annotation task was conducted in two stages: annotation of lateral gaze direction and annotation of vertical gaze direction. Lateral gaze direction was manually classified into *left*, *center*, or *right*; similarly, vertical direction into *up*, *center*, or *down*. Note that an annotation of (*center*, *center*) would indicate gaze at the interviewing clinician and *left* and *right* are directions from the perspective of the interviewing clinician. When eye gaze direction was conflated with head gaze direction, the 'absolute' direction of aversion was taken. For an illustration of sample labels from this annotation scheme, see Figure 3.1.

To evaluate the reliability of this annotation scheme, a second annotator repeated this procedure on eight sessions (approximately 38% of the dataset). Each session was segmented by the tenth of a second, and inter-annotator agreement was calculated based on classification into each of the three directional states for each dimension. This resulted in a Krippendorff's alpha coefficient of $\alpha = 0.89$ for lateral movement and $\alpha = 0.76$ for vertical movement, each of which exceeds the usual threshold for a 'reliable' level of agreement [98].

Dialogue Annotations

Interview items were grouped into two distinct categories: *introspective questions*, in which the participant is asked to examine their thoughts, feelings, or mental state, and *extrospective questions*, in which the participant is asked to describe the state of their environment. Inter-annotator agreement across four independent annotators achieved a Krippendorff's alpha coefficient of $\alpha = 0.85$, a 'reliable' level of agreement [98]. This classification is presented in Table 3.1.

Annotation of interview dialogue involved selection of the moment at which each question segment began, accurate to the tenth of a second, as well as the classification of the question itself into one of thirteen questions types (see Table 3.1). To evaluate inter-annotator agreement, a second annotator repeated this procedure on five sessions (approximately 24% of the dataset). On average, there was a difference of 1.2 seconds regarding annotation of the start of a question. There were two instances of 'missed' question annotations and one instance of disagreement on question classification, out of a total of 48.

Facial Expression Feature Extraction

Facial expression for the current analysis is defined in terms of the Facial Action Coding System (FACS), a procedure designed to describe facial expression systematically via individual muscle movements [41]. Video recordings of both clinician and participant were collected at a resolution of 1280×960 pixels at 30 frames per second. Facial action unit intensities were extracted from these videos using OpenFace, a state-of-the-art open-source facial behavior analysis toolkit [7].

TABLE 3.1
Classification of interview protocol items into introspective questions and extrospective questions.

Introspective Questions	Extrospective Questions
Has anything in particular been on your mind?	What brought you into the hospital?
What are your goals for the hospitalization?	What has the team here been helping you with?
How is your mood/spirits?	Would you say that they are doing a good job?
How is your thinking/focus?	How have people been treating you?
How is your self-confidence compared to how it usually is?	How is the food?
What changes do you observe since you were hospitalized?	How is your energy?
	How have you been sleeping?

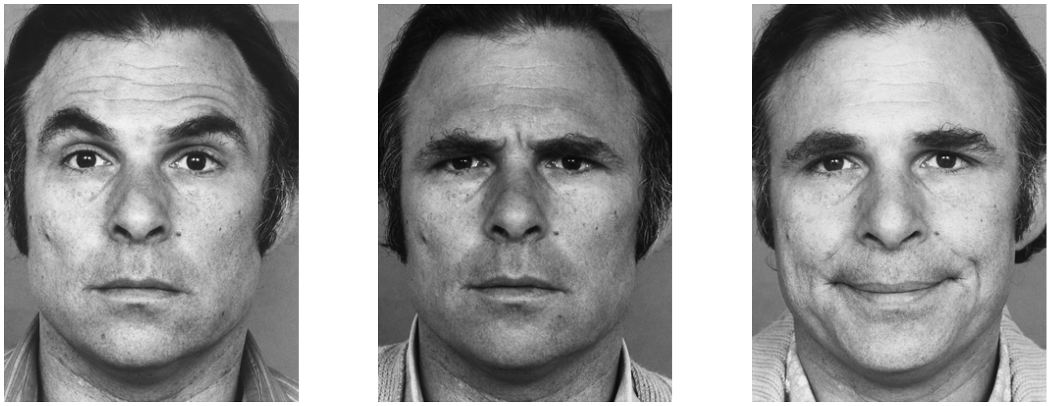
After processing with OpenFace, each frame of the video receives an intensity score $s_i \in [0, 5]$ for each of 17 facial action units, four of which are used in the present analysis. Frames with less than 70% confidence in the facial landmark detection results (often due to extreme head pose, rapid motion, or occlusion) were discarded. This threshold resulted in elimination of approximately 16% of the recorded video frames. The three facial action units most prominent in the present analysis are illustrated in Figure 3.2.

3.4 Statistical Analysis

Initial examination of the recorded interviews resulted in several qualitative observations, which informed the definition of hypotheses detailed in the following subsections. Each of these hypotheses were compared using the appropriate statistical models. Tests for normality and homoscedascity were performed before each test, and all reported p-values have been corrected using the Benjamini-Hochberg procedure for controlling the family-wise error rate within each hypothesis group. In the first section, we study overall differences in aversion behavior. The next section studies differences when contextualized within dialogue events, and the final section studies the interactions with facial expressions.

TABLE 3.2
Enumeration and brief description of a selection of symptoms contained in the PANSS positive and negative scales [87].

Scale Item	Brief Description of Behavior
Positive Scale	
Delusions	Beliefs which are unfounded, unrealistic, and idiosyncratic.
Conceptual Disorganization	Disorganized process of thinking characterized by disruption of goal-directed sequencing, e.g., circumstantiality, tangentiality, loose associations, non-sequiturs, gross illogicality, or thought block.
Hallucinatory Behavior	Verbal report or behavior indicating perceptions which are not generated by external stimuli. These may occur in the auditory, visual, olfactory, or somatic realms.
Grandiosity	Exaggerated self-opinion and unrealistic convictions of superiority, including delusions of extraordinary abilities, wealth, knowledge, fame, power, and moral righteousness.
Hostility	Verbal and nonverbal expressions of anger and resentment, including sarcasm, passive-aggressive behavior, verbal abuse, and assaultiveness.
Negative Scale	
Blunted Affect	Diminished emotional responsiveness as characterized by a reduction in facial expression, modulation of feelings, and communicative gestures.
Emotional Withdrawal	Lack of interest in, involvement with, and affective commitment to life's events.
Poor Rapport	Lack of interpersonal empathy, openness in conversation, and sense of closeness, interest, or involvement with the interviewer. This is evidenced by interpersonal distancing and reduced verbal and nonverbal communication.
Difficulty in Abstract Thinking	Impairment in the use of the abstract-symbolic mode of thinking, as evidenced by difficulty in classification, forming generalizations, and proceeding beyond concrete or egocentric thinking in problem-solving tasks.
Lack of Spontaneity and Flow of Conversation	Reduction in the normal flow of communication associated with apathy, avolition, defensiveness, or cognitive deficit. This is manifested by diminished fluidity and productivity of the verbal-interactional process.



(A) AU2 OUTER BROW RAISER **(B)** AU4 BROW LOWERER **(C)** AU14 DIMPLER

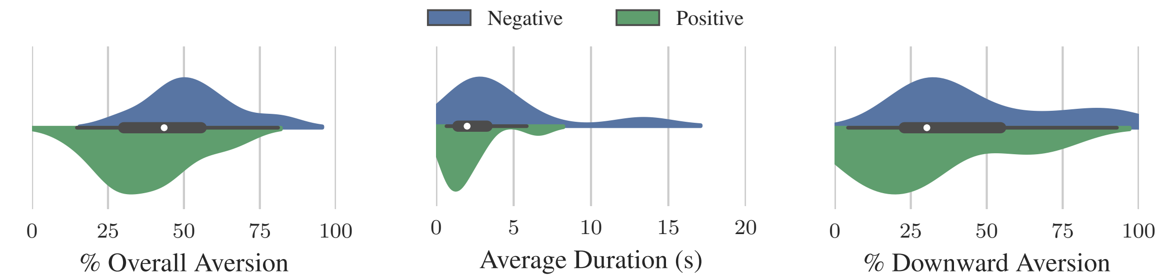
FIGURE 3.2
Illustration of the subset of facial action units used in the present analysis [41].

Aversion

The first set of hypotheses tested involved general trends in gaze aversion behaviors between individuals expressing positive and negative subtype schizophrenia.

H1.1. *Individuals expressing positive subtype schizophrenia avert their gaze less than those expressing negative subtype schizophrenia.* The first hypothesis examines the raw percentage of the interview in which participants are not averting their gaze from the interviewing clinician. This hypothesis is grounded in the understanding that individuals scoring highly on the positive symptom scale express such symptoms as hostility and suspiciousness, which may result in less gaze aversion. There was a statistically significant difference between groups at the 95% confidence level as determined by a one-way ANOVA [$F(1, 19) = 5.049$, $p = 0.037$] (see Figure 3.3a). A post-hoc comparison indicated that the average percentage of aversion over the session for individuals expressing positive subtype ($M = 38.34\%$, $SD = 14.86\%$) was significantly smaller than the average percentage for individuals expressing negative subtype ($M = 52.94\%$, $SD = 14.57\%$). This result suggests that individuals expressing positive subtype schizophrenia avert their gaze less often, in general, than individuals expressing negative subtype schizophrenia.

H1.2. *Individuals expressing negative subtype schizophrenia avert their gaze for longer pe-*



(A) **H1.1.** Percentage of the interview in which gaze was averted. $[F(1, 19) = 5.049, p = 0.037]$

(B) **H1.2.** Average duration of an aversion (in seconds). $[H(1) = 5.838, p = 0.016]$

(C) **H1.5.** Percentage of aversions that were (non-exclusively) downward. $[H(1) = 2.909, p = 0.088]$

FIGURE 3.3
Illustration of a selection of the distributions most significantly different between participants expressing positive- versus negative-subtype schizophrenic symptoms. As some distributions fail normality tests, we illustrate using the violin plot, an alternative to the traditional box plot that also accurately represents the distribution of the data using smoothed density plots. The center line represents the median and interquartile range of the dataset, much like a traditional box plot.

riods of time than individuals expressing positive subtype schizophrenia. The second hypothesis examines the average temporal length of gaze aversions when they do occur. This hypothesis is based on the defining features of negative symptoms such as poor rapport and social withdrawal, which may suggest more consistent aversion behavior. There was a statistically significant difference between groups at the 95% confidence level as determined by a Kruskal-Wallis H-test[2] $[H(1) = 5.838, p = 0.016]$ (see Figure 3.3b). A post-hoc comparison indicated that the average aversion duration for individuals expressing positive subtype schizophrenia ($M = 1.93$s, $SD = 1.63$s) was significantly smaller than for individuals expressing negative subtype ($M = 4.23$s, $SD = 3.69$s). This result suggests that when individuals expressing negative subtype schizophrenia avert their gaze, they are likely to do so for a longer period of time than individuals expressing positive subtype.

H1.3. *Individuals expressing positive subtype schizophrenia cover larger area during aversions than individuals expressing negative subtype schizophrenia.* The third hypothesis examines the average distance covered during gaze aversions. This hypothesis is based on the suggestion

[2]Both distributions failed a Shapiro-Wilk test for normality: positive subtype $[W(12) = 0.705, p = 0.001]$ and negative subtype $[W(9) = 0.705, p = 0.002]$.

that positive subtype schizophrenia involves a degree of hyperactivity and excitement, lending to fewer gaze fixations. To operationalize this definition, for each aversion event, each two-dimensional directional annotation is treated as a point in $\{-1, 0, +1\}^2$-space, and the Euclidean distance $\|\boldsymbol{x_i} - \boldsymbol{x_{i+1}}\|$ is calculated between every pair of consecutive points $\boldsymbol{x_i}$ and $\boldsymbol{x_{i+1}}$ along the aversion path. The sum of these distances results in a measure of the distance covered over the course of the aversion. There was *not* a statistically significant difference between groups at the 95% confidence level as determined by a Kruskal-Wallis H-test[3] [$H(1) = 1.823$, $p = 0.177$].

H1.4. *Individuals expressing positive subtype schizophrenia are more likely to avert their gaze laterally than individuals expressing negative subtype schizophrenia.* The fourth hypothesis examines the proportion of aversions that are (non-exclusively) lateral. Vertical aversions are often associated with anxiety, which is more canonically associated with the social withdrawal and poor rapport of negative subtype schizophrenia. There was *not* a statistically significant difference between groups at the 95% confidence level as determined by a Kruskal-Wallis H-test[4] [$H(1) = 1.548$, $p = 0.213$].

H1.5. *Individuals expressing negative subtype schizophrenia are more likely to avert their gaze downward than individuals expressing positive subtype schizophrenia.* The final hypothesis examines the proportion of aversions that are (non-exclusively) downward. Downward aversions have previously been suggested to be significantly indicative of individuals diagnosed with depression [70], which is often associated with many negative schizophrenic symptoms. There was *not* a statistically significant difference between groups at the 95% confidence level as determined by a Kruskal-Wallis H-test[5] [$H(1) = 2.909$, $p = 0.088$] (see Figure 3.3c).

[3] Both distributions failed a Shapiro-Wilk test for normality: positive subtype [$W(12) = 0.855$, $p = 0.043$] and negative subtype [$W(9) = 0.822$, $p = 0.036$].

[4] The positive subtype distribution failed a Shapiro-Wilk test for normality [$W(12) = 0.598$, $p = 0.000$].

[5] The negative subtype distribution failed a Shapiro-Wilk test for normality [$W(9) = 0.814$, $p = 0.029$].

Aversion and Dialogue

The second set of hypotheses tested involves eye contact and gaze aversion as related to dialogue and question types (see Section 3.3 for details).

H2.1. *Introspective questions result in more gaze aversion than extrospective questions.* The first hypothesis examines the difference in gaze aversion during introspective and extrospective questions. Introspective questions involve evaluating intimate details about the self, which often induces discomfort or unease. There was a statistically significant difference within subjects as determined by an ANOVA with repeated measures [$F(1, 20) = 7.347$, $p = 0.013$]. A post-hoc comparison indicated that the average proportion of aversion during introspective questions ($M = 53.70\%$, $SD = 21.21\%$) was significantly more than during extrospective questions ($M = 49.89\%$, $SD = 17.78\%$). This result suggests that regardless of subtype, individuals expressing schizophrenia are more likely to avert their gaze during introspective questions than during extrospective questions.

H2.2. *Individuals expressing negative subtype schizophrenia avert their gaze more often during introspective questions than individuals expressing positive subtype schizophrenia.* The second hypothesis suggests that individuals expressing negative subtype schizophrenia would avert their gaze more frequently during introspective questions than their positive subtype counterparts. This was informed by the prominent negative scale item involving difficulty in abstract thinking, which may result in difficulty answering this type of interview question. There was a statistically significant difference between groups as determined by a one-way ANCOVA while controlling for overall aversion percentage [$F(1, 18) = 6.486$, $p = 0.020$]. A post-hoc comparison indicated that the average proportion of aversion during introspective questions for individuals expressing positive subtype schizophrenia ($M = 41.33\%$, $SD = 13.66\%$) was significantly less than for individuals expressing negative subtype ($M = 61.81\%$, $SD = 16.12\%$). This result suggests that individuals expressing negative subtype schizophrenia are more likely to avert their gaze during introspective questions than individuals expressing positive subtype schizophrenia.

Aversion and Facial Expression

The final set of hypotheses examines the facial expressions conveyed during gaze aversions (see Section 3.3).

H3.1. *When averting gaze, individuals expressing positive subtype schizophrenia express more* AU2 OUTER BROW RAISER *than individuals expressing negative subtype schizophrenia.* The first hypothesis examines the average expression of AU2 OUTER BROW RAISER during gaze aversions. Brow raising is often associated with fear, surprise, and other spontaneous emotions [41], which may be more present in individuals expressing positive symptoms such as excitement and hyperactivity. There was a statistically significant difference between groups as determined by a one-way ANCOVA while controlling for average overall AU2 intensity [$F(1, 18) = 5.627$, $p = 0.029$]. A post-hoc comparison indicated that the average AU2 intensity expressed during aversion for individuals expressing positive subtype schizophrenia ($M = 0.847$, $SD = 0.316$) was significantly greater than for individuals expressing negative subtype ($M = 0.757$, $SD = 0.266$). This result suggests that individuals expressing positive subtype schizophrenia tend to express AU2 OUTER BROW RAISER when they avert their gaze more than individuals expressing negative subtype schizophrenia.

H3.2. *When averting their gaze, individuals expressing negative subtype schizophrenia express more* AU4 BROW LOWERER *than individuals expressing positive subtype schizophrenia.* The second hypothesis examines the average expression of AU4 BROW LOWERER during gaze aversions. Brow lowering is an expression canonically associated with negative emotions [41], which may be more present in individuals expressing negative subtype schizophrenic symptoms. There was a statistically significant difference between groups as determined by a one-way ANCOVA while controlling for average overall AU4 intensity [$F(1, 18) = 5.643$, $p = 0.029$]. A post-hoc comparison indicated that the average AU4 intensity expressed during aversion for individuals expressing negative subtype schizophrenia ($M = 0.125$, $SD = 0.053$) was significantly greater than for individuals expressing positive subtype ($M = 0.057$, $SD = 0.047$). This result suggests that individuals expressing negative subtype schizophrenia tend to express AU4 BROW LOWERER when they avert their gaze more than individuals expressing positive subtype schizophrenia. Prior work on individuals expressing schizophrenia without regard to subtype has identified this expression as generally indicative of schizophrenia [97], so the suggestion that this facial expression is expressed differently between subtypes is notable.

H3.3. *When averting their gaze, individuals expressing negative subtype schizophrenia express more* AU14 DIMPLER *than individuals expressing positive subtype schizophrenia.* The third hypothesis examines the average expression of AU14 DIMPLER during gaze aversions. AU14 DIMPLER is often associated with contempt, which may be more prevalent in individuals expressing negative subtype schizophrenia than those expressing positive subtype. There was *not* a statistically significant difference between groups as determined by a one-way ANCOVA while controlling for average overall AU14 intensity [$F(1, 18) = 3.922$, $p = 0.063$].

H3.4. *When averting their gaze, individuals expressing negative subtype schizophrenia express more* AU20 LIP STRETCHER *than individuals expressing positive subtype schizophrenia.* The final hypothesis examines the average expression of AU20 LIP STRETCHER during gaze aversions. AU20 is often likened to a 'grimace' of the face, which occurs relatively infrequently in social interaction, but prior work has suggested a particular aversion to 'negative affect' facial expressions in schizophrenia [114]. There was *not* a statistically significant difference between groups as determined by a one-way ANCOVA while controlling for average overall AU20 intensity [$F(1, 18) = 0.165$, $p = 0.689$].

3.5 Predictive Models

To approach prediction of schizophrenic symptom severity from both a typological and a dimensional assessment perspective, two sets of computational models were built. The first analysis approaches the typological perspective, with the target of predicting an individual's schizophrenic subtype based on gaze aversion behavior descriptors. The second analysis addresses the dimensional perspective, using these gaze aversion behavior descriptors to predict quantitative scores on the PANSS inventory [87].

Computational Descriptors

Based on the results of the statistical analyses conducted previously, a series of thirteen behavior descriptors were extracted from each interview session. This set of descriptors was provided as a set of features to both the typological and the dimensional predictive analyses.

Gaze aversion percentage. Over the course of the entire interview session, the percentage of time in which the participant averted their gaze from the interviewing clinician.

Gaze aversion percentage (introspective). Over the course of all introspective question segments (see Section 3.3), the percentage of time in which the participant averted their gaze from the interviewing clinician.

Aversion duration. Across the set of all aversion events, the average temporal duration of a gaze aversion.

Aversion distance. Across the set of all aversion events, the average distance covered in an aversion (see Section 3.4, H1.3. for operational definition). This allows for the distinction between fixation and gaze-wandering.

Lateral/vertical aversion percentage. (2 features) Across the set of all aversion events, the percentage of events in which the participant made a lateral/vertical aversion. A lateral/vertical aversion is an event in which the participant's gaze drifts (non-exclusively) laterally/vertically from direct gaze toward the interviewing clinician.

Directional aversion percentage. (4 features) Across the set of all aversion events, the percentage of events in which the participant made an aversion in one of the four cardinal directions: left, right, up, or down. A directional aversion is an event in which the participant's gaze drifts (non-exclusively) in that direction relative to direct gaze toward the interviewing clinician.

Average AU2 intensity during aversion. Across all aversion events, the average expressed

TABLE 3.3
Typological experiments. Performance of the automatically validated SVM classification model in terms of accuracy, Krippendorff's α, and F_1 score, as compared to a majority-class predictor baseline model.

Model	Accuracy	Krippendorff's α	F_1 Score
SVM	76.19%	0.5309	0.7597
Baseline	57.14%	−0.2424	0.3636

intensity of AU2 OUTER BROW RAISER (see Figure 3.2a).

Average AU4 intensity during aversion. Across all aversion events, the average expressed intensity of AU4 BROW LOWERER (see Figure 3.2b).

Average AU14 intensity during aversion. Across all aversion events, the average expressed intensity of AU14 DIMPLER (see Figure 3.2c).

Typological Assessment

The typological assessment is framed as a classification problem in which the target class value is either positive or negative subtype (see Section 3.3). A set of support vector machine (SVM) classifiers [39] were trained for this task using leave-one-person-out cross-testing, following leave-one-person-out cross-validation for hyperparameter tuning and feature selection using logistic regression [161]. Models were validated upon Krippendorff's α. The model was allowed to take on either a linear kernel $K(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{x}^T\boldsymbol{x'}$ or a Gaussian radial basis function (RBF) kernel $K(\boldsymbol{x}, \boldsymbol{x'}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{x'}\|^2)$, for any two feature vectors $\boldsymbol{x}, \boldsymbol{x'} \in \mathbb{R}^9$. Hyperparameters validated include $C \in \{10^{-5}, 10^{-4}, \ldots, 10^4\}$ and, in the case of the Gaussian-RBF kernel, $\gamma \in \{0.00, 0.05, \ldots, 1.00\}$.

Performance of cross-testing in terms of accuracy, Krippendorff's α, and F_1 score is displayed in Table 3.3, alongside a baseline majority-class predictor. This classification model achieved a performance well above the majority-class baseline during cross-testing. Although the Krippendorff's α does not reach a 'reliable' level of agreement [98], the moderate level of performance achieved does suggest the existence of significant information regarding the identification of

TABLE 3.4
Dimensional experiments. Performance of the automatically validated ϵ-SVR regression models in terms of Pearson's r.

Model	Pearson's r	p
Positive Score	0.5853	0.005
Negative Score	0.4330	0.049
Composite Score	0.5714	0.006

schizophrenic subtype in an individual's gaze aversion behaviors.

Dimensional Assessment

The second task of dimensional assessment is framed as a regression problem in which the target class value is either the individual's total Positive Scale score (values 7 to 49), the individual's total Negative Scale score (values 7 to 49), or the individual's composite score (values -42 to 42). A series of ϵ-support vector regressors (ϵ-SVRs; [39]) were trained for this task using leave-one-person-out cross-testing, following leave-one-person-out cross-validation for hyperparameter tuning and feature selection using LASSO [153]. Models were optimized upon Pearson's r. The model was validated upon the same hyperparameters specified in Section 3.5; in addition, the range parameter ϵ was validated within $\epsilon \in \{10^{-5}, 10^{-4}, \ldots, 10^{-1}\}$.

Performance of the best-performing regression models in terms of Pearson's r is displayed in Table 3.4. All three models were able to achieve a reasonable level of correlation with true PANSS scores during cross-testing. All of these correlations were statistically significant at the 95% confidence level. Prediction of raw dimensional scores is a more complex task than prediction of coarse typological subtype, but the promising results achieved reinforce the proposition that gaze aversion behavior is a prominent social signal containing information relevant to the identification of schizophrenic symptom severity.

3.6 Behavior Analysis

The final stage of this analysis examines one of the predictive models in detail, identifying and interpreting the significance of the most influential features. For this final step, a LASSO linear model [153] was trained upon the entire dataset, optimizing performance on composite score prediction in terms of Pearson's r. The model was limited to a selection of five features that best predicted the PANSS composite score of the participants. The LASSO model achieved a Pearson's $r = 0.65$ on the training set (compare to model performance in Section 3.5; note that this is performance on the training set, rather than leave-one-person-out validation). We review the five features selected; the model is presented in Table 3.5.

Gaze aversion during introspective questions. The most influential feature selected is the percentage of introspective question segments in which the individual is averting their gaze from the clinician. The more the participant averts their gaze during introspective questions, the lower their composite score tends to be, and by extension, the more negative symptoms they tend to express. This result was mirrored in Section 3.4, where there existed a statistically significant difference in aversion during introspective questions between individuals expressing positive subtype and negative subtype schizophrenic symptoms.

Average intensity of AU4 BROW LOWERER during gaze aversion. The next feature selected is the average intensity of AU4 BROW LOWERER (see Figure 3.2b) during aversion events. The more intense the average brow lowering during gaze aversion, the lower the participant's composite score tends to be, and the more negative symptoms they tend to express. This result was also mirrored in Section 3.4, where there existed a statistically significant difference in AU4 expression during aversion events between individuals expressing positive and negative symptoms.

Proportion of lateral gaze aversion. The only positively correlated feature selected is the proportion of gaze aversions that were (non-exclusively) lateral aversions. The more gaze aversions in which the participant's gaze moves laterally, the higher their composite score tends to

TABLE 3.5
Features selected by a LASSO linear model, when limited to five features, predicting the PANSS composite score of the participant.

PANSS Composite Score =
$-8.374 \times$ Gaze aversion during introspective questions
$-4.760 \times$ Average intensity of AU4 during gaze aversion
$+1.972 \times$ Proportion of lateral gaze aversion
$-0.725 \times$ Proportion of downward gaze aversion
$-0.001 \times$ Average gaze aversion duration
Pearson's $r = 0.653$, $p = 0.002$

be, and the more positive symptoms they tend to express. Interestingly, this descriptor was *not* discriminative on its own in the statistical analyses in Section 3.4. This may suggest that it holds more discriminative information when combined with these other features.

Proportion of downward gaze aversion. The next feature selected is the proportion of gaze aversions that were (non-exclusively) downward aversions. The more gaze aversions in which the person looks downward, the lower their composite score tends to be, and the more negative symptoms they tend to express. This descriptor was also not considered a discriminative feature in the statistical analyses in Section 3.4, although it was more significant than lateral aversion.

Average gaze aversion duration. The final feature included, with relatively little influence, is the average length of time of an aversion event. The longer periods of time a person averts their gaze, the lower their composite score tends to be, and the more negative symptoms they tend to express. Although this descriptor was significantly discriminative between individuals expressing positive and negative subtype symptoms in Section 3.4, it was not very influential in this model; this may suggest that, although this feature is still discriminative, the prior features explain the difference more accurately than gaze aversion duration.

3.7 Conclusion

Most psychiatric disorders are diagnosed with significant clinical evaluation of an individual's nonverbal and communicative behavior patterns. The present analysis aims to develop classifier models that can accurately differentiate between subtypes of schizophrenic symptoms based on the patterns of eye contact and gaze aversion expressed by an individual during a clinical interview. A strength of this work is the approach to these behaviors through an investigation of symptom severity rather than coarse-grained diagnoses; since many symptoms are shared across comorbid conditions, this work can inform systems developed toward more personalized symptom-based care.

Statistical comparisons suggest a few interesting differences in behavior between positive and negative symptoms of schizophrenia. In general, individuals expressing negative-subtype schizophrenic symptoms tend to avert their gaze from the clinician more and for longer periods of time, and this difference is even more notable during introspective questions. When these individuals do avert their gaze, they tend to lower their brows (AU4 BROW LOWERER) more than individuals expressing positive symptoms.

We have reported a predictive model able to distinguish between positive and negative subtype expressing individuals with reliable performance based on gaze aversion behaviors during a clinical interview. In addition, predictive models can reasonably predict PANSS numeric scores on the Positive Scale and the Negative Scale, as well as the composite difference score. We identify the most influential behavior descriptors and potential interactions between them; most notably, the direction of gaze aversion becomes a discriminative feature when taken in concert with other descriptors. By approaching computational identification of schizophrenic symptom intensity from both a typological and dimensional perspective, this line of work constitutes a promising step in the development of technologies to aid clinicians in diagnosis of psychiatric illnesses.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program and the BrainHub Postdoctoral Fellowship Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, positions, or policies of the National Science Foundation or BrainHub.

Chapter 4

Proposed Work

Cross-Modal Behavior Dynamics

Decades of research has emphasized that face-to-face interaction relies upon multiple modalities simultaneously, from language to gestures to visual gaze to facial expression. Having studied behavior dynamics within both verbal (Chapter 2) and nonverbal (Chapter 3) modalities, we are now in a position to consider *cross-modal* behavior dynamics. The idea of cross-modal behavior can be reasonably related to Aristotle's famous quote, "The whole is greater than the sum of its parts". Cross-modal analysis seeks to identify what knowledge can be gained from multiple modalities in unison that cannot be gained from single modalities in isolation. In particular, we aim to study cross-modal behavior through two perspectives: *monadic* cross-modal behavior (e.g., client verbal $\times$ client nonverbal), and *dyadic* cross-modal behavior (e.g., therapist verbal $\times$ client nonverbal).

We propose to take advantage of the same data used in Chapter 2 and Chapter 3. This dataset consists of a set of clinical interview series with adult individuals recently admitted to an inpatient unit at a major psychiatric hospital. These semi-structured interviews were designed to emulate existing everyday clinical interactions, such as those of brief daily check-ins by clinical staff. Each interview was recorded by video, audio, and transcription of both client and therapist behavior. From these input modalities, we can examine several behavioral modalities; for example, language use, facial expression, head and eye gaze, prosodic patterns, and gestures. Each interview was followed by administration of a set of clinical scales to evaluate multiple dimensions of psychiatric health, including symptoms associated with psychosis [87], mood disorders [116], or mania [169]. Using this rich multimodal dataset, we aim to explore the modeling of cross-modal behaviors during particular moments of interest as predictive indicators of psychiatric symptoms. We express this overarching research question:

How can we model and interpret cross-modal behavior to gain insight into the psychological health of the client that we cannot necessarily gain from unimodal behavior alone?

4.1 Cross-Modal Perspectives

We slice this multifaceted research question into two parts: (1) within-person *monadic* cross-modal behavior, and (2) between-person *dyadic* cross-modal behavior.

Monadic Cross-Modal Behavior

The monadic examination of the client's cross-modal behavior is much like the most traditional understanding of 'multimodal'. Here, we cannot presume that one modality is more influential than the others. This proposed work aims to identify particularly informative behaviors during therapy sessions that may offer valuable insight into the psychological health of the client. For example, we know that the presence of moments of heightened client emotion are powerful indicators of reduction in symptoms [101, 122]. However, we also know that it is precisely during moments of heightened emotion when the information conveyed by the verbal and nonverbal behaviors of an individual are inconsistent [9]. During these events, such inconsistency could be seen as the client speaking in one way (verbal behavior) but gesturing to imply something else (nonverbal behavior). For example, consider the case where a client tells the therapist that they're feeling confident, but they are also averting their eyes downward while they speak. We learn more about the client by the existence of this conflict than we would if we studied each component in isolation. These circumstances are the objective of our study in this proposed work on monadic cross-modal behavior.

Dyadic Cross-Modal Behavior

When studying dyadic cross-modal behaviors, we will need to consider the conversational context from a different angle. Unlike the monadic case, the relationship between modalities is likely asymmetric: the client's nonverbal behaviors in this context are reactions to an external actor instead of the self. The client cannot react before the therapist starts speaking, but it is this reaction that we aim to investigate. How does the client avert their eyes after the therapist stops speaking? How does the client offer backchanneling responses during a therapist's question?

One of the most frequently observed forms of dyadic cross-modal interaction is known as backchanneling. Backchanneling responses are non-intrusive interjections that signal the listener's attention, interest, understanding, or attitude towards the speaker's message. These responses can be verbal, non-verbal, or both. When employed appropriately, backchanneling is associated with improvements in many social outcomes, such as rapport, conflict resolution, and negotiation [38, 55, 156]. Studies have shown that psychological health affects unimodal backchanneling behaviors. For example, individuals diagnosed with social anxiety tend to less accurately reflect genuine versus polite smiles during conversation [66], while individuals with symptoms of depression generally express fewer verbal backchanneling responses compared to those without these symptoms [166]. What remains relatively understudied in this context, however, is how backchanneling across modalities informs us about psychological health; e.g., how does the frequency of the client smiling when the therapist asks about their family indicate the severity of their depressive symptoms?

4.2 Analysis

Given these two research aims, we will enumerate a set of specific behavior markers that we expect to have a strong relationship with symptom severity. These hypotheses will be drawn from the results observed in previous chapters (Chapter 2, Chapter 3) as well as from similar work in unrelated domains, such as education [37, 167] or rapport [29, 58]. Some of the behavior markers that we intend to investigate include,

- *lateral gaze aversions*, identified in Chapter 3, and the use of *words of power*, identified in Chapter 2, as indicators of the severity of positive symptoms of psychosis;
- *speaking disfluencies*, identified in Chapter 2, and *downward gaze aversions*, identified in Chapter 3, as indicators of the severity of negative symptoms of psychosis;
- *"false smiles"*, the expression of smiling with the lips but not the eyes, identified by D'Mello and Graesser [37] as an indicator of student frustration; and
- *emotional expressivity*, identified by Gratch et al. [58] as an indicator of "genuine" emotional storytelling between acquaintances.

We are primarily interested in the *intersection* of these behaviors: e.g., how emotionally expressive is the client when using words of power? During what kinds of therapist speech does the client exhibit different forms of smiling?

We will begin with a computational analysis of these behaviors as they relate to the severity of the client's symptoms. This analysis allows us to narrow our focus to behavior markers that are most influential in determining the severity of the client's symptoms. Due to the nested structure of our recorded client-therapist interactions, we will use a multilevel modeling approach to account for multiple sessions per client, building upon our initial exploration of multilevel models in Chapter 3. After identifying a salient set of behaviors to investigate further, we will then turn to the development of our principal model.

One specific approach we will consider involves the adoption of a Bayesian perspective [52].

Bayesian methods provide a means of augmenting pre-existing domain knowledge (in the form of a prior distribution) with data-driven updates (in the form of observed data) to construct more robust models than either technique can achieve individually, especially with smaller datasets [52]. Since each cross-modal behavior marker will develop its own posterior distribution (the estimated distribution after observed-data updates), Bayesian models are highly interpretable. The posterior distribution allows us to study both the central tendency and the spread of each parameter, as well as a form of "confidence" (specifically, the probability of direction, pd) of its estimates. By examining these models, we will be able to draw meaningful practical knowledge regarding the relationship between these cross-modal behavior markers and client symptoms.

Part II

Social Behavior

"To say that the human being behaves individually at one moment and socially at another is as absurd as to declare that matter follows the laws of chemistry at a certain time and succumbs to the supposedly different laws of atomic physics at another..."

– Edward Sapir, 1927 [140]

Chapter 5

Facilitative Behavior Dynamics

Clients may terminate therapy for various reasons, but one of the most common causes is the lack of a strong working alliance. The concept of working alliance captures the collaborative relationship between a client and their therapist when working toward the progress and recovery of the client seeking treatment. In this chapter, we demonstrate that analysis of the facilitative behaviors that the participants use throughout the interaction may aid in identifying a weak working alliance before early client dropout. The work in this chapter focuses on the head gestures of both the client and therapist, contextualized within conversational turn-taking actions between the pair during psychotherapy sessions. We identify multiple behavior patterns suggestive of an individual's perspective on the working alliance; interestingly, these patterns also differ between the client and the therapist. These patterns inform the development of predictive models for self-reported ratings of working alliance, which demonstrate significant predictive power for both client and therapist ratings.

The work described in this chapter first appeared in the following publication:

> Alexandria K. Vail, Jeffrey Girard, Lauren M. Bylsma, Jeffrey F. Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. Goals, Tasks, and Bonds: Toward the Computational Assessment of Therapist Versus Client Perception of Working Alliance. *Proceedings of the Sixteenth International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 2021.*
>
> https://doi.org/10.1109/FG52635.2021.9667021

5.1 Overview

Previous research has established that the strength of the relationship between a client and their therapist is a robust predictor of positive therapy outcomes [10, 74, 110]. Much of the current psychological literature on the client-therapist relationship pays particular attention to what is known as the *working alliance*. Although many variations on the definition of 'working alliance' can be found, there is a consensus on the central idea that the working alliance captures the *collaborative* aspect of the therapist-client relationship [16, 74]. Higher therapist-reported and especially client-reported ratings of the working alliance have been strongly associated with reduction of the client's symptoms and concerns [51, 73, 74], but also with other positive therapy outcomes such as reduced drug abuse and recidivism [106] and improved medication compliance [47]. Of particular note is the recognized relationship between the strength of the working alliance and client dropout [47, 95, 141]. Proactive detection is especially valuable in this case: by the time a client has decided to quit therapy, the time for potential intervention has already passed. Understanding the complexity of the therapist-client relationship is crucial for informed treatment decision-making.

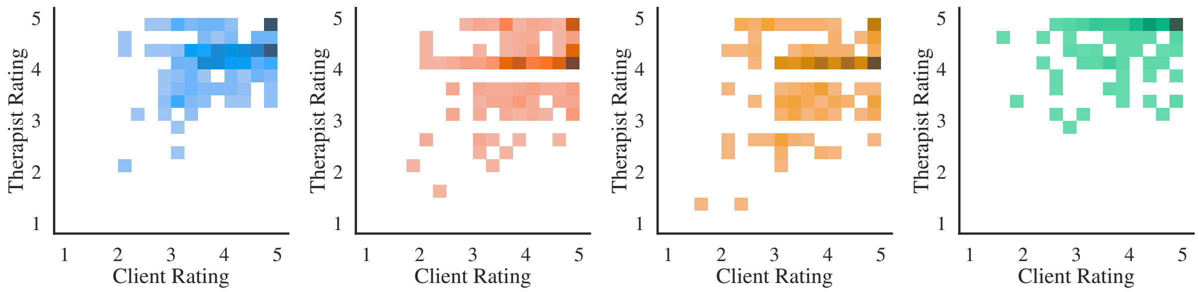
Unfortunately, measuring the strength of a working alliance faces several challenges. Most recorded ratings of the working alliance are obtained by self-reports from the client and their therapist, who are also participants in the interaction; previous research has documented significant divergences in these two participants' perception of the working alliance. Clients are often hesitant to express feedback or concerns [132, 134]: many clients do not express any concern at all until they have already decided to discontinue treatment [69]. On the other hand, therapists often miss subtle signs of client discontent during therapy sessions [132]. Alarmingly, some studies have even demonstrated that therapists may perform worse than chance at identifying signs of client frustration or annoyance [68, 110]. Several attempts have been made to evaluate the reliability of third-party human observers, but to date, observer ratings of the working alliance have repeatedly emerged as the least valuable predictors of therapy outcomes [74, 110].

The primary aim of this paper is to explore the use of computational behavior analysis to overcome the obstacles facing the objective measurement of the working alliance. Our analysis focuses primarily on head gestures and turn-taking behaviors, as these features have been identified as essential signals in the detection of similar measures of relationship [29, 58]. We begin with a set of inferential analyses to explore general trends in behavior that may indicate a participant's perception of the working alliance. Given these identified patterns, we develop a series of predictive models to estimate the working alliance ratings provided by the therapist and the client. Following this, we perform a set of ablation studies to examine the value of including specific categories of behavioral features, such as therapist behavior versus client behavior or head gesture features versus turn-taking features. Finally, we conclude by discussing some of the most notable takeaways revealed by these results and the promising directions for future work.

5.2 Related Work

To date, there has been little to no computational behavior analysis of working alliance in psychotherapy. However, there is a large volume of published studies in the computational literature that explores a similar construct: *rapport*, which can broadly be defined as mutual attentiveness, amiability, and receptivity between interaction participants [154]. Rapport differs from the working alliance in several fundamental ways, but one of the most notable differences is that rapport is generally considered to be 'other-focused', in which the primary goal is to develop a relationship between participants [154]. In contrast, the working alliance is 'task-focused', in which developing the relationship is secondary to the accomplishment of mutual goals [16]. The working alliance is more commonly described in asymmetric interactions, such as between therapist and client or teacher and student [74]. However, both concepts are related to relationship-building, and given the relative paucity of studies investigating working alliance computationally, we draw insight from the considerable amount of literature on the similar concept of rapport.

In previous studies of dyadic interaction, different behaviors have been shown to be related



(A) Overall Scores **(B)** Goal Subscale Scores **(C)** Task Subscale Scores **(D)** Bond Subscale Scores

FIGURE 5.1
Heatmap distributions of client and therapist ratings of working alliance and its subscales.

to rapport-building. One such behavior is head gestures: nodding is recognized as one of the most valuable indicators of rapport between human participants [154]. To a lesser extent, head shakes are also related to rapport in therapeutic contexts [156]. A growing body of literature has investigated the incorporation of rapport-building when designing virtual agents; gestures of both head and hands have been identified as some of the most influential behaviors for inclusion [58, 136].

Significant attention has also been paid to turn-taking behaviors in 'listening' agents [30]. Appropriate backchanneling (verbal and nonverbal) is critical to developing user trust [14]. Similarly, increased pauses have been recognized as positively impacting rapport-building, in terms of waiting to 'grab the floor' after a partner's dialogue turn but also within a turn, allowing the partner to 'grab the floor' themselves [28]. Taking longer dialogue turns — speaking for longer periods before transitioning to the partner — significantly impairs the development of rapport between participants [29]. Given that the therapeutic setting is an asymmetric interaction, 'listening' behaviors are especially pertinent in this context.

5.3 Dataset

Audiovisual recordings were collected from 266 therapy sessions between 39 unique clients and 11 unique therapists. Each therapist met with an average of 3.6 unique clients, and each client

TABLE 5.1
Sample items from both therapist and client versions of the Working Alliance Inventory.

Goal Subscale	Task Subscale	Bond Subscale
[Therapist] and I collaborate on setting goals for my therapy.	What I am doing in therapy gives me new ways of looking at my problem.	I believe [Therapist] likes me.
[Therapist] and I have established a good understanding of the kind of changes that would be good for me.	[Therapist] and I agree on what is important for me to work on.	I feel that [Therapist] appreciates me.
We are working towards mutually agreed upon goals.	[Client] and I agree about the steps to be taken to improve his/her situation.	I feel [Therapist] cares about me even when I do things that he/she does not approve of.
[Client] and I have a common perception of his/her goals.	[Client] and I both feel confident about the usefulness of our current activity in therapy.	I appreciate [Client] as a person.
		[Client] and I respect each other.

participated in an average of 6.8 sessions lasting between 40 and 60 minutes each (average 50.3 minutes).

Potential participants were recruited from a research registry, printed material advertising the study, and word-of-mouth. To be included in the study, participants had to be adults aged 18–65, meet DSM-V criteria for a major depressive disorder, currently experience at least moderate depressive symptoms (as measured by a Hamilton Rating Scale for Depression score ≥ 14; [61]), and be willing and able to provide informed consent. Individuals with a comorbid psychotic disorder, active suicidal or homicidal ideation, chronic depression, or current substance or alcohol abuse were excluded from the study. If an individual was suspected of experiencing psychosis or active suicidal ideation with intent or plan to harm themselves, the investigator terminated the screening interview and ensured that the individual obtained appropriate care, including but not limited to a referral to the psychiatric emergency room.

Included clients ranged from 22 to 65 years of age; 77% identified as female, and 62% identified as white. Clients were randomly assigned to an eight-session course of one of two psychotherapy conditions: cognitive behavioral therapy (CBT; 21 clients, 6 therapists) or interpersonal psychotherapy (IPT; 18 clients, 5 therapists).[1]

Ratings of Working Alliance

Following the conclusion of each therapy session, both therapist and client participants completed the therapist and client versions of the revised short-form Working Alliance Inventory (WAI; [65]), a widely used measure of alliance in therapy. The WAI consists of three subscales capturing three aspects of working alliance:

- the *goal* subscale, which assesses the individual's belief that participants agree on the overall objectives of the treatment;
- the *task* subscale, which assesses the individual's belief that participants agree on the steps required to reach the goals mentioned above; and
- the *bond* subscale, which assesses the individual's respect and trust for the other participant in an emotional sense.

Each subscale consists of a set of statements which the individual rates on a five-point Likert-type scale ranging from 'seldom true' to 'always true'. Representative items for each subscale are presented in Table 5.1, and the distribution of scores observed in our dataset is illustrated in Figure 5.1.

Head Gesture Annotation

Head motion for each participant was automatically measured using the OpenFace facial behavior analysis toolkit [8]. Gestures of interest in the present study were limited to head nods (vertical motion along the pitch dimension) and head shakes (horizontal motion along the yaw dimension). A low-resource algorithm was selected to classify head gestures based on prior work using basic dimensions of motion [82, 86, 164]. Although these works derived head motion from

[1]There were no statistically significant differences in working alliance ratings observed between the two treatment conditions.

$$
\begin{aligned}
\text{WAI} \sim 1 + \text{feature}_{\text{avg}} &+ \underbrace{(\text{feature}_{\text{dev}})}_{\text{session-level component}} \\
&+ \underbrace{(1 + \text{feature}_{\text{dev}} \mid \text{therapist})}_{\text{therapist-level component}} \\
&+ \underbrace{(1 + \text{feature}_{\text{dev}} \mid \text{therapist} : \text{client})}_{\text{client-level component}}
\end{aligned}
$$

FIGURE 5.2
Inferential model specification in formula notation.

the motion of a particular facial landmark between the eyes, our implementation instead incorporates head motion derived from the head tracking features provided by OpenFace [170]. Total distance traveled along each dimension was calculated over a rolling window of one second, and gestures were detected based on the top quartile of distance traveled within one second.

Speaking Turn Annotation

We define a 'speaking turn' as a contiguous speech segment from a single speaker until a non-speaking pause longer than one second. To determine speaking turns throughout the session, we performed speaker diarization (i.e., identifying when each speaker is actively speaking) using a voice activity detection algorithm available through openSMILE [43]. By applying this detection algorithm to each of the two participant microphones (client and therapist), the resulting annotations indicate whether the client or the therapist is presently speaking or, occasionally, if both are speaking.

5.4 Analysis

The present analysis consists of three stages. We begin with a set of inferential models to identify meaningful relationships between participant behaviors and working alliance ratings. We then incorporate these behaviors into a set of predictive models to estimate working alliance ratings.

TABLE 5.2
Summary statistics for features derived from head gestures and turn-taking behaviors.

	Client Behavior		Therapist Behavior	
	Mean	SD	Mean	SD
Head Nods (#)	208.25	47.75	208.89	50.04
Head Shakes (#)	162.07	71.58	167.51	52.17
Turn Length (s)	2.817	1.007	3.333	3.173
Wait Time (s)	1.305	1.899	1.854	1.483
Listening Nods (%)	0.229	0.070	0.236	0.078
Listening Shakes (%)	0.184	0.081	0.221	0.065

Finally, we perform a set of ablation studies to examine the value of including specific categories of behavior features: (1) client behavior vs. therapist behavior, and (2) head gestures vs. turn-taking behaviors.

Our feature set is primarily composed of the two sets of features derived from head gestures and speaking turns, as described in Section 5.3. Four additional features were derived from head gestures and turn-taking behaviors to identify head gestures while listening. We define therapist 'listening nods' as the percentage of client turns during which the therapist nods their head; a similar feature for client 'listening nods' is also computed for the client. We also define two 'listening shakes' features in the same manner for the head shake gestures of either client or therapist while listening. Our complete feature set, computed at the session level, consists of six features: head nods, head shakes, speaking turn length, wait time (pause length between the end of the partner's turn and the start of the speaker's), listening nods, and listening shakes. Summary statistics for all features are presented in Table 5.2.

Inferential Analysis

Due to the nested structure of our recorded client-therapist interactions, we utilize a multilevel modeling approach to account for multiple sessions per client and multiple clients per therapist.

TABLE 5.3

Client Ratings — Population-level effects from inferential models of working alliance ratings.

		Client Behavior			Therapist Behavior		
		Median	89% HDI	Sig.	Median	89% HDI	Sig.
OVERALL SCALE	Head Nods (#)	5.93	[0.00, 9.06]	⋆	−2.42	[−7.89, −0.15]	
	Head Shakes (#)	−6.89	[−6.02, −2.43]	⋆⋆	−2.20	[−6.89, 2.58]	
	Turn Length (s)	−0.14	[−0.48, 0.17]		−0.05	[−0.27, 0.17]	
	Wait Time (s)	−0.01	[−0.11, 0.10]		−0.01	[−0.12, 0.09]	
	Listening Nods (%)	4.29	[1.57, 7.11]	⋆⋆	−1.66	[−5.18, 1.68]	
	Listening Shakes (%)	−4.17	[−6.05, −2.15]	⋆⋆	−3.16	[−6.70, 0.48]	
GOAL SUBSCALE	Head Nods (#)	3.27	[−1.83, 6.24]		−4.18	[−9.79, −0.08]	⋆
	Head Shakes (#)	−6.58	[−6.35, −0.26]	⋆⋆	−1.25	[−7.99, 3.22]	
	Turn Length (s)	−0.18	[−0.53, 0.17]		−0.11	[−0.34, 0.12]	
	Wait Time (s)	0.01	[−0.11, 0.12]		−0.01	[−0.12, 0.10]	
	Listening Nods (%)	3.62	[0.52, 6.71]	⋆	−2.67	[−6.45, 0.98]	
	Listening Shakes (%)	−4.37	[−6.43, −2.21]	⋆⋆	−3.42	[−7.25, 0.48]	
TASK SUBSCALE	Head Nods (#)	4.54	[−0.31, 10.47]		−5.32	[−9.24, 0.10]	⋆
	Head Shakes (#)	−7.27	[−10.22, −3.16]	⋆⋆	−2.48	[−7.88, 0.28]	
	Turn Length (s)	−0.12	[−0.48, 0.25]		−0.02	[−0.27, 0.23]	
	Wait Time (s)	−0.03	[−0.15, 0.10]		−0.04	[−0.16, 0.08]	
	Listening Nods (%)	5.51	[2.34, 8.48]	⋆⋆	−1.27	[−5.04, 2.58]	
	Listening Shakes (%)	−4.67	[−6.91, −2.40]	⋆⋆	−3.83	[−7.82, 0.31]	
BOND SUBSCALE	Head Nods (#)	4.45	[−0.83, 7.83]	⋆	−1.84	[−4.54, 4.27]	
	Head Shakes (#)	−4.71	[−6.51, 0.24]	⋆	−1.89	[−7.01, 6.01]	
	Turn Length (s)	−0.18	[−0.52, 0.15]		−0.04	[−0.27, 0.19]	
	Wait Time (s)	−0.01	[−0.12, 0.11]		0.01	[−0.09, 0.12]	
	Listening Nods (%)	3.57	[0.41, 6.60]	⋆	−0.87	[−4.28, 2.53]	
	Listening Shakes (%)	−3.07	[−5.36, −0.77]	⋆	−2.70	[−6.37, 1.12]	

HDI = highest density interval, Sig. = significance, $\star\ pd > 95\%$, $\star\star\ pd > 99\%$.

TABLE 5.4

Therapist Ratings — Population-level effects from inferential models of working alliance ratings.

		Client Behavior			Therapist Behavior		
		Median	89% HDI	Sig.	Median	89% HDI	Sig.
OVERALL SCALE	Head Nods (#)	−1.04	[−2.72, 1.73]		0.85	[0.87, 4.18]	
	Head Shakes (#)	−2.33	[−4.79, −0.79]		−1.25	[−3.20, 0.10]	⋆⋆
	Turn Length (s)	0.07	[−0.03, 0.17]		−0.07	[−0.24, 0.11]	
	Wait Time (s)	−0.02	[−0.07, 0.03]		−0.02	[−0.07, 0.03]	
	Listening Nods (%)	−0.82	[−2.51, 0.84]		2.13	[0.87, 3.36]	⋆⋆
	Listening Shakes (%)	−1.36	[−3.13, 0.43]		−0.87	[−1.93, 0.17]	
GOAL SUBSCALE	Head Nods (#)	0.76	[−3.76, 3.30]		1.94	[0.10, 4.54]	⋆
	Head Shakes (#)	−0.21	[−4.02, 0.15]		0.70	[−2.35, −0.99]	
	Turn Length (s)	0.14	[0.02, 0.26]	⋆	0.01	[−0.20, 0.24]	
	Wait Time (s)	−0.06	[−0.12, 0.01]		−0.06	[−0.12, 0.01]	
	Listening Nods (%)	−0.33	[−2.41, 1.82]		2.67	[1.11, 4.18]	⋆⋆
	Listening Shakes (%)	−0.93	[−3.17, 1.30]		−0.54	[−1.87, 0.80]	
TASK SUBSCALE	Head Nods (#)	−1.39	[−2.60, 1.03]		1.76	[0.30, 4.53]	
	Head Shakes (#)	−4.14	[−6.64, −1.10]	⋆	−3.73	[−3.51, −1.02]	⋆⋆
	Turn Length (s)	0.15	[0.03, 0.27]	⋆	−0.01	[−0.24, 0.21]	
	Wait Time (s)	−0.05	[−0.12, 0.01]		−0.06	[−0.12, 0.01]	
	Listening Nods (%)	0.02	[−2.17, 2.21]		2.97	[1.36, 4.54]	⋆⋆
	Listening Shakes (%)	−1.17	[−3.47, 1.22]		−0.96	[−2.36, 0.37]	
BOND SUBSCALE	Head Nods (#)	−2.15	[−3.05, 1.10]		0.27	[−0.95, 2.78]	
	Head Shakes (#)	−1.08	[−4.32, 0.99]		−0.72	[−1.83, −1.73]	⋆⋆
	Turn Length (s)	−0.07	[−0.15, 0.02]		−0.21	[−0.35, −0.07]	⋆⋆
	Wait Time (s)	0.04	[0.01, 0.08]	⋆	0.05	[0.01, 0.09]	⋆
	Listening Nods (%)	−2.24	[−3.43, −1.06]	⋆⋆	0.81	[−0.37, 1.93]	
	Listening Shakes (%)	−1.92	[−3.38, −0.52]	⋆	−1.16	[−1.94, −0.36]	⋆

HDI = highest density interval, Sig. = significance, $\star$ $pd > 95\%$, $\star\star$ $pd > 99\%$.

Recognizing the multilevel structure of such interactions is critical, as these observations are not wholly independent, and such dependencies could bias parameter estimation or model building during training time [35]. We follow an established method for decomposing longitudinal data into three separate components [59].

- The *session-level* components capture how each session attended by a particular client compares to the other sessions attended by that client. Features at this level are those described in the previous section.
- The *client-level* components capture how each client compares to the other clients interacting with the same therapist. Features at this level aggregate all sessions attended by the same client.
- The *therapist-level* components capture whether each therapist's sessions tend to have higher or lower measures than the other therapists' sessions. Features at this level aggregate all sessions conducted by a given therapist, including all of their clients.

We approach our models from a Bayesian perspective. Bayesian methods provide a means of augmenting pre-existing domain knowledge (in the form of a prior distribution) with data-driven updates (in the form of observed data) to construct more robust models than either technique can achieve individually [52]. These analyses were performed using the `bambi` Python package [27], a high-level interface for the probabilistic programming framework PyMC3 [139]. Models were estimated through Markov chain Monte Carlo [119] via the No-U-Turn Sampler algorithm [71]. The model specification is presented in Figure 5.2. This equation describes the form of the model, in which each term includes an implied coefficient: these coefficients are parameters estimated during training time.

Interpretation of these models requires examining the resulting posterior distribution (the estimated distribution after observed-data updates) for each model parameter. To quantify these posterior distributions, we measure the posterior median and the 89% highest density interval (HDI). These two measures help us study the central tendency and spread, respectively, for each of the model parameters (also known as *effects*). The posterior median minimizes absolute error; the 89% HDI is common in Bayesian analysis, as it is more stable than the 95% HDI [99]. To understand the significance of the observed results, we also calculate the probability of direction (pd), a metric ranging between 50% and 100%, indicating the probability that a given parameter has the same sign as the posterior median [107]. We interpret pd values greater than 95% as 'significant' and pd values greater than 99% as 'highly significant'. Table 5.3 and Table 5.4 present the results obtained from the inferential analyses of client and therapist working alliance ratings, respectively. Note that each row of the table indicates a separate model, and that client behavior models were examined independently of therapist behavior models.

We observe that head gestures when listening are some of the client's most significant predictors of higher working alliance ratings. On the other hand, therapist behaviors had fewer significant associations with therapist ratings: the turn-taking features (turn length and wait time) were more strongly associated with working alliance ratings from the therapist. In both cases, the working alliance ratings were more associated with the behavior of the person providing the ratings than with the behavior of their partner.

Predictive Models

To evaluate the predictive power of head gestures and turn-taking behaviors in estimating working alliance ratings, we developed a set of models targeting each WAI subscale. Using the therapist-level, client-level, and session-level aggregated features (see Section 5.4 for details), we evaluated three predictive modeling procedures: support vector regression (SVR; [39]), Elastic Net [171], and random forests [19]. These algorithms were selected based on their ability to perform well on small datasets.

Model hyperparameters were automatically selected using a nested leave-one-therapist-out cross-validation approach to minimize train-test data contamination. For each therapist ($n = 11$), all sessions conducted by that therapist were designated as the test set, while all other sessions

TABLE 5.5
Performance metrics of predictive models: Root Mean Square Error, median and standard deviation.

	Client Ratings			
	Overall	Goal	Task	Bond
Baseline	0.82 (0.21)	0.86 (0.21)	0.94 (0.24)	0.85 (0.22)
SVR	**0.63 (0.22)**	0.69 (0.22)	0.74 (0.19)	**0.60 (0.25)**
Elastic Net	0.65 (0.23)	**0.66 (0.22)**	**0.68 (0.23)**	0.65 (0.23)
Random Forest	0.72 (0.18)	0.73 (0.20)	0.73 (0.18)	0.77 (0.19)

	Therapist Ratings			
	Overall	Goal	Task	Bond
Baseline	0.39 (0.31)	0.61 (0.36)	0.61 (0.42)	0.36 (0.29)
SVR	**0.31 (0.27)**	0.42 (0.30)	**0.50 (0.36)**	**0.30 (0.23)**
Elastic Net	0.37 (0.25)	**0.42 (0.31)**	0.53 (0.35)	0.32 (0.23)
Random Forest	0.38 (0.21)	0.43 (0.26)	0.58 (0.28)	0.36 (0.29)

were designated as the training set. Within the training set, validation for each fold was performed similarly: the sessions from one therapist were used for validation, while the remaining sessions were used for training. Features were recomputed for each training run to ensure that they do not rely on values from the test set. Prediction performance during validation and testing was measured using the root mean squared error (RMSE) metric. A benefit of RMSE over other similar metrics (e.g., the coefficient of determination R^2) is its definition in the same units as the output variable — in this case, working alliance ratings — and its stability in smaller datasets. Table 5.5 compares the test-set performance for each prediction model. For comparison, we also include a baseline model predicting the mean from the training set. All three models performed above the baseline model: the SVR and Elastic Net models tended to achieve the lowest RMSE.

Table 5.6
Performance metrics of ablation studies: Root Mean Square Error, median and standard deviation.

	Client Ratings			
	Overall	Goal	Task	Bond
Client Behavior	**0.64 (0.25)**	**0.69 (0.23)**	**0.71 (0.26)**	**0.70 (0.28)**
Therapist Behavior	0.95 (0.29)	1.02 (0.30)	1.04 (0.28)	1.03 (0.30)
Gesture Features	0.70 (0.23)	0.75 (0.25)	0.76 (0.26)	0.78 (0.30)
Turn-Taking Features	0.71 (0.25)	0.77 (0.23)	0.78 (0.33)	**0.73 (0.29)**
Gest. + Turn. Features	**0.67 (0.27)**	**0.74 (0.26)**	**0.73 (0.27)**	0.74 (0.25)

	Therapist Ratings			
	Overall	Goal	Task	Bond
Client Behavior	0.64 (0.31)	0.80 (0.44)	0.84 (0.46)	0.64 (0.33)
Therapist Behavior	**0.44 (0.34)**	**0.55 (0.40)**	**0.71 (0.45)**	**0.38 (0.30)**
Gesture Features	0.51 (0.33)	0.61 (0.40)	0.63 (0.47)	0.47 (0.36)
Turn-Taking Features	0.53 (0.36)	0.64 (0.37)	0.65 (0.47)	**0.44 (0.30)**
Gest. + Turn. Features	**0.49 (0.34)**	**0.59 (0.38)**	**0.60 (0.42)**	0.45 (0.31)

Ablation Studies

Following evaluation of the predictive models, we wanted to understand better the predictive value of including specific categories of features. We formulated two ablation studies to investigate: (1) behavior features from the therapist alone versus features from the client alone, and (2) head gesture features versus turn-taking features. Therapist-only features included features derived only from the therapist's behavior, and likewise for the client. Head gesture features are derived from head gestures alone (nods, shakes), independent of turn-taking behaviors (turn length, wait time). For comparison, we also present a third condition (referred to as 'Gest. + Turn. Features' in Table 5.6): the inclusion of both gestures and turn-taking features, but without the listening nods and listening shakes features that are derived from their combination. Table 5.6 compares the predictive performance of each of these models for both ablation studies.

5.5 Discussion

The present analysis sought to assess the value of computational nonverbal behavior analysis in estimating working alliance strength between therapists and clients. In this work, we investigated this proposition in three aspects: (1) a series of inferential analyses to identify general trends in behavior, (2) predictive model training to assess the ability to estimate working alliance ratings, and (3) a set of ablation studies to examine the significance of particular feature subsets. From these results, we identified some overall trends of note.

Participant ratings of the working alliance are largely uninformed by the behavior of the other participant. A consistent theme throughout these results is the suggestion that client behaviors do not offer much insight into therapist ratings, and similarly that therapist behaviors do not offer much insight into client ratings. This result corroborates prior work suggesting a frequent disconnect between therapist and client perception of the alliance [132, 134]. Also of note is the trend that client behaviors appear to hold more predictive power toward client ratings than therapist behaviors hold toward therapist ratings. This result is a valuable finding, as previous work has established that client ratings of the working alliance are the most reliable indicators of positive therapy outcomes, compared to therapist and observer ratings [74].

Head gestures tend to be more reflective of the task-oriented components of the working alliance, while turn-taking behaviors tend to be more reflective of the relationship-oriented component. As in many similar multimodal analyses [125, 137], our results identify trends in the salience of particular behavioral signals during the prediction of different outcome measures (Table 5.6). We note that turn-taking behaviors (speaking turn length and wait time) were primarily associated with the relationship-oriented component of the working alliance ratings — the bond subscale. On the other hand, head gestures (head nods and head shakes) were associated mainly with the working alliance ratings' task-oriented components — the goal and task subscales. There are similarities between these connections and those identified in studies of rapport, which recognize head gestures as more 'contentful' interaction signals [58, 156] and turn-taking patterns as more indicative of trust and respect [154]. We also note that the derived features (listening nods and listening shakes) were more predictive of the goal and task subscales than the bond subscale. This result could be attributed to prioritization among behavior signals, indicating that head gestures are a 'stronger' signal than turn-taking behaviors.

Beyond simply being *uninformed* by the partner's behavior, in certain cases, working alliance ratings are *misinformed* by the partner's behavior. A comparison of the behavior patterns associated with client ratings (Table 5.3) and therapist ratings (Table 5.4) reveals a few notable divergences. In one case, an increase in nodding on the part of the therapist was generally associated with the therapist providing *higher* ratings on the goal subscale. However, this same therapist behavior was associated with *lower* goal and task subscale ratings from the client. Similarly, when clients nodded more frequently when listening, clients tended to provide *higher* ratings on all subscales, but therapists tended to provide *lower* ratings. These results seem to be consistent with other research, which found that therapy participants often 'misread' the behavioral cues of their partner [68, 132]. Despite this, our computational models were capable of predicting both participants' self-reported ratings of working alliance with moderate accuracy (Table 5.5).

5.6 Conclusion

The *working alliance* is a critical piece of the interaction between client and therapist that captures the collaborative aspect of the therapeutic relationship. A strong working alliance has been associated with several measures of positive therapy outcomes, but is often difficult to identify, as its definition relies on the subjective perspectives of both the client and the therapist. Further complexity is introduced by participant unawareness and misunderstanding of partner behaviors during the interaction.

Together, these results provide important insights into the challenges facing assessment of the working alliance during therapy and how computational behavior analysis holds promise for addressing these obstacles. Further research might explore the role of personal characteristics (e.g., personality, sociodemographic) or the client's psychiatric concerns (e.g., anxiety, depression), as the influence of these factors on nonverbal behavior is well-established [33, 120]. Although the sample of participants in this work is diverse and representative of the population in one community, generalizations to broader populations dissimilar to this one will require additional data collection and repeat analysis. A natural progression of this work would also include other behavioral signals, such as facial expressions or acoustic patterns in speech. The understanding gained through this line of research can foster the development of systems providing early detection of a weak working alliance, allowing for preemptive intervention and reduction in the barriers facing clients seeking treatment.

Acknowledgments

This material is based upon work partially supported by The Center for Machine Learning and Health at Carnegie Mellon University, National Institutes of Health awards R01MH125740, R01MH096951, UL1TR001857, and U01MH116925, and National Science Foundation awards #1722822 and #1750439. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred.

Chapter 6

Convergent Behavior Dynamics

As discussed in prior chapters, a strong working alliance has been extensively linked to many positive therapeutic outcomes. We can learn much about the strength of this alliance through participants' behavior, even at the fundamental level of facilitative behavior, but we now narrow our focus to convergent behavior. Although many aspects of therapy sessions are worth thorough examination, language use is of particular interest given its recognized relationship to similar dyadic concepts such as rapport, cooperation, and affiliation. Specifically, in this chapter we study language entrainment, which measures how much the therapist and client adapt toward each other's use of language over time. We explore these questions through the use of structural equation modeling (SEM) techniques, which allow for both multilevel and temporal modeling of the relationship between the quality of the therapist-client working alliance and the participants' language entrainment.

The work described in this chapter first appeared in the following publication:

> Alexandria K. Vail, Jeffrey Girard, Lauren M. Bylsma, Jeffrey F. Cohn, Jay Fournier, Holly Swartz, and Louis-Philippe Morency. Toward Causal Understanding of Therapist-Client Relationships: A Study of Language Modality and Social Entrainment. *Proceedings of the Twenty-Fourth International Conference on Multimodal Interaction (ICMI 2022)*, Bangalore, India, 2022.
>
> https://doi.org/10.1145/3536221.3556616

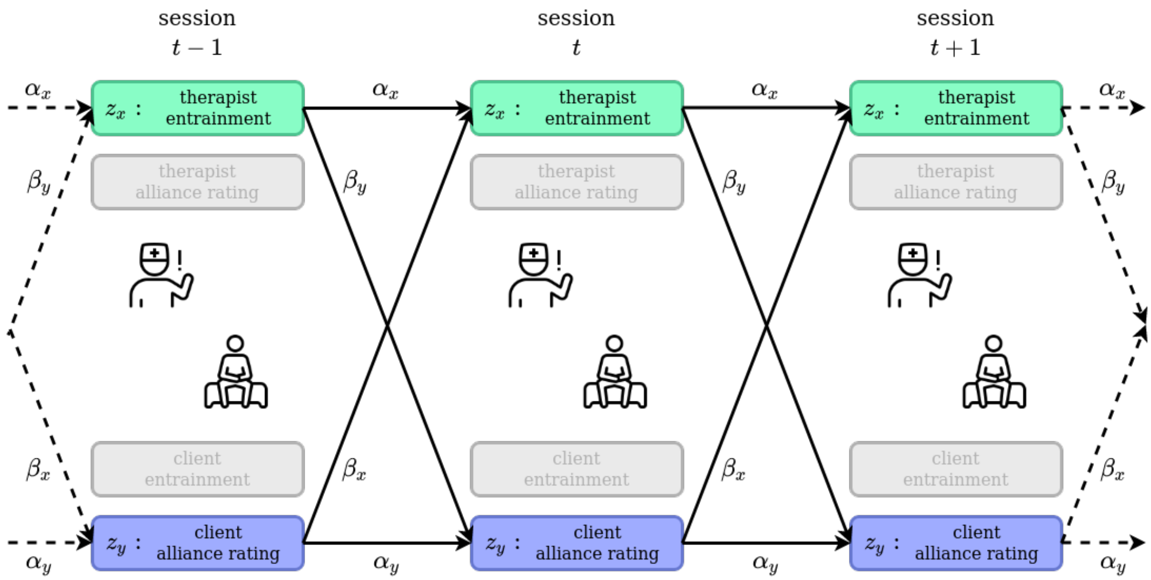


FIGURE 6.1
Example illustration of the structure of the therapist-entrainment/client-alliance analysis. During each session, we calculated an entrainment score (style and content) based on each participant's behavior, and after each session, each participant provided a rating of the working alliance (goal, task, and bond subscales). Edge labels (α_x, α_y, β_x, β_y) and node labels (z_x, z_y) correspond to the parameters introduced in Section 6.5 and Figure 6.2. A similar structure was mirrored for the therapist-entrainment/therapist-alliance, client-entrainment/client-alliance, and client-entrainment/therapist-alliance analyses.

6.1 Overview

Evidence suggests that the quality of the relationship between a client and their therapist is one of the most critical factors in determining treatment success [74, 110]. Concretely, much of the current psychological literature on the client-therapist relationship focuses on what is known as the *working alliance* [73]. This concept aims to capture the *collaborative* aspect of the therapist-client relationship. The working alliance is generally considered consisting of three components: agreement on the overall goal of the treatment, agreement on the tasks required to reach that goal, and the feeling of emotional bond between the participants. A positive working alliance between client and therapist plays a crucial role in fostering numerous positive therapeutic outcomes, including reduction of the client's symptoms and concerns [51, 73, 74], reduced drug abuse and recidivism [106] and improved medication compliance [47]. Of particular note is the recognized relationship between the quality of the working alliance and client dropout [47, 95, 141].

Proactive detection of a poor working alliance is especially valuable in this case: by the time a client has decided to quit therapy, the time for potential intervention has already passed. Understanding the complexity of the therapist-client relationship is crucial for informed treatment decision-making.

While working alliance and therapist-client relationships are decidedly multimodal concepts, the modality of language use is of particular interest given its importance in understanding similar forms of dyadic interaction [25, 78, 79, 127]. Relatively few studies have examined approaches for evaluating the working alliance beyond explicit questionnaires. More importantly, no previous work has studied the causal direction of the relationship between language and working alliance. Studying this relationship through the lens of causality allows us to go beyond correlation and address a broader range of research questions, such as the ones we focus on in this paper: does language behavior affect how the working alliance is perceived, or does working alliance perception affect how language is used?

This paper builds upon structural equation modeling (SEM) techniques to investigate the causal relationship between language use and working alliance. In particular, we introduce a specific method of structuring this model that allows us to study both relationships over time (temporal modeling) and patterns within individuals (multilevel modeling). Given the highly social nature of therapy sessions, we focus on *entrainment* in participant language. Linguistic entrainment is the process of multiple interlocutors (in our case, a client and their therapist) converging toward each other's use of language. We study linguistic entrainment in terms of both stylistic properties and content properties.

The structure of this paper consists of eight sections. In the next section, we review previous literature on behavior detection, working alliance, and linguistic entrainment (Section 6.2), and the following section provides a brief overview of the dataset used in this analysis (Section 6.3). Section 6.4 describes the definition and computation of our features and labels. We then devote Section 6.5 to an in-depth explanation of the SEM-based model we use in our analysis. The primary contributions of the paper lie in the next two sections: Section 6.6 evaluates the performance of this model in relation to other commonly used modeling techniques, while Section 6.7 interprets the model's conclusions and discusses the implications of these results for behavioral research. The final section summarizes the main findings of this work and identifies areas for further research.

6.2 Related Work

Interpersonal coordination is a behavioral phenomenon where multiple interacting individuals adapt their behavior together over time [157], which can take many forms [21]. Previous research has demonstrated that humans will coordinate their movements [3], voices [77, 130], and other communicative behaviors [105] to match each other during an interaction. A considerable amount of work has been published on the relationship between prosocial outcomes and behavioral coordination: increased interpersonal coordination during interaction leads to improved cooperation and collaboration [165], as well as higher self-reported ratings of rapport [133] and affiliation [76].

Despite this growing body of literature, relatively little work has focused on the role of interpersonal coordination in psychotherapy (cf. [1, 2, 128, 166]). Within this area of research, most prior work on therapy sessions has focused primarily on movement synchrony [96, 131]. In this analysis, we draw from related literature in social psychology that examines the role of *language entrainment* as a predictor of prosocial outcomes. Significant evidence exists to suggest that increased language style matching, in particular, leads to higher ratings of social intimacy, stability, and involvement [78, 79]. Language entrainment has also been linked to increased perception of support [127] and the general positivity of the interaction in question [25]. In long-term social relationships, language entrainment has also been shown to predict child attachment security significantly in parental relationships [17]. Inspired by this adjacent literature, this analysis explores whether language entrainment can also serve as a reliable and objective indicator of the quality of the therapeutic working alliance.

6.3 Dataset

Audiovisual recordings were collected from 266 therapy sessions between 39 unique clients and 11 unique therapists. Each therapist met with an average of 3.6 unique clients, and each client participated in an average of 6.8 sessions lasting between 40 and 60 minutes each (average 50.3 minutes).

Potential participants were recruited from a research registry, printed material advertising the study, and word-of-mouth. To be included in the study, participants had to be adults aged 18–65, meet DSM-5 criteria for a major depressive disorder[1], currently experience at least moderate depressive symptoms (as measured by a Hamilton Rating Scale for Depression score ≥ 14; [61]), and be willing and able to provide informed consent. Individuals with a comorbid psychotic disorder, active suicidal or homicidal ideation, chronic depression, or current substance or alcohol abuse were excluded from the study. If an individual was suspected of experiencing psychosis or active suicidal ideation with intent or plan to harm themselves, the investigator terminated the screening interview and ensured that the individual obtained appropriate care, including but not limited to a referral to the psychiatric emergency room.

Included clients ranged from 22 to 65 years of age; 77% identified as female, and 62% identified as White. Clients were randomly assigned to an eight-session brief course of one of two empirically supported psychotherapy conditions: cognitive behavioral therapy (CBT; 21 clients, 6 therapists) or interpersonal psychotherapy (IPT; 18 clients, 5 therapists).[2]

[1]The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; [5]) is a taxonomy of psychiatric disorders published by the American Psychiatric Association. This manual serves as the primary diagnostic tool for psychiatric diagnosis and treatment in the United States.

[2]There were no statistically significant differences in working alliance ratings observed between the two treatment conditions.

TABLE 6.1
Sample items from both therapist and client versions of the Working Alliance Inventory.

Goal Subscale	Task Subscale	Bond Subscale
[Therapist] and I collaborate on setting goals for my therapy. [Therapist] and I have established a good understanding of the kind of changes that would be good for me. We are working towards mutually agreed upon goals. [Client] and I have a common perception of his/her goals.	What I am doing in therapy gives me new ways of looking at my problem. [Therapist] and I agree on what is important for me to work on. [Client] and I agree about the steps to be taken to improve his/her situation. [Client] and I both feel confident about the usefulness of our current activity in therapy.	I believe [Therapist] likes me. I feel that [Therapist] appreciates me. I feel [Therapist] cares about me even when I do things that he/she does not approve of. I appreciate [Client] as a person. [Client] and I respect each other.

6.4 Language Entrainment and Working Alliance

Ratings of Working Alliance

Following the conclusion of each therapy session, both therapist and client participants completed the therapist and client versions of the revised short-form Working Alliance Inventory (WAI; [65]), a widely used measure of alliance in therapy. The WAI consists of three subscales capturing three aspects of a working alliance:

- the *goal* subscale, which assesses the individual's belief that participants agree on the overall objectives of the treatment;
- the *task* subscale, which assesses the individual's belief that participants agree on the steps required to reach the goals mentioned above; and
- the *bond* subscale, which assesses the individual's respect and trust for the other participant in an emotional sense.

Each subscale consists of statements that the individual rates on a five-point Likert-type scale ranging from 'seldom true' to 'always true'; the inventory contains 12 items for the client and 10 items for the therapist. Representative items for each subscale are presented in Table 6.1.

Language Style and Content Metrics

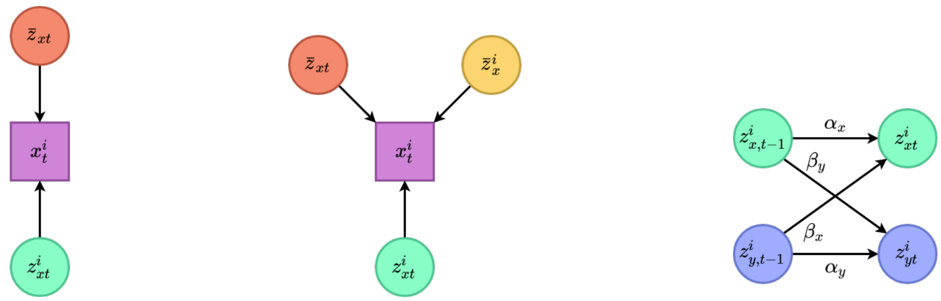
Language entrainment is the process of multiple interlocutors adapting toward each other's use of language throughout an interaction. Although there exist many operational definitions to measure this construct, we leverage and expand upon a metric called *reciprocal linguistic style matching* (rLSM; [118]). The original definition of rLSM utilizes the Linguistic Inquiry and Word Count dictionary (LIWC; [123]), a well-validated and established lexicon that organizes approximately 6,400 English words into several semantically or functionally similar categories. In particular, we use LIWC "function word" categories: pronouns, articles, prepositions, auxiliary verbs, adverbs, conjunctions, and negations. Function words are useful to examine because they are independent of context, and their use is often less conscious. The benefit of rLSM over other metrics is the reciprocal component, which aims to measure how much the interlocutors change toward each other over time, rather than how much they may coincidentally speak with a similar style.

The rLSM score is initially calculated at the utterance level. Consider a therapist's response (T) to an utterance by the client (C): we aim to calculate the rLSM metric for the therapist's utterance. Since utterance T is a response to utterance C, we define rLSM$_T$ as follows:

$$\text{rLSM}_T(S) = 1 - \frac{|S_C - S_T|}{S_C + S_T + 0.0001} \tag{6.1}$$

Here S represents any LIWC category score (e.g., negations) computed for client and therapist utterances (S_C and S_T, respectively). The total rLSM score for a statement is the average score of all function word categories. This score is calculated for each utterance during the session, and all utterance scores from each session are then averaged to determine each participant's session-level rLSM score.

We also propose an extension to rLSM, which studies the "content" component of language for contrast against the "style" component of language. We approximate this content component



(A) $x_t^i = \bar{z}_{xt} + z_{xt}^i$
The CLPM measurement component for x_t^i. A similar structure is mirrored for y_t^i.

(B) $x_t^i = \bar{z}_{xt} + \bar{z}_x^i + z_{xt}^i$
The RI-CLPM measurement component for x_t^i. A similar structure is mirrored for y_t^i.

(C) $z_{xt}^i = \boldsymbol{\alpha_x} z_{x,t-1}^i + \boldsymbol{\beta_x} z_{y,t-1}^i + e_{xt}^i$
The temporal deviation component common to both CLPM and RI-CLPM.

FIGURE 6.2
Breakdown of the essential components of the CLPM and RI-CLPM techniques for a given session t.

using the following LIWC categories: affective words; social words (family, friends); words relating to cognitive, perceptual, and biological processes (seeing, feeling; health); and words relating to motivation/drives and personal concerns (risk, reward; leisure, religion). We term this new metric rLCM — reciprocal linguistic content matching.

6.5 Causal Model Introduction

Our model was designed with several desired principles in mind. First, we needed our model to be *interpretable*. Although the nature of the present analysis is primarily exploratory, we begin with some degree of expert domain knowledge and initial hypotheses as to the underlying structure of the data. For example, we expect that some individuals will adapt their language more than others ([126]; requiring multilevel modeling) and that working alliance ratings tend to increase over time ([67]; requiring temporal modeling).

To leverage these existing theoretical foundations, we turn to structural equation modeling (SEM) techniques [40]. SEM is a set of multivariate techniques that are generally confirmatory in nature, aiming to test whether a particular model structure fits a given dataset [102]. Unlike traditional machine learning models, SEM primarily leverages not the raw data provided to it but the covariance matrix: the goal is to minimize the distance between the observed and model-implied matrices. SEM also offers some advantages in our particular case. Given the additional overhead and sensitivity required to collect rich healthcare data, such as our own dataset introduced in Section 6.3, these healthcare datasets are often of a smaller size than those in other domains of multimodal research. In reducing the number of estimated variables by imposing a theoretical structure, SEM also allows us to explicitly account for the variance due to the inevitable measurement error present in psychological data. These features allow us to attain greater statistical power with fewer samples.

Given that we pursue the use of SEM for our analysis, we must design the underlying structure fundamental to these techniques. We intend to evaluate the relationship between the participants' perception of the working alliance and the adaptation of their language use toward their conversational partner, and in particular, the direction of this relationship: we are interested in *causality* in the data. Given the longitudinal nature of our dataset, the standard practice is to turn to the family of cross-lagged panel models (CLPMs; [24]). Finally, we must consider that our observations follow the same individuals over time, so we must also include consideration for participant-level patterns. We expect that participants will differ in their personal tendencies simply due to personality or other individual characteristics; for example, some people may be more inclined to adapt their language than others. This final consideration leads us to a modern hierarchical extension of the CLPM: random intercept cross-lagged panel modeling (RI-CLPM; [60]). The following subsections describe the intuitions and definitions of the classic CLPM as well as the improvements and benefits introduced by the RI-CLPM extension.

Cross-Lagged Panel Modeling

Cross-lagged panel models (CLPM; [24]) involve the evaluation of the effect of two (or more) variables on each other over time. Consider x and y as two distinct variables (e.g., entrainment score and working alliance rating) from participant i measured over multiple time points (sessions) t. We aim to evaluate the relationship between x and y. The first important intuition behind CLPM techniques is the idea that a measured variable x (or y) is composed of a mean and a variation from that mean. This intuition can be formulated as follows (see Figure 6.2a for an illustrated breakdown):

$$x_t^i = \bar{z}_{xt} + z_{xt}^i; \qquad y_t^i = \bar{z}_{yt} + z_{yt}^i; \tag{6.2}$$

where z_{xt}^i and z_{yt}^i represent the participant's temporal deviations from the temporal group means $\bar{z}_{xt}$ and $\bar{z}_{yt}$, respectively.

The second important intuition behind this model is that these temporal deviations z_{xt}^i are affected not only by previous temporal instances of itself, but also previous temporal instances of the other variable, z_{yt}^i; the same concept applies symmetrically for temporal variations of the other measured variable. This intuition is where the "cross-lagged" term in this approach originates. We can formally model these temporal deviations on the latent variables z_{xt}^i and z_{yt}^i as follows (Figure 6.2c):

$$z_{xt}^i = \alpha_x z_{x,t-1}^i + \beta_x z_{y,t-1}^i + e_{xt}^i, \tag{6.3}$$

$$z_{yt}^i = \alpha_y z_{y,t-1}^i + \beta_y z_{x,t-1}^i + e_{yt}^i. \tag{6.4}$$

The parameters α_x and α_y are autoregressive parameters that account for the temporal stability of these constructs: that is, the closer these parameters are to one, the more stable the rank order of individuals across time points. The parameters e_{xt}^i and e_{yt}^i represent residuals. The cross-lagged parameters β_x and β_y are fundamental to this family of models — by comparing the crossed effects of x on y (and vice versa), we can identify evidence to suggest the causal predominance of one direction over the other.

Random Intercept Cross-Lagged Panel Modeling

Following Hamaker et al. [60], we use an extension of CLPM that allows each participant to have their own individual variation compared to the group-level means $\bar{z}_{xt}$ and $\bar{z}_{yt}$. This model is named the random intercept cross-lagged panel model (RI-CLPM). RI-CLPM is a multilevel model where observations are nested within individuals. This model includes a random intercept that allows it to account not only for temporal stability, but also trait-level stability. With this in mind, Equation 6.2 can be rewritten as follows (see Figure 6.2b for an illustrated breakdown):

$$x_t^i = \bar{z}_{xt} + \bar{z}_x^i + z_{xt}^i, \qquad y_t^i = \bar{z}_{yt} + \bar{z}_y^i + z_{yt}^i, \tag{6.5}$$

where the added parameters $\bar{z}_x^i$ and $\bar{z}_y^i$ represent the participant's individual trait-level deviations from the existing temporal group means. In this case, the parameters z_{xt}^i and z_{yt}^i now represent the participant's temporal deviations from their personalized expected scores (i.e., $\bar{z}_{xt} + \bar{z}_x^i$ and $\bar{z}_{yt} + \bar{z}_y^i$) rather than deviation from the temporal group mean (i.e., $\bar{z}_{xt}$ and $\bar{z}_{yt}$). We can now express these deviations as follows (Figure 6.2c):

$$z_{xt}^i = \alpha_x z_{x,t-1}^i + \beta_x z_{y,t-1}^i + e_{xt}^i, \tag{6.6}$$

$$z_{yt}^i = \alpha_y z_{y,t-1}^i + \beta_y z_{x,t-1}^i + e_{yt}^i. \tag{6.7}$$

The autoregressive parameters α_x and α_y no longer represent merely the rank order of participants over time, but the degree of the within-person carry-over effect. For example, if this parameter is positive, it suggests that if a participant scored higher than their expected score at time point t, they are likely to also score higher than their expected score at time point $t + 1$.

One advantage of using the RI-CLPM over the CLPM is that it is effectively a generalization of the CLPM: if the additional elements are determined to be unnecessary, the additions tend toward zero and the model essentially 'collapses' to the base CLPM. Furthermore, in the case of the present analysis, we can reasonably assume that the effect the variables x and y have on

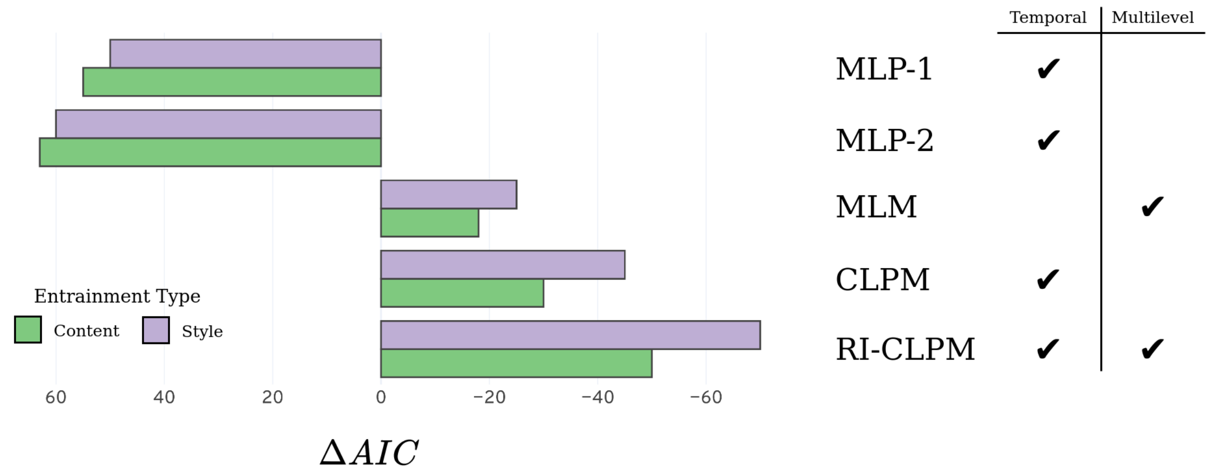


FIGURE 6.3
Comparative performance of baseline models relative to the linear model (LM). Note that AIC is a relative metric, and has no meaning in absolute terms: there are no "good" or "bad" AIC scores, only "better" or "worse" than another. Therefore, lower ΔAIC scores (further right in the chart) are better.

each other over time remains stable: our observed time points are roughly evenly spaced, and we do not perform any midpoint 'intervention' that would suggest that any particular interval differs from the other intervals. As a result, we tie parameters (i.e., α and β) across time points, providing us with many more degrees of freedom in our model and parameters that are more straightforward to interpret.

6.6 Prediction Experiment

Our first set of experiments compares RI-CLPM performance against other commonly used models, such as neural networks. As a reminder, an important goal when designing our model based on RI-CLPM was to leverage domain knowledge to reduce complexity and hopefully improve performance. Our model integrates inductive biases (domain knowledge) for both the temporal and the multilevel aspects of the data.

Baseline Models

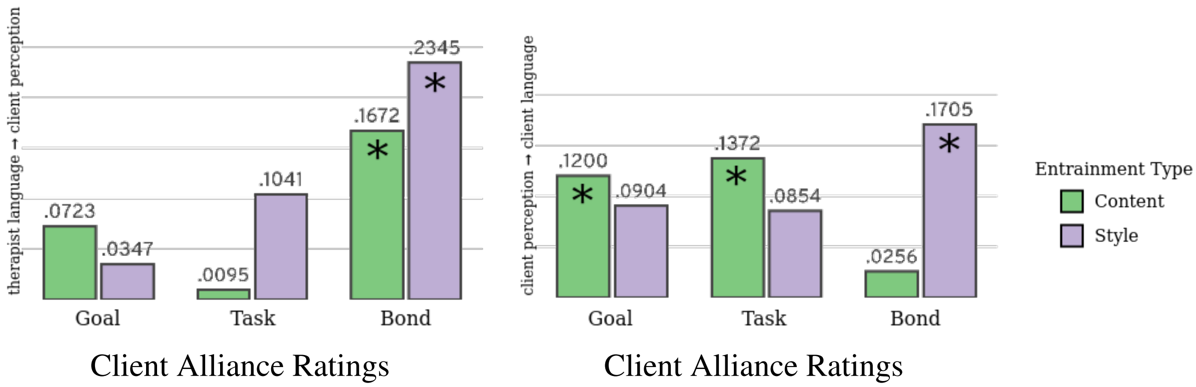
We compare our model with several commonly used machine learning models. We begin with neural networks: given the small number of data samples, we constrained ourselves to multi-layer perceptrons. We included two variants with one or two hidden layers (MLP-1 and MLP-2, respectively). To study the relative importance of the two inductive biases we included in our model, we included as baselines a multilevel linear model (MLM) and the standard CLPM. The comparison with the CLPM allows us to evaluate the importance of including the random intercept component. All models were compared in terms of the performance of a simple linear model (LM), which can also perform effectively with small datasets.

Prediction Metrics

One of the challenges when evaluating all of these models is selecting a metric that will be fair and comparable across models. Although many commonly used models (such as MLP models) are generally trained and evaluated in terms of their predictive performance (e.g., accuracy), SEM-based models have no directly corresponding notion of "prediction". Therefore, for this comparison, we rely on a metric revolving around *model fit*: Akaike's information criterion (AIC; [36]), which evaluates how well a given model's implied structure matches a given dataset. Rather than providing an "absolute" score, it instead offers evidence for the preference of one model over a set of others: in other words, there are no "good" or "bad" AIC scores, only scores that are "better" or "worse" than that of another model. This metric can be expressed as follows:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}), \tag{6.8}$$

where k is the number of estimated parameters in the model and $\hat{L}$ is the maximum value of its likelihood function.



(A) Estimated cross-lagged parameters (β) evaluating the effect of the **therapist's entrainment** behavior on the **client's alliance** ratings. *How does the therapist's language affect the client's perception of the working alliance?*

(B) Estimated cross-lagged parameters (β) evaluating the effect of the **client's alliance** ratings on the **client's entrainment** behavior. *How does the client's perception of the working alliance affect their language?*

FIGURE 6.4
Highlighted results from the language analysis described in Section 6.7. Asterisks (*) indicate parameters statistically significantly different from zero ($p < 0.05$).

Results and Discussion

Figure 6.3 presents an overview of the performance of all models. Given that AIC is a relative metric, all scores are interpreted in terms of difference from the baseline model, the linear model. From this figure, it becomes apparent that the general pattern of better performance is achieved with the addition of temporal and multilevel elements — for such a relatively small but rich dataset, the importance of leveraging expert knowledge of both domain and dataset structure is evident.

6.7 Language Analysis

Our second set of experiments analyses the learned cross-lagged parameters (β_x and β_y) of the RI-CLPM model. Our goal is to study the relative effects of a participant's perception of the working alliance on their linguistic entrainment behavior. One benefit of our approach is the ability to distinguish directional effects — that is, whether working alliance perception affects linguistic entrainment, or if linguistic entrainment affects working alliance perception.

Working alliance ratings were collected from both client and therapist at the end of each session: these working alliance ratings are divided into agreement on goals, agreement on tasks, and agreement on bond. We also calculated both a stylistic entrainment score and a content entrainment score for each participant during each session (see Section 6.4 for more details on the calculation of these metrics). We fit an RI-CLPM to each combination of language behavior and working alliance ratings. From these fitted models, we primarily examine the cross-lagged parameters that estimate the relationship between the two measured variables: see Section 6.5 for more details on the model.

Results

Highlighted results are presented in Figure 6.4. Numerous significant effects can be observed from these results. In general, the client's perception of the working alliance results in an increase in their style and content entrainment (Figure 6.4b). In particular, the client's perception of bond results in an increase in their stylistic entrainment, while their perception of the goal and task aspects of the working alliance result in an increase in their content entrainment.

From Figure 6.4a, we can see that the client's perception of bond is significantly influenced by both content and stylistic linguistic entrainment on the part of the therapist. On the other hand, the therapist's perception of the working alliance appears less impacted by linguistic entrainment: the only significant association observed is that an increase in the client's content matching results in an increase in the therapist's perception of task agreement ($\beta = 0.1179$).

Discussion

The present analysis was designed to determine the effect of language entrainment during therapy sessions on the participants' perception of the working alliance, and vice versa. The results provide preliminary evidence to suggest a bidirectional but asymmetric relationship between these two constructs.

Stylistic entrainment is generally associated with perception of bond, while content matching is generally associated with perception of task and goal. By examining working alliance ratings at this granular level, we can observe that stylistic entrainment seems associated mainly with the perception of bond. In contrast, content matching appears primarily associated with the perception of task and goal.

Therapy clients express their perception of the working alliance through linguistic entrainment. Perhaps the most compelling finding to emerge from this analysis is the suggestion that the client appears to demonstrate their current perception of the working alliance through their linguistic entrainment behavior, as seen in Figure 6.4b.

Therapist linguistic entrainment has a notable impact on the client's perception of the working alliance bond. Finally, a third notable takeaway is that the therapist's language entrainment behavior seems to have a substantial impact on the client's perception of the working alliance, and particularly, their impression of the bond (Figure 6.4a).

These results, particularly those discussed in the latter two points, also demonstrate the importance of considering causality when investigating these relationships. A model that explores only correlation, as most commonly used models, would be unable to ascertain, for example, whether a client's linguistic entrainment affects their perception of the alliance or if their perception affects their entrainment.

6.8 Conclusion

The working alliance is a multifaceted concept that captures the collaborative aspect of the relationship between a therapist and their client. We use structural equation modeling (SEM) techniques to study the causal relationship between working alliance and language entrainment behaviors. We demonstrate that this kind of modeling can achieve excellent performance compared to other standard machine learning models, with the added benefit of interpretability and causal analysis. Interpretation of the model reveals valuable insights into the dyadic interaction between therapist and client during therapy. In general, the language entrainment of the therapist can have an impact on the client's perception of the alliance, and the client's perception of the alliance is often reflected in their own language use.

Future work includes exploring the relationship between working alliance and other social behaviors, such as gestures, prosody, and facial expression; bringing these modalities together into a multimodal approach is also of great interest. Examining the relationship between these behaviors and the alliance throughout a single interaction at a more granular level may also reveal exciting relationships. Such findings could eventually be implemented in the form of a computer-mediated feedback system, aiding the therapist in recognizing the deterioration of the working alliance in the moment and allowing for more immediate intervention to address client concerns. Multimodal behavior analysis in therapy has many promising future paths: the ensuing enhancement of therapeutic interaction will help ensure that more people seeking therapy receive the treatment they need.

Acknowledgments

This material is based upon work partially supported by U.S. National Science Foundation awards #1722822 and #1750439, and U.S. National Institute of Health awards R01MH125740, R01MH096951, and U01MH116925. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors, and no official endorsement should be inferred.

Chapter 7

Proposed Work

Divergent Behavior Dynamics

The dynamics of collaborative social interaction can be illustrated at a high level as multiple timelines developing in parallel, moving toward the same general objective. In Chapter 5, we discussed the modeling of facilitative behavior, through which the participants maintain the flow of conversation, such as through turn-taking patterns. In Chapter 6, we discussed the modeling of convergent behavior, through which participants (consciously or subconsciously) coordinate their behavior, such as through linguistic entrainment. The final remaining dynamic we aim to explore in this thesis is divergent behavior. Divergent behavior occurs with an increase in contrast between participants' behavior. These behaviors could appear as overt as open conflict, or could manifest more subtly, such as through withdrawal behaviors.

We propose an expansion of the work on the data introduced in Chapter 5 and Chapter 6 (we refer the reader to these chapters for a detailed description of the dataset). This dataset consists of a set of recorded therapy sessions with adult individuals recruited from the local community. Individuals participating in the study met the standard criteria for a major depressive disorder [5] at the beginning of treatment and were experiencing at least moderate symptoms of depression at each subsequent session [61]. The therapy sessions adhered to a brief eight-week course of treatment with a trained therapist, under no specific constraints beyond the overarching therapeutic approach (cognitive behavioral therapy [12] or interpersonal psychotherapy [108]). Sessions were recorded using audio and video, and a transcript was generated for each session. This data allows us to investigate several modalities, such as facial expression, gestures, language use, postural changes, and prosodic patterns. After each session, the client and therapist completed the Working Alliance Inventory [65], a widely used measure of alliance in therapy. We will use this data to explore divergent moments of interest as indicators of the strength of the working alliance between the client and therapist. We express this overarching research question:

How can we identify and computationally represent behavioral markers of divergence that relate to the perception of working alliance from both therapist and client?

7.1 Computational Representation of Divergence Behaviors

Our first major research task is to computationally represent behavioral patterns of divergence. Behavioral divergence reflects an increase in contrast between participants' behavior. This contrast between behaviors can be seen through either nonverbal behaviors, verbal behaviors, or both.

Dyadic divergence. Our primary research goal is to study dyadic divergence, where the behavioral measures are derived from the behavior of both participants. One way to describe this concept would be to define dyadic divergence as changes in the difference between the client's and therapist's behavior (e.g., when the client is behaving more differently than the therapist than they were previously). We also keep the door open to a secondary and complementary option, which is to study monadic divergence – when a participant expresses a behavior pattern that is different from their own past behaviors.

We intend to prioritize nonverbal behavioral markers of divergence, to complement Chapter 6, which studied verbal behavior (i.e., linguistic entrainment) for the corresponding concept of convergence. However, we are also open to the inclusion of verbal behaviors as part of our divergence analysis, with the specific goal of contextualizing nonverbal behavior within verbal context. To restate in more concise terms, the primary goal of the proposed work described in this chapter is to identify and computationally represent dyadic and monadic behavioral markers of divergent moments as they relate to perception of working alliance.

Working alliance constructs. Behavioral markers are designed to indicate behaviors that are predictive of a broader, more abstract construct. In our case, we are studying markers related to participant perception of the working alliance. Following our work from Chapter 5 and Chapter 6, we will focus on the same three sub-constructs of working alliance: goal, task, and bond. We intend to include both therapist and client perspectives of working alliance.

Dyadic behavior metrics. A dyadic behavior metric evaluates observed behavior of both participants over a short period of time (e.g., over three speaking turns). Our first behavior of interest is head gestures. Our results from Chapter 5 demonstrated that head nods and head shakes can be used to predict working alliance sub-constructs. Another interesting takeaway was that there was a difference in how they could be used to predict client versus therapist ratings. We intend to represent the dyadic patterns related to head gestures using a methodology similar to that used in Chapter 6, particularly by extending reciprocal linguistic style matching (rLSM) to nonverbal behaviors (e.g., head gestures). rLSM features are computed over a sliding time window spanning three speaking turns: (1) a speaking turn of the participant of interest, (2) a speaking turn of the other participant in the dyad, and (3) a second speaking turn for the participant of interest. This window allows rLSM to assess changes in the participant's language use toward (or away from) the other participant.

One possible approach would be to replace verbal features of a speaking turn with nonverbal behaviors during a speaking turn, similar to how we used lexicon-based features of linguistic style and content in Chapter 6. Our initial study will use the head gesture features from Chapter 5, but we also intend to extend this analysis to eye gaze patterns, as in Chapter 3. We plan to characterize eye gaze patterns using the same methodology used in Chapter 3, in which we automatically describe gaze aversion patterns of the client. While we will initially focus our analyses on nonverbal behaviors, we will also keep as an option inclusion of verbal behaviors in the analysis. Since we have already processed our dataset (details in Chapter 6) to include linguistic features using a lexicon-based approach (LIWC), we have an existing method of representing stylistic and content-oriented linguistic features.

Session-level features. Our psychotherapy dataset contains multiple therapy sessions per client, usually eight. The working alliance ratings are defined at the session level, after each session, from both the client and therapist. Our objective is to identify behavioral markers that summarize behavior within each session, and study their link to working alliance ratings of goal, task, and bond. To summarize the session-level patterns, we intend to pursue three strategies:

- **Averaged session-level features.** One strategy is to take a more holistic approach, like the one we used in Chapter 6, by simply averaging the behavior measures (nonverbal rLSM metrics) over the whole session. Recall that rLSM metrics are computed over three speaking turns. The simplest approach is to slide the window of speaking turns through the entire session, keeping the focus on one specific participant (client or therapist), and then average the resulting metrics. This gives us a measure of the average trend of convergence or divergence.

- **Time-slice patterns.** A more fine-grained approach would be to summarize the dyadic behavior metrics (e.g., rLSM metrics) over a short time window. As a starting point, we intend to consider five-minute time slices. The sessions usually last for about 60 minutes, which allows for approximately 12 slices per session. Within each slice, we intend to summarize dyadic behavior metrics, and then characterize the temporal patterns (e.g., linear and non-linear slope) within the session.

- **Moments of change**. We intend to explore whether our computational metrics can identify specific moments of change: moments that differ from the more "typical" behaviors (or difference in behavior). We intend to examine two patterns of change: rapid changes and permanent changes. Rapid changes are characterized by a strong change in behavior that eventually returns to the behavioral pattern previously observed. Permanent changes also include a strong change in behavior, but in this case, the change persists until the end of the session. We plan to begin by fitting a linear regression model on the initial time slices of the session (e.g., the first five time slices), and study the difference from this linear model and the remaining time slices. However, we want to emphasize the exploratory nature of this third level of analysis.

7.2 Analysis of Divergence Markers and Working Alliance

The second major research task will be to study the relationship between the dyadic behavioral markers computed in the first task and the perception of the working alliance. We plan to build upon the analytical methodology we developed in Chapter 6, through which we studied the link between dyadic behavior and working alliance using structural equation models (SEM). SEM is a set of multivariate techniques that are generally confirmatory in nature, aiming to test whether a particular model structure fits a given dataset [102]. There are some advantages to confronting this topic with SEM. Since the collection of such rich healthcare data tends to require additional overhead and sensitivity, the size of healthcare datasets is often smaller than in other domains of multimodal research. As a result, many commonly used machine learning models are not suited for this use case, having been optimized for processing tremendous amounts of data. Since SEM-based models involve the use of domain knowledge to impose a theoretical structure, it allows us to attain greater statistical power with fewer samples.

We are intending to use a particular kind of SEM model called the cross-lagged panel model (CLPM), a variant of which is used in the convergent behavior analysis (RI-CLPM; more details regarding this specific SEM model can be found in Chapter 6). One advantage of our proposed methodology is that it gives us an opportunity to study some causal aspects of the problem: how well divergence markers of the previous session can predict the working alliance ratings of the current sessions, but also how working alliance ratings can predict future divergent behaviors.

The RI-CLPM allows us to study this temporal and causal relationship while also taking into consideration the multi-level aspect of the study (i.e., the same therapist will meet with multiple unique clients, and the same client will attend multiple unique sessions).

Evaluation of SEM-based models is less straightforward than most common machine learning models. Although many frequently used models (such as multilayer perceptron models) are generally trained and evaluated in terms of their predictive performance (e.g., accuracy), SEM-based models have no directly corresponding notion of "prediction". The "success" of a SEM-model is established by the degree of model fit: how well a given model's implied structure matches a given dataset. Through the development of an abstract structure for an SEM-based model, we are encoding several hypotheses about the underlying structure within the data. By training and evaluating through model fit, we will be able to interpret the model parameters to answer many essential research questions regarding the relationship between divergent moments and the client-therapist perception of the working alliance.

Bibliography

[1] Katie Aafjes-van Doorn and Lena Müller-Frommeyer. Reciprocal language style matching in psychotherapy research. *Counselling and Psychotherapy Research*, 20(3):449–455, September 2020. ISSN 1473-3145, 1746-1405. doi: 10.1002/capr.12298. 6.2

[2] Katie Aafjes-van Doorn, John Porcerelli, and Lena Christine Müller-Frommeyer. Language style matching in psychotherapy: An implicit aspect of alliance. *Journal of Counseling Psychology*, 67(4):509–522, July 2020. ISSN 1939-2168, 0022-0167. doi: 10.1037/cou0000433. 6.2

[3] Drew H. Abney, Alexandra Paxton, Rick Dale, and Christopher T. Kello. Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. *Cognitive Processing*, 16(4):325–332, November 2015. ISSN 1612-4782, 1612-4790. doi: 10.1007/s10339-015-0648-2. 6.2

[4] Robert Allan and Stephen Scheidt. Group Psychotherapy for Patients with Coronary Heart Disease. *International Journal of Group Psychotherapy*, 48(2):187–214, April 1998. ISSN 0020-7284. doi: 10.1080/00207284.1998.11491536. 1

[5] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, fifth edition edition, May 2013. ISBN 978-0-89042-555-8 978-0-89042-557-2. doi: 10.1176/appi.books.9780890425596. 1, 1.1, 1, 7

[6] Dane Archer and Robin M. Akert. Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, 35(6):443–449, 1977. ISSN 1939-1315. doi: 10.1037/0022-3514.35.6.443. 1.1

[7] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, March 2016. doi: 10.1109/WACV.2016.7477553. 3.3

[8] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the Thirteenth IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, May 2018. doi: 10.1109/fg.2018.00019. 5.3

[9] Eunice Barbosa, Maria Amendoeira, Tiago Ferreira, Ana Sofia Teixeira, José Pinto-Gouveia, and Joào Salgado. Immersion and distancing across the therapeutic process: Relationship to symptoms and emotional arousal. *Research in Psychotherapy:*

Psychopathology, Process and Outcome, 20(2), July 2017. ISSN 2239-8031. doi: 10.4081/ripppo.2017.258. 1.1, 1.3, 1.3, 4.1

[10] Michael Barkham, Wolfgang Lutz, Louis Georges Castonguay, Allen E. Bergin, and Sol L. Garfield, editors. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. Wiley, Hoboken, NJ, 7th edition, 50th anniversary edition edition, 2021. ISBN 978-1-119-53658-1. 5.1

[11] D. H. Barlow, J. M. Gorman, M. K. Shear, and S. W. Woods. Cognitive-behavioral therapy, imipramine, or their combination for panic disorder: A randomized controlled trial. *JAMA*, 283(19):2529–2536, May 2000. ISSN 0098-7484. doi: 10.1001/jama.283.19.2529. 1

[12] Judith S. Beck. *Cognitive Behavior Therapy: Basics and Beyond*. The Guilford Press, New York, third edition edition, 2021. ISBN 978-1-4625-4419-6. 7

[13] Helena Fatouros Bergman, Gunilla Preisler, and Andrzej Werbart. Communicating with patients with schizophrenia: Characteristics of well functioning and poorly functioning communication. *Qualitative Research in Psychology*, 3(2):121–146, January 2006. ISSN 1478-0887. doi: 10.1191/1478088706qp047oa. 1, 3.1, 3.2

[14] Timothy Bickmore and Justine Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 396–403, New York, NY, USA, March 2001. Association for Computing Machinery. ISBN 978-1-58113-327-1. doi: 10.1145/365024.365304. 5.2

[15] J. J. Blanchard, K. T. Mueser, and A. S. Bellack. Anhedonia, positive and negative affect, and social functioning in schizophrenia. *Schizophrenia Bulletin*, 24(3):413–424, 1998. ISSN 0586-7614. doi: 10.1093/oxfordjournals.schbul.a033336. 2.1, 2.2, 2.4, 2.4

[16] Edward S. Bordin. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3):252–260, 1979. ISSN 0033-3204(Print). doi: 10.1037/h0085885. 5.1, 5.2

[17] Jessica L. Borelli, Kizzann A. Ramsook, Patricia Smiley, David Kyle Bond, Jessica L. West, and Katherine H. Buttitta. Language Matching Among Mother-child Dyads: Associations with Child Attachment and Emotion Reactivity: Mother-child LSM. *Social Development*, 26(3):610–629, August 2017. ISSN 0961205X. doi: 10.1111/sode.12200. 6.2

[18] A. L. Bouhuys and R. H. van den Hoofdakker. The interrelatedness of observed behavior of depressed patients and of a psychiatrist: An ethological study on mutual influence. *Journal of Affective Disorders*, 23(2):63–74, October 1991. ISSN 0165-0327. doi: 10.1016/0165-0327(91)90093-8. 1

[19] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/a:1010933404324. 5.4

[20] Benjamin Buck, Kyle S. Minor, and Paul H. Lysaker. Lexical Characteristics of Anticipatory and Consummatory Anhedonia in Schizophrenia: A Study of Language in Spontaneous Life Narratives. *Journal of Clinical Psychology*, 71(7):696–706, July 2015. ISSN

1097-4679. doi: 10.1002/jclp.22160. 2.1, 2.2, 2.4

[21] Emily A. Butler. Temporal Interpersonal Emotion Systems: The "TIES" That Form Relationships. *Personality and Social Psychology Review*, 15(4):367–393, November 2011. ISSN 1088-8683, 1532-7957. doi: 10.1177/1088868311411164. 6.2

[22] Lauren M. Bylsma, Bethany H. Morris, and Jonathan Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical Psychology Review*, 28(4): 676–691, April 2008. ISSN 0272-7358. doi: 10.1016/j.cpr.2007.10.001. 1

[23] Jane Cahill, Michael Barkham, Gillian Hardy, Anne Rees, David A. Shapiro, William B. Stiles, and Norman Macaskill. Outcomes of patients completing and not completing cognitive therapy for depression. *British Journal of Clinical Psychology*, 42(2):133–143, June 2003. ISSN 01446657. doi: 10.1348/014466503321903553. 1

[24] Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago, 1963. ISBN 978-0-395-30787-8. 6.5, 6.5

[25] Kaitlin Cannava and Graham D. Bodie. Language use and style matching in supportive conversations between strangers and friends. *Journal of Social and Personal Relationships*, 34(4):467–485, June 2017. ISSN 0265-4075, 1460-3608. doi: 10.1177/0265407516641222. 6.1, 6.2

[26] Maria R. Capecelatro, Matthew D. Sacchet, Peter F. Hitchcock, Samuel M. Miller, and Willoughby B. Britton. Major depression duration reduces appetitive word use: An elaborated verbal recall of emotional photographs. *Journal of Psychiatric Research*, 47(6): 809–815, June 2013. ISSN 00223956. doi: 10.1016/j.jpsychires.2013.01.022. 1.1

[27] Tomás Capretto, Camen Piho, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A. Martin. Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. *Journal of Statistical Software*, 103:1–29, August 2022. ISSN 1548-7660. doi: 10.18637/jss.v103.i15. 5.4

[28] Justine Cassell. Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105, January 2004. ISSN 0193-3973. doi: 10.1016/j.appdev.2003.11.003. 5.2

[29] Justine Cassell and Timothy Bickmore. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132, February 2003. ISSN 1573-1391. doi: 10.1023/A:1024026532471. 1, 4.2, 5.1, 5.2

[30] Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13 (4-5):519–538, May 1999. ISSN 0883-9514. doi: 10.1080/088395199117360. 5.2

[31] Alex S. Cohen, Annie St-Hilaire, Jennifer M. Aakre, and Nancy M. Docherty. Understanding anhedonia in schizophrenia through lexical analysis of natural speech. *Cognition and Emotion*, 23(3):569–586, April 2009. ISSN 0269-9931. doi: 10.1080/02699930802044651. 2.1, 2.2, 2.4, 2.4

[32] Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. *Applied Multiple*

Regression/Correlation Analysis for the Behavioral Sciences. Routledge, New York, third edition, July 2002. ISBN 978-0-203-77444-1. doi: 10.4324/9780203774441. 2.5

[33] Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, September 2009. doi: 10.1109/ACII.2009.5349358. 5.6

[34] Jeffrey F. Cohn, Nicholas Cummins, Julien Epps, Roland Goecke, Jyoti Joshi, and Stefan Scherer. Multimodal assessment of depression from behavioral signals. In *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2*, volume 21, pages 375–417. Association for Computing Machinery and Morgan & Claypool, October 2018. ISBN 978-1-970001-71-6. 1, 1.1

[35] Jan De Leeuw and Erik Meijer, editors. *Handbook of Multilevel Analysis*. Springer-Verlag, New York, 2008. ISBN 978-0-387-73183-4. doi: 10.1007/978-0-387-73186-5. 5.4

[36] J. deLeeuw. Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 599–609. Springer New York, New York, NY, 1992. ISBN 978-0-387-94037-3 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_37. 6.6

[37] Sidney K. D'Mello and Arthur Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, June 2010. ISSN 1573-1391. doi: 10.1007/s11257-010-9074-4. 1, 4.2

[38] Aimee L. Drolet and Michael W. Morris. Rapport in Conflict Resolution: Accounting for How Face-to-Face Contact Fosters Mutual Cooperation in Mixed-Motive Conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50, January 2000. ISSN 0022-1031. doi: 10.1006/jesp.1999.1395. 4.1

[39] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. 2.5, 3.5, 3.5, 5.4

[40] Otis Dudley Duncan. *Introduction to Structural Equation Models*. Studies in Population. Academic Press, New York, 1975. ISBN 978-0-12-224150-5. 6.5

[41] Paul Ekman and Wallace V. Friesen. Facial Action Coding System, 1978. (document), 3.3, 3.2, 3.4

[42] Ralph Waldo Emerson, publisher Ticknor and Fields, and printer H. O. Houghton (Firm). *The Conduct of Life*. Boston, Ticknor and Fields, 1860. I

[43] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *Proceedings of the Thirty-eighth IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pages 483–487, May 2013. doi: 10.1109/

icassp.2013.6637694. 5.3

[44] Lynn A. Fairbanks, Michael T. McGuire, and Candace J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology*, 91:109–119, 1982. ISSN 1939-1846. doi: 10.1037/0021-843X.91.2.109. 1

[45] Barry A. Farber. Patient self-disclosure: A review of the research. *Journal of Clinical Psychology*, 59(5):589–600, 2003. ISSN 1097-4679. doi: 10.1002/jclp.10161. 1.1, 1.3

[46] Melissa Fisher, Kelly McCoy, John H. Poole, and Sophia Vinogradov. Self and other in schizophrenia: A cognitive neuroscience perspective. *The American Journal of Psychiatry*, 165(11):1465–1472, November 2008. ISSN 1535-7228. doi: 10.1176/appi.ajp.2008.07111806. 2.2, 2.4

[47] Arlene F. Frank. The Role of the Therapeutic Alliance in the Treatment of Schizophrenia: Relationship to Course and Outcome. *Archives of General Psychiatry*, 47(3):228, March 1990. ISSN 0003-990X. doi: 10.1001/archpsyc.1990.01810150028006. 1.1, 5.1, 6.1

[48] Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa M. Bedore, and Elizabeth D. Peña. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 46–55, USA, May 2009. Association for Computational Linguistics. ISBN 978-1-932432-41-1. 2.2, 2.4

[49] Shuang Gao, Vince D. Calhoun, and Jing Sui. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, 24(11):1037–1052, 2018. ISSN 1755-5949. doi: 10.1111/cns.13048. 1

[50] Jane R. Garrison, Emilio Fernandez-Egea, Rashid Zaman, Mark Agius, and Jon S. Simons. Reality monitoring impairment in schizophrenia reflects specific prefrontal cortex dysfunction. *NeuroImage. Clinical*, 14:260–268, 2017. ISSN 2213-1582. doi: 10.1016/j.nicl.2017.01.028. 2.2

[51] Louise Gaston, Charles Marmar, Dolores Gallagher, and Larry Thompson. Alliance Prediction of Outcome Beyond in-Treatment Symptomatic Change as Psychotherapy Processes. *Psychotherapy Research*, 1(2):104–112, August 1991. ISSN 1050-3307, 1468-4381. doi: 10.1080/10503309112331335531. 1.1, 5.1, 6.1

[52] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, third edition, July 2015. ISBN 978-0-429-11307-9. doi: 10.1201/b16018. 4.2, 5.4

[53] Charles J. Gelso, Dennis M. Kivlighan Jr., and Rayna D. Markin. The real relationship and its role in psychotherapy outcome: A meta-analysis. *Psychotherapy*, 55:434–444, 2018. ISSN 1939-1536. doi: 10.1037/pst0000183. 1.1

[54] J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, March 1992. doi: 10.1109/ICASSP.1992.225858. 2.4

[55] Stephen B. Goldberg. The Secrets of Successful Mediators. *Negotiation Journal*, 21(3): 365–376, 2005. ISSN 1571-9979. doi: 10.1111/j.1571-9979.2005.00069.x. 4.1

[56] Charles Goodwin. The interactive construction of a sentence in natural conversation. *Everyday Language: Studies in Ethnomethodology*, 1979. 2.2

[57] Kimberley M. Gordon and Shaké G. Toukmanian. Is *how* it is said important? The association between quality of therapist interventions and client processing. *Counselling and Psychotherapy Research*, 2(2):88–98, June 2002. ISSN 1473-3145, 1746-1405. doi: 10.1080/14733140212331384867. 1

[58] Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. Creating Rapport with Virtual Agents. In Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé, editors, *Intelligent Virtual Agents*, Lecture Notes in Computer Science, pages 125–138, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-74997-4. doi: 10.1007/978-3-540-74997-4_12. 1, 4.2, 5.1, 5.2, 5.5

[59] Ellen L. Hamaker and Bengt Muthén. The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25(3):365–379, June 2020. ISSN 1939-1463. doi: 10.1037/met0000239. 5.4

[60] Ellen L. Hamaker, Rebecca M. Kuiper, and Raoul P. P. P. Grasman. A critique of the cross-lagged panel model. *Psychological Methods*, 20(1):102–116, 2015. ISSN 1939-1463, 1082-989X. doi: 10.1037/a0038889. 6.5, 6.5

[61] M. Hamilton. A Rating Scale for Depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1):56–62, February 1960. ISSN 0022-3050. doi: 10.1136/jnnp.23.1.56. 5.3, 6.3, 7

[62] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, 45(1):1–23, December 2007. ISSN 0163-853X, 1532-6950. doi: 10.1080/01638530701739181. 1.1

[63] David L. Hare, Samia R. Toukhsati, Peter Johansson, and Tiny Jaarsma. Depression and cardiovascular disease: A clinical review. *European Heart Journal*, 35(21):1365–1372, June 2014. ISSN 1522-9645. doi: 10.1093/eurheartj/eht462. 1

[64] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, NY, 2nd edition edition, January 2016. ISBN 978-0-387-84857-0. 2.5

[65] Robert L. Hatcher and J. Arthur Gillaspy. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research*, 16(1):12–25, January 2006. ISSN 1050-3307. doi: 10.1080/10503300500352500. 5.3, 6.4, 7

[66] Erin A. Heerey and Ann M. Kring. Interpersonal consequences of social anxiety. *Journal of Abnormal Psychology*, 116(1):125–134, February 2007. ISSN 0021-843X. doi: 10.1037/0021-843X.116.1.125. 1.3, 4.1

[67] Anne Grete Hersoug, Jon T. Monsen, Odd E. Havik, and Per Høglend. Quality of Early Working Alliance in Psychotherapy: Diagnoses, Relationship and Intrapsychic Variables

as Predictors. *Psychotherapy and Psychosomatics*, 71(1):18–27, 2002. ISSN 0033-3190, 1423-0348. doi: 10.1159/000049340. 6.5

[68] Clara Hill, Barbara Thompson, and Maureen Corbett. The Impact of Therapist Ability to Perceive Displayed and Hidden Client Reactions on Immediate Outcome in First Sessions of Brief Therapy. *Psychotherapy Research*, 2(2):143–155, January 1992. ISSN 1050-3307. doi: 10.1080/10503309212331332914. 5.1, 5.5

[69] Clara E. Hill, Elizabeth Nutt-Williams, Kristin J. Heaton, Barbara J. Thompson, and Renee H. Rhodes. Therapist retrospective recall impasses in long-term psychotherapy: A qualitative analysis. *Journal of Counseling Psychology*, 43(2):207–217, 1996. ISSN 1939-2168(Electronic),0022-0167(Print). doi: 10.1037/0022-0167.43.2.207. 5.1

[70] M. Hinchliffe, M. Lancashire, and F. J. Roberts. Eye-contact and depression: A preliminary report. *The British Journal of Psychiatry: The Journal of Mental Science*, 117(540): 571–572, November 1970. ISSN 0007-1250. doi: 10.1192/bjp.117.540.571. 3.2, 3.4

[71] Matthew D. Hoffman and Andrew Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15 (1):1593–1623, January 2014. ISSN 1532-4435. 5.4

[72] Kai Hong, Ani Nenkova, Mary E. March, Amber P. Parker, Ragini Verma, and Christian G. Kohler. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry Research*, 225(1-2):40–49, January 2015. ISSN 1872-7123. doi: 10.1016/j.psychres.2014.10.002. 2.1, 2.2

[73] Adam O. Horvath and Leslie S. Greenberg. The development of the Working Alliance Inventory. In *The Psychotherapeutic Process: A Research Handbook*, Guilford Clinical Psychology and Psychotherapy Series, pages 529–556. Guilford Press, New York, NY, US, 1986. ISBN 978-0-89862-651-3. 1.1, 5.1, 6.1

[74] Adam O. Horvath and B. Dianne Symonds. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, 38(2):139–149, 1991. ISSN 1939-2168(Electronic),0022-0167(Print). doi: 10.1037/0022-0167.38.2.139. 1.1, 5.1, 5.2, 5.5, 6.1

[75] Julian Hough and David Schlangen. Joint, Incremental Disfluency Detection and Utterance Segmentation from Speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 326–336, Valencia, Spain, April 2017. Association for Computational Linguistics. 2.4

[76] Michael J. Hove and Jane L. Risen. It's All in the Timing: Interpersonal Synchrony Increases Affiliation. *Social Cognition*, 27(6):949–960, December 2009. ISSN 0278-016X. doi: 10.1521/soco.2009.27.6.949. 6.2

[77] Zac E. Imel, Jacqueline S. Barco, Halley J. Brown, Brian R. Baucom, John S. Baer, John C. Kircher, and David C. Atkins. The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, 61(1):146–153, 2014. ISSN 1939-2168, 0022-0167. doi: 10.1037/a0034943. 6.2

[78] Molly E. Ireland and James W. Pennebaker. Language style matching in writing: Syn-

chrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99(3):549–571, 2010. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0020386. 6.1, 6.2

[79] Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, 22(1):39–44, January 2011. ISSN 0956-7976, 1467-9280. doi: 10.1177/0956797610392928. 6.1, 6.2

[80] Louise C. Johns, S. Rossell, C. Frith, F. Ahmad, D. Hemsley, E. Kuipers, and P. K. McGuire. Verbal self-monitoring and auditory verbal hallucinations in patients with schizophrenia. *Psychological Medicine*, 31:705–715, 2001. ISSN 1469-8978. doi: 10.1017/S0033291701003774. 2.2

[81] Marcia K. Johnson and Carol L. Raye. Reality monitoring. *Psychological Review*, 88: 67–85, 1981. ISSN 1939-1471. doi: 10.1037/0033-295X.88.1.67. 2.2, 2.4, 2.4

[82] Bart Jongejan, Patrizia Paggio, and Costanza Navarretta. Classifying head movements in video-recorded conversations based on movement velocity, acceleration and jerk. In *Proceedings of the Fourth European and Seventh Nordic Symposium on Multimodal Communication*, page 8, 2016. 5.3

[83] Doerteu Junghaenel, Joshua M. Smyth, and Laura Santner. Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology*, 27: 36–55, 2008. ISSN 1943-2771. doi: 10.1521/jscp.2008.27.1.36. 2.1, 2.2

[84] Jeffrey H. Kahn, Renée M. Tobin, Audra E. Massey, and Jennifer A. Anderson. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, 120(2):263–286, 2007. ISSN 0002-9556. 2.4

[85] John M. Kane, Delbert G. Robinson, Nina R. Schooler, Kim T. Mueser, David L. Penn, Robert A. Rosenheck, Jean Addington, Mary F. Brunette, Christoph U. Correll, Sue E. Estroff, Patricia Marcy, James Robinson, Piper S. Meyer-Kalos, Jennifer D. Gottlieb, Shirley M. Glynn, David W. Lynde, Ronny Pipes, Benji T. Kurian, Alexander L. Miller, Susan T. Azrin, Amy B. Goldstein, Joanne B. Severe, Haiqun Lin, Kyaw J. Sint, Majnu John, and Robert K. Heinssen. Comprehensive Versus Usual Community Care for First-Episode Psychosis: 2-Year Outcomes From the NIMH RAISE Early Treatment Program. *The American Journal of Psychiatry*, 173(4):362–372, April 2016. ISSN 1535-7228. doi: 10.1176/appi.ajp.2015.15050632. 2.1

[86] Ashish Kapoor and Rosalind W. Picard. A real-time head nod and shake detector. In *Proceedings of the Workshop on Perceptive User Interfaces (PUI 2001)*, pages 1–5, New York, NY, USA, November 2001. Association for Computing Machinery. ISBN 978-1-4503-7473-6. doi: 10.1145/971478.971509. 5.3

[87] S. R. Kay, A. Fiszbein, and L. A. Opler. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, 13(2):261–276, 1987. ISSN 0586-7614. doi: 10.1093/schbul/13.2.261. (document), 2.2, 2.3, 2.2, 2.4, 2.4, 2.4, 2.4, 3.1, 3.3, 3.2, 3.5, 4

[88] Richard S. E. Keefe, Miriam C. Arnold, Ute J. Bayen, Joseph P. McEvoy, and William H.

Wilson. Source-monitoring deficits for self-generated stimuli in schizophrenia: Multinomial modeling of data from three sources. *Schizophrenia Research*, 57(1):51–67, September 2002. ISSN 0920-9964. doi: 10.1016/s0920-9964(01)00306-1. 2.2, 2.4

[89] M. B. Keller, J. P. McCullough, D. N. Klein, B. Arnow, D. L. Dunner, A. J. Gelenberg, J. C. Markowitz, C. B. Nemeroff, J. M. Russell, M. E. Thase, M. H. Trivedi, and J. Zajecka. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *The New England Journal of Medicine*, 342(20):1462–1470, May 2000. ISSN 0028-4793. doi: 10.1056/NEJM200005183422001. 1

[90] Anke Kirsch and Stefan Brunnhuber. Facial expression and experience of emotions in psychodynamic interviews with patients with PTSD in comparison to healthy subjects. *Psychopathology*, 40(5):296–302, 2007. ISSN 0254-4962. doi: 10.1159/000104779. 1

[91] Edward B. Klein, Walter N. Stone, Mitchell W. Hicks, and Ian L. Pritchard. Understanding Dropouts. *Journal of Mental Health Counseling*, 25(2):89–100, April 2003. ISSN 1040-2861. doi: 10.17744/mehc.25.2.xhyreggxdcd0q4ny. 1

[92] Chris L. Kleinke. Gaze and eye contact: A research review. *Psychological Bulletin*, 100:78–100, 1986. ISSN 1939-1455. doi: 10.1037/0033-2909.100.1.78. 3.1

[93] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning, Boston, MA, USA, eighth edition, 2014. ISBN 978-1-133-31159-1. 1.1

[94] Anna M. Kokotovic and Terence J. Tracey. Premature termination at a university counseling center. *Journal of Counseling Psychology*, 34(1):80–82, 1987. ISSN 0022-0167. doi: 10.1037/0022-0167.34.1.80. 1

[95] Anna M. Kokotovic and Terence J. Tracey. Working alliance in the early phase of counseling. *Journal of Counseling Psychology*, 37(1):16–21, 1990. ISSN 1939-2168, 0022-0167. doi: 10.1037/0022-0167.37.1.16. 1.1, 5.1, 6.1

[96] Sander L. Koole and Wolfgang Tschacher. Synchrony in Psychotherapy: A Review and an Integrative Framework for the Therapeutic Alliance. *Frontiers in Psychology*, 7:1–17, June 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.00862. 6.2

[97] A. M. Kring and J. M. Neale. Do schizophrenic patients show a disjunctive relationship among expressive, experiential, and psychophysiological components of emotion? *Journal of Abnormal Psychology*, 105(2):249–257, May 1996. ISSN 0021-843X. doi: 10.1037//0021-843x.105.2.249. 3.4

[98] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc, Los Angeles, Calif., third edition edition, April 2012. ISBN 978-1-4129-8315-0. 3.3, 3.3, 3.5

[99] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, Boston, edition 2 edition, 2015. ISBN 978-0-12-405888-0. 5.4

[100] R. D. Laing. *The Divided Self: An Existential Study in Sanity and Madness*. Penguin Books, revised ed. edition edition, August 1965. ISBN 978-0-14-013537-4. 3.2

[101] Richard D. Lane, Lee Ryan, Lynn Nadel, and Leslie Greenberg. Memory reconsolidation, emotional arousal, and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, 38:e1, 2015/ed. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X14000041. 1.1, 1.3, 1.3, 4.1

[102] Brett Paul Laursen, Todd D. Little, and Noel A. Card, editors. *Handbook of Developmental Research Methods*. Guilford Press, New York, NY, 2012. ISBN 978-1-4625-1393-2. 6.5, 7.2

[103] I. Leudar, P. Thomas, and M. Johnston. Self-repair in dialogues of schizophrenics: Effects of hallucinations and negative symptoms. *Brain and Language*, 43(3):487–511, October 1992. ISSN 0093-934X. doi: 10.1016/0093-934x(92)90114-t. 2.2

[104] Willem J. M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, July 1983. ISSN 0010-0277. doi: 10.1016/0010-0277(83)90026-4. 2.2, 2.4

[105] Max M. Louwerse, Rick Dale, Ellen G. Bard, and Patrick Jeuniaux. Behavior Matching in Multimodal Communication Is Synchronized. *Cognitive Science*, 36(8):1404–1426, November 2012. ISSN 03640213. doi: 10.1111/j.1551-6709.2012.01269.x. 6.2

[106] Lester Luborsky. Therapist Success and Its Determinants. *Archives of General Psychiatry*, 42(6):602, June 1985. ISSN 0003-990X. doi: 10.1001/archpsyc.1985.01790290084010. 1.1, 5.1, 6.1

[107] Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10:2767, December 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.02767. 5.4

[108] J. C. Markowitz, M. Svartberg, and H. A. Swartz. Is IPT time-limited psychodynamic psychotherapy? *The Journal of Psychotherapy Practice and Research*, 7(3):185–195, 1998. ISSN 1055-050X. 7

[109] Ellen C. Marks, Clara E. Hill, and Dennis M. Kivlighan Jr. Secrets in psychotherapy: For better or worse? *Journal of Counseling Psychology*, 66:70–82, 2019. ISSN 1939-2168. doi: 10.1037/cou0000311. 1.1

[110] Daniel J. Martin, John P. Garske, and M. Katherine Davis. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, 68(3):438–450, 2000. ISSN 1939-2117, 0022-006X. doi: 10.1037/0022-006x.68.3.438. 1.1, 5.1, 6.1

[111] Rosemarie McCabe, Patrick G. T. Healey, Stefan Priebe, Mary Lavelle, David Dodwell, Richard Laugharne, Amelia Snell, and Stephen Bremner. Shared understanding in psychiatrist-patient communication: Association with treatment adherence in schizophrenia. *Patient Education and Counseling*, 93(1):73–79, October 2013. ISSN 1873-5134. doi: 10.1016/j.pec.2013.05.015. 2.2

[112] Albert Mehrabian and Susan R. Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3):248–252, 1967. ISSN 0095-8891. doi: 10.1037/h0024648. 1.1

[113] Rena Menke. Examining Nonverbal Shame Markers Among Post-pregnancy Women with Maltreatment Histories. Master's thesis, Wayne State University, United States – Michigan, 2011. 1

[114] Jörg Merten and Rainer Krause. What Makes Good Therapists Fail? In Pierre Philippot, Robert S. Feldman, and Erik J. Coats, editors, *Nonverbal Behavior in Clinical Settings*, pages 111–124. Oxford University Press, October 2003. ISBN 978-0-19-514109-2. doi: 10.1093/med:psych/9780195141092.003.0005. 3.4

[115] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1202. 2.1, 2.2

[116] S. A. Montgomery and M. Asberg. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry: The Journal of Mental Science*, 134:382–389, April 1979. ISSN 0007-1250. doi: 10.1192/bjp.134.4.382. 4

[117] Kim T. Mueser, Alan S. Bellack, Margaret S. Douglas, and Randall L. Morrison. Prevalence and stability of social skill deficits in schizophrenia. *Schizophrenia Research*, 5(2): 167–176, September 1991. ISSN 0920-9964. doi: 10.1016/0920-9964(91)90044-R. 2.2

[118] Lena C. Müller-Frommeyer, Niels A. M. Frommeyer, and Simone Kauffeld. Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, 51(3):1343–1359, June 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-1078-8. 6.4

[119] Radford M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical report, University of Toronto, 1993. 5.4

[120] Stephen Nowicki and Marshall P. Duke. Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, 18(1):9–35, March 1994. ISSN 1573-3653. doi: 10.1007/BF02169077. 5.6

[121] L. A. Opler, S. R. Kay, V. Rosado, and J. P. Lindenmayer. Positive and negative syndromes in chronic schizophrenic inpatients. *The Journal of Nervous and Mental Disease*, 172(6): 317–325, June 1984. ISSN 0022-3018. doi: 10.1097/00005053-198406000-00002. 3.1

[122] Antonio Pascual-Leone and Nikita Yeryomenko. The client "experiencing" scale as a predictor of treatment outcomes: A meta-analysis on psychotherapy process. *Psychotherapy Research*, 27(6):653–665, November 2017. ISSN 1050-3307. doi: 10.1080/10503307.2016.1152409. 1.1, 1.3, 4.1

[123] James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*, 2015. doi: 10.15781/t25p41. 2.4, 6.4

[124] John E. Perez and Ronald E. Riggio. Nonverbal Social Skills and Psychopathology. In *Nonverbal Behavior in Clinical Settings*, Series in Affective Science, pages 17–44. Oxford University Press, New York, NY, US, 2003. ISBN 978-0-19-514109-2. doi: 10.1093/med:

psych/9780195141092.003.0002. 1

[125] Emily Mower Provost, Yuan Shangguan, and Carlos Busso. UMEME: University of Michigan Emotional McGurk Effect Data Set. *IEEE Transactions on Affective Computing*, 6(4):395–409, October 2015. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2407898. 5.5

[126] Bernd Puschner, Stephanie Bauer, Leonard M. Horowitz, and Hans Kordy. The relationship between interpersonal problems and the helping alliance. *Journal of Clinical Psychology*, 61(4):415–429, April 2005. ISSN 0021-9762, 1097-4679. doi: 10.1002/jclp.20050. 6.5

[127] Stephen A. Rains. Language Style Matching as a Predictor of Perceived Social Support in Computer-Mediated Interaction Among Individuals Coping With Illness. *Communication Research*, 43(5):694–712, July 2016. ISSN 0093-6502, 1552-3810. doi: 10.1177/0093650214565920. 6.1, 6.2

[128] Fabian Ramseyer and Wolfgang Tschacher. Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, 79(3):284–295, 2011. ISSN 1939-2117, 0022-006X. doi: 10.1037/a0023419. 6.2

[129] C. Radhakrishna Rao, editor. *Linear Statistical Inference and Its Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, April 1973. ISBN 978-0-470-31643-6 978-0-471-70823-0. doi: 10.1002/9780470316436. 2.5

[130] Catherine M. Reich, Jeffrey S. Berman, Rick Dale, and Heidi M. Levitt. Vocal Synchrony in Psychotherapy. *Journal of Social and Clinical Psychology*, 33(5):481–494, May 2014. ISSN 0736-7236. doi: 10.1521/jscp.2014.33.5.481. 6.2

[131] Katharina C. H. Reinecke, Peter Joraschky, and Hedda Lausberg. Hand movements that change during psychotherapy and their relation to therapeutic outcome: An analysis of individual and simultaneous movements. *Psychotherapy Research*, 32(1):104–114, January 2022. ISSN 1050-3307, 1468-4381. doi: 10.1080/10503307.2021.1925989. 6.2

[132] David L. Rennie. Clients' deference in psychotherapy. *Journal of Counseling Psychology*, 41(4):427–437, October 1994. ISSN 0022-0167. doi: 10.1037/0022-0167.41.4.427. 5.1, 5.5

[133] Miriam Rennung and Anja S. Göritz. Prosocial Consequences of Interpersonal Synchrony: A Meta-Analysis. *Zeitschrift für Psychologie*, 224(3):168–189, July 2016. ISSN 2190-8370, 2151-2604. doi: 10.1027/2151-2604/a000252. 6.2

[134] Renee H. Rhodes, Clara E. Hill, Barbara J. Thompson, and Robert Elliott. Client retrospective recall of resolved and unresolved misunderstanding events. *Journal of Counseling Psychology*, 41(4):473–483, 1994. ISSN 1939-2168(Electronic),0022-0167(Print). doi: 10.1037/0022-0167.41.4.473. 5.1, 5.5

[135] Virginia P. Richmond, James C. McCroskey, and Mark Hickson. *Nonverbal Behavior in Interpersonal Relations*. Pearson/Allyn & Bacon, Boston, seventh edition, 2012. ISBN 978-0-205-04230-2. 1.1

[136] Laurel D. Riek, Philip C. Paul, and Peter Robinson. When my robot smiles at me:Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3(1):99–108, March 2010. ISSN 1783-8738. doi: 10.1007/s12193-009-0028-2. 5.2

[137] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, AVEC '17, pages 3–9, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-5502-5. doi: 10.1145/3133944.3133953. 5.5

[138] D. R. Rutter. Visual interaction in recently admitted and chronic long-stay schizophrenic patients. *British Journal of Social & Clinical Psychology*, 15:295–303, 1976. ISSN 0007-1293. doi: 10.1111/j.2044-8260.1976.tb00037.x. 3.2

[139] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, April 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.55. 5.4

[140] Edward Sapir. *The Unconscious Patterning of Behavior in Society.*, pages 114–142. Alfred A. Knopf, New York, 1927. doi: 10.1037/13401-006. II

[141] Stephen M. Saunders, Kenneth I. Howard, and David E. Orlinsky. The Therapeutic Bond Scales: Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 1(4):323–330, 1989. ISSN 1939-134X, 1040-3590. doi: 10.1037/1040-3590.1.4.323. 1.1, 5.1, 6.1

[142] Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, 53(2):361–382, 1977. ISSN 0097-8507. doi: 10.2307/413107. 2.2

[143] J. T. Schelde. Major depression: Behavioral markers of depression and recovery. *The Journal of Nervous and Mental Disease*, 186(3):133–140, March 1998. ISSN 0022-3018. doi: 10.1097/00005053-199803000-00001. 1

[144] Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. Automatic behavior descriptors for psychological disorder analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013. doi: 10.1109/FG.2013.6553789. 1

[145] Elizabeth Ellen Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. University of California, Berkeley, 1994. 2.4

[146] M. M. Singh, S. R. Kay, and L. A. Opler. Anticholinergic-neuroleptic antagonism in terms of positive and negative symptoms of schizophrenia: Implications for psychobiological subtyping. *Psychological Medicine*, 17(1):39–48, February 1987. ISSN 0033-2917. doi: 10.1017/s0033291700012964. 3.1

[147] Richard B. Slatcher, Simine Vazire, and James W. Pennebaker. Am "I" more important

than "we"? Couples' word use in instant messages. *Personal Relationships*, 15(4):407–424, December 2008. ISSN 13504126, 14756811. doi: 10.1111/j.1475-6811.2008.00207.x. 1.1

[148] T. Solorio, M. Sherman, Y. Liu, L. M. Bedore, E. D. Peña, and A. Iglesias. Analyzing language samples of Spanish–English bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17(3):367–395, July 2011. ISSN 1469-8110, 1351-3249. doi: 10.1017/S1351324910000252. 2.2

[149] John Sommers-Flanagan and Rita Sommers-Flanagan. *Clinical Interviewing*. Wiley, Hoboken, New Jersey, 6th edition edition, November 2016. ISBN 978-1-119-21558-5. 3.1

[150] Letizia Squarcina, Filippo Maria Villa, Maria Nobile, Enrico Grisan, and Paolo Brambilla. Deep learning for the prediction of treatment response in depression. *Journal of Affective Disorders*, 281:618–622, February 2021. ISSN 0165-0327. doi: 10.1016/j.jad.2020.11.104. 1

[151] Substance Abuse and Mental Health Services Administration. Key Substance Use and Mental Health Indicators in the United States: Results from the 2019 National Survey on Drug Use and Health. *Security Research Hub Reports*, January 2020. 1

[152] Linda Teri, Laura E. Gibbons, Susan M. McCurry, Rebecca G. Logsdon, David M. Buchner, William E. Barlow, Walter A. Kukull, Andrea Z. LaCroix, Wayne McCormick, and Eric B. Larson. Exercise Plus Behavioral Management in Patients With Alzheimer Disease: A Randomized Controlled Trial. *JAMA*, 290(15):2015–2022, October 2003. ISSN 0098-7484. doi: 10.1001/jama.290.15.2015. 1

[153] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. 3.5, 3.6

[154] Linda Tickle-Degnen and Robert Rosenthal. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, 1(4):285–293, October 1990. ISSN 1047-840X. doi: 10.1207/s15327965pli0104_1. 5.2, 5.5

[155] Ladislav Timulak and Daragh Keogh. The client's perspective on (experiences of) psychotherapy: A practice friendly review. *Journal of Clinical Psychology*, 73(11):1556–1567, 2017. ISSN 1097-4679. doi: 10.1002/jclp.22532. 1.1, 1.3

[156] Philip Tsui and Gail L. Schultz. Failure of rapport: Why psychotherapeutic engagement fails in the treatment of Asian clients. *The American Journal of Orthopsychiatry*, 55(4):561–569, October 1985. ISSN 0002-9432. doi: 10.1111/j.1939-0025.1985.tb02706.x. 4.1, 5.2, 5.5

[157] M. T. Turvey. Coordination. *American Psychologist*, 45(8):938–953, 1990. ISSN 1935-990X, 0003-066X. doi: 10.1037/0003-066x.45.8.938. 6.2

[158] Alexandria K. Vail, Tadas Baltrusaitis, Luciana Pennant, Elizabeth Liebson, Justin Baker, and Louis-Philippe Morency. Visual attention in schizophrenia: Eye contact and gaze aversion during clinical interactions. In *2017 Seventh International Conference on Af-*

fective Computing and Intelligent Interaction (ACII), pages 490–497, San Antonio, TX, October 2017. IEEE. ISBN 978-1-5386-0563-9. doi: 10.1109/acii.2017.8273644. 2.1

[159] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, November 2009. ISSN 02628856. doi: 10.1016/j.imavis.2008.11.007. 1.1

[160] Alessandro Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder. Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, January 2012. ISSN 1949-3045. doi: 10.1109/t-affc.2011.27. 1.1

[161] Strother H. Walker and David B. Duncan. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*, 54(1/2):167–179, 1967. ISSN 0006-3444. doi: 10.2307/2333860. 3.5

[162] Peter Waxer. Nonverbal cues for depression. *Journal of Abnormal Psychology*, 83:319–322, 1974. ISSN 1939-1846. doi: 10.1037/h0036706. 1, 3.1, 3.2

[163] Patricia Webbink. *The Power of the Eyes*. Springer Pub Co, 1st edition edition, January 1986. ISBN 978-0-8261-2670-2. 3.2

[164] Haolin Wei, Patricia Scanlon, Yingbo Li, David S. Monaghan, and Noel E. O'Connor. Real-time head nod and shake detection for continuous human affect recognition. In *Proceedings of the Fourteenth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2013)*, pages 1–4, July 2013. doi: 10.1109/wiamis.2013.6616148. 5.3

[165] Scott S. Wiltermuth and Chip Heath. Synchrony and Cooperation. *Psychological Science*, 20(1):1–5, January 2009. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.1467-9280.2008.02253.x. 6.2

[166] Travis J. Wiltshire, Johanne Stege Philipsen, Sarah Bro Trasmundi, Thomas Wiben Jensen, and Sune Vork Steffensen. Interpersonal Coordination Dynamics in Psychotherapy: A Systematic Review. *Cognitive Therapy and Research*, 44(4):752–773, August 2020. ISSN 0147-5916, 1573-2819. doi: 10.1007/s10608-020-10106-3. 1.3, 4.1, 6.2

[167] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. Affect-aware tutors: Recognising and responding to student affect. *International Journal of Learning Technology*, 4(3/4):129–164, October 2009. ISSN 1477-8386. doi: 10.1504/IJLT.2009.028804. 1, 4.2

[168] Torsten Wörtwein, Tadas Baltrušaitis, Eugene Laksana, Luciana Pennant, Elizabeth S. Liebson, Dost Öngür, Justin T. Baker, and Louis-Philippe Morency. Computational Analysis of Acoustic Descriptors in Psychotic Patients. In *Interspeech 2017*, pages 3256–3260. ISCA, August 2017. doi: 10.21437/Interspeech.2017-466. 2.1

[169] R. C. Young, J. T. Biggs, V. E. Ziegler, and D. A. Meyer. A Rating Scale for Mania: Reliability, Validity and Sensitivity. *The British Journal of Psychiatry*, 133(5):429–435, November 1978. ISSN 0007-1250, 1472-1465. doi: 10.1192/bjp.133.5.429. 4

[170] Amir Zadeh, Yao Chong Lim, Tadas Baltrušaitis, and Louis-Philippe Morency. Convo-

lutional Experts Constrained Local Model for 3D Facial Landmark Detection. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2519–2528, October 2017. doi: 10.1109/ICCVW.2017.296. 5.3

[171] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. 5.4