Automated Assessment of Social Cognition in People with a Schizophrenia Spectrum Disorder

Jake Portanova

A dissertation submitted
in partial fulfillment of the requirements
for a degree of

Doctor of Philosophy

University of Washington
2022

Reading Committee
Trevor Cohen, Chair
Ellen Bradley
Ben Buck

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Automated Assessment of Social Cognition in People with a Schizophrenia Spectrum Disorder

Jake Portanova

Chair of the Supervisory Committee:

Trevor Cohen

Department of Biomedical Informatics and Medical Education

Social cognitive deficits are core features of schizophrenia spectrum disorders (SSD) or schizophrenia, amongst other conditions. These deficits limit overall functioning, arguably the most important outcome when treating the mentally ill. However, as noted by former National Institute of Mental Health director Thomas Insel, “one cannot treat what they cannot measure” , and these deficits are difficult to measure in consistent and scalable ways. In this work, I leverage neural language representations (word embeddings and a deep neural network) to derive four novel measures of social cognition from transcribed responses to two video stimuli - one designed to evoke emotions, and the other representing intentions. The resulting measures are evaluated for their ability to distinguish patients with SSD from neurotypical controls, their relationships to validated measures of social cognition and SSD symptomatology, and their ability to detect the effects of an experimental therapeutic agent intended to enhance social cognitive abilities. The resulting automated measures of social cognition can mediate new approaches to the diagnosis, monitoring, treatment, and rehabilitation of people with an SSD, and other conditions involving social cognitive deficits.

For the Portanova family. Thank you for always being there for me.

Acknowledgments

The past five years have been the most influential and formative of my life, but they haven't come without significant struggle. My biggest regret during my Ph.D. is making my problems other people's problems and being a one-way friend. I would like to take this moment to acknowledge all the people that have been there for me in the dark times, listened to my problems, and put up with my lability.

Despite my best efforts, my family has never left my side. First off, thank you, Jaidyn. You are the reason I am alive. I appreciate all the on-demand therapy. You are going to be a great psychologist. Thank you, Steve and Janet, for giving me life and housing me during the pandemic. Thank you Jaclyn for helping me with the politics of graduate school. Thank you, Joey, for housing me when no one else would. At my very lowest, I turn to you. Thank you to Joey's children and my cousins, Jason and Sarah. You both have helped me tremendously improve as a human being and as an adult. Thank you to Jaye and Bruce for being there for me in the downs and celebrating every little thing I accomplish. Thank you, Jolie and Shayna, for being the best emotional support and giving the best life advice when I reach out. Thank you, Carolyn, for being my mom when I needed a mom, and thank you, Peter, for encouraging me to go to graduate school and teaching me the value of work. This would not have happened without either of you. I hope you are proud. Thank you for everything.

Thank you, Trevor Cohen. Thank you for taking me under your wing. Thank you for showing me how much I am capable of by pushing me past my limits and then pushing me further and further. Thank you for employing me at two of the lowest points in my life. Thank you for shaping my worldview and for being an outstanding role model. The qualities I admire most

from you are your loyalty, rationality, steadiness, and your prodigious intelligence, and I hope to take those with me throughout my career. You are a better mentor, boss, and advisor than I could have ever asked for. Thank you for everything. I truly believe you saved my life.

I want to thank Ellen Bradley, Ben Buck, and Ian Painter. Thank you for teaching me how to do science. I appreciate you challenging me at every turn and making me a better scientist. You are all role models, and I hope you remain in my life past this dissertation. Thank you for taking me under your wing and giving more to me than I ever could have hoped.

I would also like to thank and apologize to my friends and partners of the past five years. I appreciate you being there for me when I was such a mess, and I am sorry for not vibing. It has been a lot. I would like to point out four people from my fraternity that have truly stuck with me at low points: Brendon Josephson, Daniel Politis, Garrett Hardy, and Tony Zozaya. Thank you so much. You helped me more than you know. I'd also like to thank Hannah Burkhardt and Kathleen Muenzen from my cohort. Thank you for being besties, and thank you for sticking with me virtually while I finished in California. I hope to be friends with all of you for life. I'd also like to thank and apologize to my romantic partners during the period of my program. I am sorry for dragging you through my prioritization of academia over our relationships and my mental health. One of the greatest tragedies of my life is the loss of your friendship. Thank you for letting me lean on you until you fell over.

I am not proud of how I handled the tribulations handed to me over the past five years. However, I am very proud of who I am today, this dissertation, and what it means for science. I passionately believe that this dissertation is a foundational step in my career, which I have dedicated to improving the diagnosis, monitoring, treatment, and rehabilitation of psychotic spectrum disorders. To everyone mentioned above, I sincerely thank you for sacrificing your

time and energy over the past five years. You all have walked me through the most traumatic events I could imagine. Your support means the world to me. I hope the quality of this dissertation and my actions in life from this day forward show you that I have grown from your consistent wisdom into a person that is able to function in society and participate in a relationship. I promise to continue to grow along this journey, which I believe is a journey that will eventually lead to curing schizophrenia, and thus I wanted to acknowledge your role in improving the lives of people with schizophrenia and improving my life as well. Thank you all for everything. Your understanding and support helped me form a life worth living, a life dedicated to slowly and methodically reducing the individual burden of the horrible disorder of schizophrenia. I know it's not much, but all I can offer you is my dissertation, this acknowledgment, and a promise to be better.

Table of Contents

Chapter 1: Introduction and Overview

1.1 Introduction

Social cognitive deficits are deficits in the mental operations that underlie social interactions (Green et al., 2008). In schizophrenia, these deficits can manifest in the inability to process or produce emotions and intentions appropriately (Green et al., 2019). This dissertation concerns social cognitive deficits occurring in the psychiatric disorder of schizophrenia: a disorder that has high personal, financial, and societal costs (Jin & Mosweu, 2017). Clinical assessment – as well as pharmacological management – of patients with schizophrenia primarily addresses so-called "positive" symptoms (such as auditory hallucinations and delusional beliefs) (Kane & Marder, 1993). Less attention is given to "negative" symptoms (such as reduced initiative or affective flattening), or deficits in emotion, affect processing, and expression. Antipsychotics show limited efficacy in treating negative symptoms (Aleman et al., 2017). Furthermore, as compared with positive symptoms, deficits in social cognition are more subtle, harder to measure and less responsive to treatments (antipsychotic medications) (Kucharska-Pietura & Mortimer, 2013).

Assessment of speech in schizophrenia has been implemented using computational linguistics methods to estimate the frequencies of particular categories of words in transcribed speech (Buck et al., 2015b; Minor et al., 2015). Such automated measurement of speech has shown promise as a means to relate observable linguistic differences to known characteristics of schizophrenia. For example, investigators have found significant associations between lexical characteristics of language and negative symptoms and functioning (Cohen et al., 2009; Minor et al., 2015). Furthermore, one study indicated that frequencies of particular parts of speech are associated with scores from validated social cognition scales (Buck et al., 2015b). Taken together, these

studies suggest that computational linguistics methods have utility as a means to detect clinically meaningful linguistic differences. Furthermore, these studies have generally not involved what are by now state-of-the-art computational linguistic methods. Moving past dictionary-based methods may reveal improved signal beyond that detectable by previous methods. Nonetheless, given the correlations between social cognition deficits and negative symptoms (Piskulic & Addington, 2011) and functioning (Fett et al., 2011), the results from this work provide a promising indication that computational linguistics methods can be leveraged to identify manifestations of social cognition in people with schizophrenia.

Related work on the automated measurement of clinical manifestations in schizophrenia has focused on assessments of formal thought disorder, manifesting as disorganized speech (Elvevåg et al., 2007; Mota et al., 2012; Xu et al., 2022). While this work has resulted in a range of measures of speech coherence and connectedness, these measures are not focused on social cognition, and have not been assessed for their correlation with established measurements for social cognitive deficits in schizophrenia. There have not been equivalent attempts to create an automated, scalable assessment for social cognitive deficits.

In the current work, I address this gap in the literature by creating measures to assess social cognitive deficits from transcribed speech samples quantitively – based on the idea of measuring the *alignment* between participant responses to a stimulus and what that stimulus is expected to evoke. Alignment quantifies social cognition in relation to a normative response, rather than as a measurement on a predefined scale. Further, I evaluate the sensitivity of these measures to an experimental intervention that acutely enhances social cognitive abilities in schizophrenia: intranasal oxytocin administration (Woolley et al., 2014). With the first of these measures, termed *emotional alignment,* I use an emotionally evocative video task as a stimulus and

transcriptions of patients' spoken responses as a data source to create an automated assessment of how well a participant's described emotions align with the emotions expected to be elicited by the stimulus. Second, to measure *intentional alignment,* I use transcriptions of spoken responses to videos of geometric shapes to measure how well people understand the intentions these videos were designed to express (White et al., 2011). To evaluate responses to these stimuli, I developed two paradigms: Alignment with Assigned Labels (AAL) and Alignment with Neurotypical Controls (ANC). The proposed approach will provide two scores for each social cognitive deficit: emotion processing (perceiving and using emotions adaptively) and Theory of Mind (mental state attribution) (Green & Horan, 2010).

The overarching Alignment Paradigm is a framework for creating measurements of the appropriateness of a response. I adapt the Alignment Paradigm to create two subcomponents for measuring the similarity of a response to a normative or expected response: Alignment with Neurotypical Controls (ANC) and Alignment with Assigned Labels (AAL). Alignment with Neurotypical Controls is a simple, elegant, and flexible method that can be used to measure the appropriateness of a response when the range of expected responses is characterized with high granularity, and does not have a one-to-one relationship with each stimulus. Alignment with Assigned Labels measures the appropriateness of a response when there is a unique and pre-assigned expected response to each stimulus. In this work, I leverage these two adaptations of the Alignment Paradigm to measure emotion processing and mentalizing. However, they can be applied to other areas in mental health informatics, and perhaps health informatics in general, where responses to a stimulus can be compared to a normative response.

1.2 Research Question and Hypotheses

The research question I address in both aims of this dissertation is: "How can measurable deficits in social cognition be assessed objectively using Natural Language Processing (NLP)?" To further specify the problem of measuring social cognition, I will describe the characteristics of an optimal measure.

The desiderata for an ideal social cognition measure are as follows.

1: The measure should be automated.

2: The measure should be scalable.

3: The measure should differentiate between patients and controls in disorders with known deficits of this nature.

4: The measure should agree with existing psychometrically validated measures of social cognition.

5: The measure should correlate with symptom and endophenotype scales related to the illness in question.

6: The measure should discriminate from neurocognition.

7: The measure should be responsive to treatments for social cognition.

Several hypotheses follow from the desiderata of a social cognition measure when statistical methods are applied to them. For the emotional or intentional alignment measures developed during the course of my research, I hypothesize that:

1. There will be significant group differences between patients and controls.
2. The automated NLP measures will significantly correlate with analogous existing measures that rely on human raters.
3. The measure in question will be significantly inversely correlated with negative symptoms, and positively correlated with functioning.

4. The measure in question will retain its relationships to social cognition and functional outcomes when controlling for neurocognition.
5. The measure will detect changes after the administration of oxytocin.

1.3 Specific Aims

The driving hypotheses of this work, therefore, are that these novel measures of alignment can differentiate between patients with schizophrenia and healthy controls (H1); converge with existing measures of social cognition (H2) which is referred to as *convergent validity*; correlate with symptoms associated with social cognition (H3) which is referred to here as *criterion validity*; not correlate significantly with neurocognition (to show they are measuring a different aspect of the disorder) (H4) which is referred to as *discriminant validity*; and detect improvements from known treatments (H5). I evaluated these hypotheses with the following Specific Aims.

Aim 1: To develop an automated assessment of Emotional Alignment

To meet this aim, I created methods to automatically measures deviations from the intended or normative social response (e.g. 'joy' or the combination of several emotions) to an emotionally evocative video clip. The clips concerned were selected by the research team at the University of California, San Francisco (UCSF). First, I developed a method to measure the *Emotional Alignment with Assigned Labels (EmoAAL)* – the extent to which spoken responses align with the emotions a stimulus is expected (normatively) to evoke – of each participant with valence labels assigned to video stimuli by members of the research team. Then, I developed a method to measure *Emotional Alignment with Neurotypical Controls (EmoANC)* – the extent to which spoken responses align with those of control participants. Both of these methods rely on a deep

neural network that has been fine-tuned to recognize expressions of emotion in narrative text (Demszky et al., 2020) – it is the alignment between the emotions extracted by this neural network and either valence labels (EmoAAL) or emotions extracted from normative responses (EmoANC) that is measured. Since there are known emotion processing deficits in people with schizophrenia, I assessed differences in Emotional Alignment (EA) between cases and controls (H1). I also assessed the validity of the EA measurements through associations with validated measures of symptom severity, functioning, quality of life, and cognition. In doing so, I characterized further what these metrics do and do not measure, to evaluate the hypotheses that EA will have convergent (H2), criterion (H3), and discriminant (H4) validity. I also evaluated the responsiveness of EA measures to an oxytocin intervention expected to result in changes in social cognition with the hypothesis these measures will show an increase in emotional alignment after drug administration (H5).

Aim 2: To develop an automated assessment of assessing Intentional Alignment

To meet this aim, I created models that automatically measure whether or not a participant is appropriately recognizing the intentions (e.g. 'mocking', ‘coaxing’, ‘seducing’) represented by short video clips of animated shapes. I did so using the computational linguistics technique of neural word embeddings (vector representations of words that can be used to measure semantic relatedness) (Mikolov et al., 2013). This allowed me to measure the alignment between participant responses and labels assigned to the video clips by their designers, *Intentional Alignment with Assigned Labels (IntAAL)*. I also developed a measure of the *Intentional Alignment with Neurotypical Controls (IntANC)* – the extent to which spoken responses align with the responses of control participants. For evaluation, I calculated the correlation between these *Intentional Alignment (IA)* measures and additional data available for participants collected

using the widely-used Hinting Scale for mentalization or understanding the intentions of others. As with EA, I used t-tests and ROC curves to evaluate how each IA measure differentiates cases from controls with the hypothesis that IA will show group differences (H1). I also assessed the validity of the IA measurements through associations with validated measures of symptom severity, functioning, and cognition to further characterize what these metrics are measuring, and evaluate the hypotheses that Intentional Alignment measures have convergent (H2), criterion (H3), and discriminant (H4) validity. Lastly, I evaluated the responsiveness of IA measures to an oxytocin intervention expected to result in changes in social cognition with the hypothesis these measures will show an increase in intentional alignment after drug administration (H5).

1.4 Roadmap

Chapter 2 presents a scoping review of the importance of assessing social cognition in schizophrenia and examines existing NLP approaches to identify measures for comparison with them. The conclusions from this review provide the motivation for creating the automated measurements developed in this dissertation.

Chapter 3 provides an overview of the methods I have developed and a discussion of the methodological innovation involved. This chapter introduces the Alignment Paradigm, a formal approach to identifying how similar a participant's response is to a value representing the average of all neurotypical responses.

Chapter 4 describes the automated measurement of emotion processing in archival data from a double-blind crossover study. This was accomplished by creating two measures of emotional alignment (EA). The measures were evaluated for their utility as psychological assessments and their ability to detect the effect of an oxytocin intervention.

Chapter 5 focuses on work using data from mentalization assessments (i.e., Abell's geometric shapes videos; Happe, 1994) to develop automated mentalization assessments that consider the intentional alignment (IA) of a response to a video stimulus. This chapter discusses the development and evaluation of two novel methods to assess mentalization from transcripts of participant responses automatically. These methods are evaluated for their utility as a psychological assessment, and for their ability to detect a response to oxytocin.

Finally, Chapter 6 provides a discussion of my work's critical contributions, including innovation, limitations, and clinical and informatics implications.

Chapter 2: Psycholinguistic Associations with Social Cognition in People with Schizophrenia

This chapter originated as a thesis entitled "Psycholinguistic Associations with Social Cognition in People with Schizophrenia: A Scoping Review",

submitted in partial fulfillment of the requirements for the degree of Master of Science University of Washington 2021. My committee members were Trevor Cohen and Ben Buck.

2.1 Introduction

Schizophrenia is a disorder that affects 0.7% of the population in the United States (Saha et al., 2005). Social cognition, or the mental operations that underlie social interactions (Green et al., 2008), is impaired in over 90% of people with schizophrenia (Fiszdon et al., 2013). These deficits are associated with functional outcomes (Fett et al., 2011); however, only a small number of standardized assessments are available to measure social cognition (Pinkham et al., 2017) and these measurements are dependent on human ratings, limiting their scalability and objectivity. However, differences in social cognition have been shown to correspond with changes in language use (Buck et al., 2015b). Therefore, automated computational linguistics measures to detect social cognitive impairments may help the functional rehabilitation of people with schizophrenia by providing a way to identify those likely to benefit most from treatments targeting these impairments, and monitoring responses to these treatments over time.

Social cognition is especially important because it relates to both the functional outcomes and the severity of illness (Fett et al., 2011). In some cases, these relationships follow from the role of social-cognitive abilities in elements of the symptom scales concerned. For example, the first

item on the Cognitive Assessment Interview for Negative Symptoms (CAINS) scale (Kring et al., 2013) for negative symptoms concerns motivation to be in a romantic relationship. With deficits in aspects of social cognition such as emotion processing, romantic relationships may be more challenging for patients. This, in turn, may reduce their motivation to pursue such relationships, leading to the measurable manifestation of this deficit. Despite these research findings, many barriers prevent the widespread measurement of social cognition in treatment settings (Pinkham et al., 2017).

While social cognitive deficits are recognized as critical predictors of functioning (Fett et al., 2011), there is a lack of validated, scalable measures and consistently replicated psychosocial interventions to assess and improve social cognition, respectively. This review aims to identify the current state-of-the-art in measuring social cognition to identify gaps and opportunities for future work to address existing barriers for scalable, psychometrically valid measurements.

The primary motivating question for this review is *what linguistic measures can be used to measure known social cognitive deficits*? The scope of this review includes literature from biomedical informatics, computer science, and linguistics. The focus of the review is on linguistic manifestations of social cognition. The purpose of this review is to assess how one might measure social cognition for the diagnosis, monitoring, treatment, and rehabilitation of people with schizophrenia. The questions motivating this review of the literature are: (1) What linguistic measures are associated with social cognition deficits in people with schizophrenia? (2) Where are papers on linguistic indicators of social cognition in schizophrenia published? (3) When did research at the intersection of social cognition in schizophrenia and linguistics commence, and what methods are used in this research to identify social cognitive variables

(including direct and indirect measurements of social cognition)? As such, this review includes both methods currently used to measure social cognition, and others more generally related to it.

2.2 Significance

Social cognitive deficits impede social relationships and are related to functioning and neurocognition (Fett et al., 2011). There are barriers to assessing social cognition in clinical settings on an ongoing basis to support diagnosis, monitoring, or treating people with social cognitive deficits in schizophrenia. One barrier pertains to difficulties in measurement. (Pinkham, 2017). Linguistic measurement of mentalizing and emotional processing presents an opportunity for scalable and precise measurement (Buck et al., 2015b). Developing measures with these technologies presents opportunities to improve the quality and scalability of diagnostic procedures, as well as the capacity to evaluate the effects of social cognitive treatments objectively.

2.3 Current Evidence

Research on automated language analysis in people with schizophrenia has focused predominantly on formal thought disorder, which manifests as changes in language structure or form (Corcoran et al., 2020). Thought disorder is a feature of psychotic disorders and is detectable in spoken language (Andreasen, 1986a). Natural language processing approaches to measuring *disorganization*, an aspect of thought disorder, were first applied to people with schizophrenia in the seminal work of Elvevåg et al. (2007). The method used for this work was Latent Semantic Analysis (LSA), which creates high-dimensional vector representations of words and documents to facilitate the measurement of the relatedness between text units (Deerwester et al., 1990). Elvevåg and colleagues applied LSA to text data (i.e., transcribed

speech) from patients with schizophrenia for a variety of tasks. LSA vector representations of words are learned from large amounts of unlabeled text, such that words occurring in similar contexts will have similar vectors. The underlying assumption is that words that occur in similar contexts are meaningfully related to one another. Therefore, language that is *not* meaningfully related, as one might expect to encounter in the context of disorganized thinking, would be represented by vectors that are far apart in an LSA space. Relatedness between sequential words, phrases, or sentences can be estimated based on this "semantic distance," with larger units of text represented as the vector average of the words they contain. Elvevåg and colleagues (2007) exploited this capability to measure the coherence of language from a word association task, a verbal fluency task, a story-telling task, and structured interviews. The resulting measure of coherent speech differentiated patients from controls, with patients having a lower mean coherence in the experiments. In subsequent work, Elvevåg and colleagues (2014) derived semantic features using LSA to show group differences between patients and their siblings and controls (both patients and siblings showed a deficit, though the deficit with siblings was not significantly different from controls). Elvevåg and colleagues' methods were the first measures of disorganized speech using Natural Language Processing (NLP). Later work, described in greater detail in a subsequent section of this review, has shown that the resulting measurements can serve as features for machine learning models to predict the onset of psychosis in individuals at risk (Bedi et al., 2015).

In a 2012 paper, Mota and colleagues describe an alternative approach to characterizing speech in people with schizophrenia, using *speech graphs* (Mota et al., 2012). Speech graphs embody a graph-theoretic approach, representing words in speech as nodes, and the connections between proximal words as edges. Mota and colleagues used network analysis methods applied to the

resulting speech graphs to differentiate between people with schizophrenia and bipolar disorder, reaching a sensitivity and specificity of more than 93%, compared to 62.5% sensitivity and specificity with two interview-based symptom assessments – the Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962) and the Positive and Negative Syndrome Scale (PANSS; Kay et al., 1987). The same group extended this approach to differentiate between bipolar disorder and schizophrenia using text from dream reports (Mota et al., 2014). Subsequently, Mota et al. (2017) leveraged speech graphs to classify negative symptoms and predict the diagnosis of schizophrenia six months in advance in a cohort of 60 individuals from a public child psychiatric clinic with an accuracy of 85% in the same cohort by using the speech graph variables of Largest Connected Component (the component with the largest number of nodes when each pair has a node between them) and Largest Strongly Connected Component (the component with the largest number of nodes where each pair of nodes has a mutually reachable path) as predictors. In a 2018 paper, Mota and colleagues report results showing that deviations in speech, as measured using speech graphs, are predictors of psychosis in a cohort of 135 neurotypical individuals and 65 people who experienced psychosis.

Beyond speech graphs, others have attempted to apply models of the linguistic manifestations of schizophrenia and, more specifically, formal thought disorder. Bedi et al. predicted the conversion of high-risk individuals to psychosis using a machine learning classifier with features derived using LSA vector representations to quantify coherence (with the approach established by Elvevåg and colleagues), as well as features from part of speech tagging, to obtain perfect accuracy in predicting the onset of psychosis in a small sample of 34 youths (Bedi et al., 2015). Bedi and colleagues describe work in which they used leave-one-out cross-validation to validate their machine learning methods, and assessed their metric in terms of its relationship to

symptoms, identifying an association between their metric and the symptoms related to the disorder. In follow-up work, these results were validated in the context of a larger cohort, as well as across cohorts, with a similar method which included features of decreased semantic coherence, greater variance in that coherence, and reduced usage of possessive pronouns as variables to obtain an 83% accuracy in predicting psychosis onset across cohorts (intra-protocol) (Corcoran, Carrillo, Fernández-Slezak, et al., 2018).

Xu et al. (2020) furthered work using vector representations of speech to identify formal thought disorder by creating an automated method to predict a discrete thought disorder variable, focusing on the *global coherence* (the relatedness between elements of a narrative and its central theme) as an alternative to the sequential metrics (typically measuring relatedness between utterances that occur in sequence) that have predominated in previous work. The resulting "centroid" based methods aligned more closely with the human judgment of derailment in a set of recordings of participants describing their experiences of auditory verbal hallucinations, with the best performance for detection of severe thought disorder obtained by comparing neural embedding (a type of word vector representations) based representations of individual sentences in a transcript with their centroid, or vector average, and using the lowest of the resulting similarity scores (the minimum coherence - which was also the best aggregation method in Bedi's work) to represent a transcript (Xu et al., 2020).

Identification of formal thought disorder also does not address *poverty of content*: the occurrence of large amounts of speech without substantive meaning. In 2019, Rezaii and colleagues automatically measured poverty of content with neural word embeddings by developing an approach to measure "semantic density" in which the vector sum of word vectors representing a sentence is used as a target for reconstruction using word vectors from a fixed vocabulary, to

determine the number of word "meanings" required to represent the meaning of the sentence. The ratio of the number of word vectors required for this reconstruction to the original number of words gives the semantic density. In addition, the authors of this paper developed a method called "latent content analysis", which involves comparing the extent to which the meanings of different stimulus words are represented in text from participants. Of these methods, semantic density, as the sole predictive feature, identified semantic differences in language that predicted conversion to psychosis with an accuracy of 86.7% in a 30-patient training set, and 80% in a 10-patient validation set (Rezaii et al., 2019). Thus, the use of word vectors to measure linguistic differences that mark, and at times anticipate, psychosis is well supported by prior work. However, this work has not focused on quantifying social cognitive deficits (Corcoran et al., 2020). The lack of methods for automated quantification makes it difficult to measure social cognition related outcomes in drug and psychotherapy trials.

Researchers typically measure social cognition with questionnaires or tasks that present stimuli that elicit specific forms of social thinking. A range of social cognitive measurements in people with schizophrenia were evaluated in a series of studies in the Social Cognition Psychometric Evaluation (SCOPE) study (Pinkham et al., 2013). Pinkham et al. surveyed experts in social cognition to identify key domains of social cognition and measurements of these domains. The authors identified *emotion processing*, *attributional bias*, *social perception*, and *theory of mind* as the four main domains of social cognition most relevant to schizophrenia research. Relevant domains of social cognition, neurocognition, and perception are illustrated in **Table 1**, which is adapted from a table presented in Green et al.'s review (2019), including only those elements relevant to the current topic of discussion.

Table 1: Relevant domains of social cognition

Domain of Social Cognition	Description	Example
Emotion Processing	Ability to effectively identify emotions (e.g., facial expression) in others and to manage one's own emotions	Being able to identify from your boss' face whether he/she is angry at you
Social perception	Ability to identify social roles, rules and context from non-verbal cues including body language, prosody and social schema knowledge	Figuring out the relationship between two people based on a brief sample of conversation
Attribution bias	The way in which individuals explain the causes and make sense of social events or interactions	Jumping to the conclusion that you are in danger when you feel fearful
Mentalizing (or theory of mind)	Ability to represent the mental states of others and make inferences about their intentions and beliefs	Being able to take another person's perspective during a conversation

The SCOPE study identified four measures of emotion processing, two measures of attributional bias, three measures of social perception, and eight measures of theory of mind. In 2015, Pinkham and colleagues then analyzed the psychometric properties of eight tasks (identified by the RAND panel consensus ratings in the previous study) used to measure social cognition for the four domains they previously identified and found only the Bell Lysaker Emotion Recognition Task (BLERT) and the Hinting Task were recommended for use in clinical trials when examined with six different metrics. These metrics were (i) test-retest reliability, (ii) utility as a repeated measure, (iii) relationship to functional outcome, (iv) practicality and tolerability, (v) sensitivity to group differences, and (vi) internal consistency. Lastly, Pinkham and colleagues performed a further evaluation that identified the Penn Emotional Recognition Task as a third acceptable measure (Pinkham et al., 2018). Thus, the results of the SCOPE study revealed only two measures for emotion processing (BLERT and ER-40) and one measure for theory of mind

(Hinting Task) that were deemed acceptable for use (Pinkham et al., 2018). A follow-up study focused specifically on first-episode psychosis was reported in Ludwig et al. (2018). Using similar psychometric criteria, authors concluded that only the Hinting Task was acceptable for use. Another study that evaluated these and other tasks identified the Ambiguous Intentions Hostility Questionnaire (AIHQ), BLERT, and the Hinting Task as acceptable for use in clinical trials (Halverson et al., 2022). Taken together, this research identifies a gap in the availability of scalable and reliable measures of various social cognitive deficits. This review will help to address this gap by aggregating the existing linguistic measures to guide future research.

While social cognition has proven difficult to measure, it may be more important than other potentially more readily assessable measures such as the evaluation of neurocognition and symptom scales. Fett et al. found social cognition was associated with functional outcomes, neurocognition, and community functioning (as defined by independent living, working productivity, and functional outcomes) in their review of forty-eight independent meta-analyses (Fett et al., 2011). In this work, social cognition was defined to include theory of mind, emotion processing, and social perception. In a 2018 review, Javed and Charles summarized research about the role of cognition and social functioning, highlighting social cognition's importance in schizophrenia (Javed and Charles, 2018). The review underscores the importance of social cognition, and advances work in this area by highlighting natural language processing for social cognition as an area for future growth.

2.4 Limitations of Current Evidence

Automated assessment of formal thought disorder is methodologically intensive, involving the application of complex configurations of computational linguistics methods, with repeated validation of the ability to differentiate patient versus control status in several studies. However,

these methods were not designed to measure social cognitive deficits. The studies above by Pinkham and colleagues were comprehensive in their surveying and testing of psychological measures providing a clear picture of the relevant domains of social cognition and which scales constitute acceptable measurements from the investigators' perspective. However, only a small number of validated instruments were recommended.

A shared limitation of the linguistically-oriented studies is their relatively small sample sizes. In every study cited above related to formal thought disorder, the sample size (including healthy controls) is less than 100. Schizophrenia is a heterogeneous condition, and these samples may not capture the full breadth of the disorders. Furthermore, while the methodologies used in the research on computational measurements of formal thought disorder are relatively advanced, few describe relationships with social cognition.

2.5 Critique

By focusing on thought disorganization, natural language processing researchers have neglected social cognition in schizophrenia and, by extension, functional outcomes. One potential reason for neglecting social cognition research with natural language processing methods is that it is complex to model because it requires knowledge of multiple domains (such as emotion processing and theory of mind), and as such, unlike semantic coherence, is difficult to reduce to a single quantitative estimate for group comparisons. Furthermore, according to the SCOPE studies, the psychological community has only created a few recommended measurements of social cognition, even though those are expensive and time-consuming to administer.

2.6 Synthesis

Schizophrenia is a disorder marked by significant deficits in social cognition associated with functional outcomes. While several linguistic biomarkers for the disorder have been identified, there is no previously published review of linguistic methods applied to social cognition in schizophrenia. The current review fills this gap, by synthesizing current evidence to inform computational linguistics researchers of the constructs associated with social cognition in people with schizophrenia and thus assist in the development of automated validated metrics.

2.7 Methods

2.7.1 Study design

To select the publications in the field of social cognition in schizophrenia, I followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Extension for Scoping Reviews. I conducted a PRISMA scoping review to identify linguistic associations with social cognition in people with schizophrenia and to characterize the literature. The PRISMA scoping review provides a feasible, structured method for gathering a comprehensive set of papers from the literature and identifying key characteristics that may interest emerging NLP researchers. It also identifies articles that can be synthesized so that researchers can identify novel constructs for NLP modeling.

2.7.2 Papers Included

The papers included were articles from the PubMed database. I searched the title and abstract [tiab] to acquire these results.

Search	Results

“schizophrenia” AND “linguistic” AND (“cognition” or “cognitive”) AND “social”	43

Table 2: The left column shows the Boolean logic entered into the PubMed search, and the right column shows the results from the corresponding search.

Inclusion Criteria

I aimed to include all papers that included (1) a primary population of people with schizophrenia or schizoaffective disorder, (2) a linguistic measure, (3) a focus on cognition, and (4) a relationship to social cognition or social functioning. Social functioning was also included because of the lack of social cognition related research, and the association between social cognition and social functioning.

Exclusion criteria

I excluded papers that did not discuss linguistic indicators of social cognition in the title or abstract. For example, papers focused on speech prosody were excluded from the analysis.

I obtained my sample by applying the inclusion and exclusion criteria to the PubMed database to acquire a subset of articles that indicate linguistic indicators of social cognitive ability in schizophrenia.

2.7.3 Data collection procedures

The abstracts of the articles that meet the criteria were screened to assess appropriateness per inclusion and exclusion criteria. The psycholinguistic properties concerned and their social cognitive associations were identified for a final synthesis table. A psycholinguistic was defined

as something that possessed a linguistic property or was measured via speech. A social cognitive domain was defined as a measure of social cognition or social functioning. I aimed to identify journals in which the research was published, years of publication, and patterns in modeling methods to identify useful baseline measures for the research described in the following chapters. The data analysis was performed with R (R Core Team, 2021).

2.8 Results

2.8.1 Selection of Sources of Evidence

Table 3 summarizes the procedures used to identify articles, and the number of articles that remained after applying them. Forty-three articles were identified of which twelve were eligible. However, two were duplicates, leaving the remaining ten articles used in the review.

Table 3: Articles retrieved at each stage of the search procedure.

	Identification	Eligibility	Screening for Duplicates	Included
Search 1	43	12	2	
Total				10

2.8.2 Characteristics of Sources of Evidence

Figure 1 details the publications resulting from the PubMed search. The x-axis, Year, shows the year the publications concerned were published. The y-axis, Counts, shows the number of publications published that year.

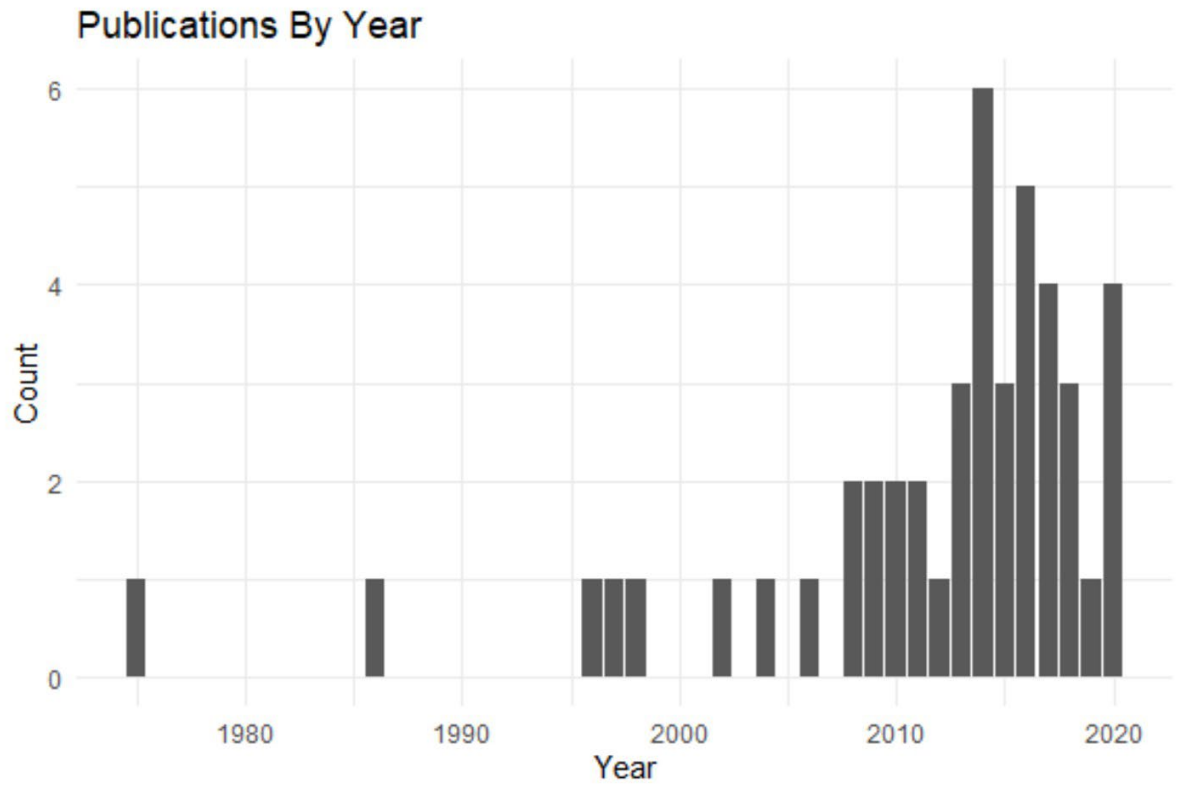


Figure 1: Number of publications meeting the search criteria in **Table 3** by year (1975 - 2020).

Journals containing papers meeting the inclusion and exclusion criteria are included in the plot shown in **Figure 1** and presented in more detail in Table 3. The journal articles come from psychiatry journals not centered around schizophrenia (n=3) and schizophrenia journals (n=5). *Journal of Schizophrenia Research* (n=3) was the best-represented journal for publishing research on linguistic manifestations in people with schizophrenia. Next, *Schizophrenia Bulletin* (n=2) and *Journal of Nervous and Mental Disorder* were the second best-represented. The *British Journal of Psychiatry* (n=1), *Psychiatric Rehabilitation Journal* (n=1), and *Psychiatry Polish* (n=1) were all tied with one observation.

Table 4 shows the sources of the publications identified by the PubMed search in Table 3.

Table 4: Publication sources.

Journal	Count
Schizophrenia Research	3
Schizophrenia Bulletin	2
Journal of Nervous and Mental Disease	2
Psychiatry Polish	1
British Journal of Psychiatry	1
Psychiatric Rehabilitation Journal	1

2.8.3 Results of Individual Sources of Evidence

Table 5 summarizes the features of the remaining 10 papers. The first column shows the linguistic indicator concerned. The second column shows whether this psycholinguistic property was related to the applied social cognitive measure. The third shows the domain of social cognition measured. The fourth defines the measure. The fifth column provides the authors of the work concerned. Twelve different psycholinguistic indicators (measures that quantify some aspect of language use) were shown to have a relationship to social cognition (**Table 6**).

Table 5: Linguistic Indicators of Social Cognition in People with Schizophrenia

Psycholinguistic indicator	Change in the psycholinguistic property as social cognitive measure increases	Social Cognitive Domain	Social Cognitive Measure	Authors
Metaphorical language and humor	Decrease	Theory of Mind, Social functioning and coping	The False-belief Task	Wyszomirska et al.
Communication failure (unclear language)	Increase	Cognition and functioning	CPT, WCST, Shipley Institute of Living Scale	N Docherty
Communication failure (unclear language)	Increase	Employment	N/A	Adamczyk et al.
Speech complexity, error	Less complexity, More errors	Cognition	Reverse Digit Span Test, Shape Cancelling Test	Thomas et al.
Linguistic complexity	Mixed results. An increase in complexity is associated with the identification of where the social problem occurred.	Neurocognition and social problem solving	MCCB, AIPSS	Moe et al.
Verbally articulated anxiety	Increase in anxiety associated with being prone to hospitalization	Isolation	Hospitalization frequency	Grand et al.
Word memory	Increase in the association of word memory and social cognitive measures	Facial affect processing and Theory of Mind	FEIT, FEDT, Hinting Task	Horton and Silverstein
Sign language acquisition and vocabulary	Increase	Theory of Mind	Hinting Task	Horton
Pronouns, second-person pronouns, prepositions, articles	Decrease in pronouns, Increase in prepositions and articles	Facial affect processing, Theory of Mind	Bell-Lysaker Emotion Recognition Task, Hinting Task	Buck et al.

Table 6: Psycholinguistic Indicators Related to Social Cognition. LIWC = Linguistic Inquiry and Word Count. NLTK = Natural Language Toolkit. Definitions largely come from the Oxford English Dictionary.

Indicator	Definition	Measurements and Citations
Metaphors	A figure of speech in which a in which the meaning of a word or phrase is not its literal meaning	Number of metaphors, experiencing amusement resulting from inaccuracies, different semantic interpretations, irony, and sarcasm
Communication Failure (unclear language)	When an utterance is unintentionally misleading, ineffective, or offensive	Communicative Development Inventory (CDI). *(Fenson et al., 2007)*
Communication Skills	The ability of a participant to communicate effectively	Right-Hemisphere Language Battery (RHLB) *(Bryan, 1995)*
Syntactic Complexity	The number of immediate constituents of a syntactic construction	Brief Syntactic Analysis (BSA), CPIDR (idea density) *(Covington, 2012)*
Error	Whether a sentence has a grammatical error	Counting deviant sentences
Anxious Language	Language indicating anxiety	Manifest Anxiety Scale (MAS)
Word memory	The ability of an individual to remember a word	Rey Auditory Verbal Learning Test *(Rey, 1964)*
Sign Language Acquisition	Whether or not someone is fluent in sign language	Patient-reported age of language acquisition
Referential Cohesion	The use of the pronouns this or that	Counting ‘this’ and ‘that.’
Pronouns and second-person pronouns	A part of speech that substitutes for nouns or noun phrases (e.g. I, me, this, that)	LIWC or NLTK Dictionary of Parts-of-Speech *(Pennebaker, 2015; Loper and Bird, 2002)*
Prepositions	Words expressing a relation to another element (e.g. in, on, by)	LIWC Dictionary of Prepositions *(Pennebaker, 2015)*
Articles	The part of speech used to indicate nouns (e.g. the, a)	LIWC Dictionary of Articles *(Pennebaker, 2015)*

Communication failure or vague language (n=2), as identified by Docherty et al. (2012) and Adamczyk et al. (2016), and syntactic complexity, as identified by Thomas et al. (1996) and Moe et al. (2018), were the most frequent psycholinguistic indications in the reviewed papers.

Metaphorical language and humor (n=1), Complexity (n=1), Anxiety words (n=1), word memory (n=1), poverty of content (n=1), referential cohesion (n=1), and descriptive statistics of lexical characteristics (i.e. prepositions, pronouns, second-person pronouns, and articles) (n=1) were each identified once in my literature review.

Several psycholinguistic indicators were associated with social cognition or social functioning. Social functioning was the often cited domain (n=4). Social functioning was identified in relation to coping skills (using humor to cope) by Wyszomirska et al. (2020), unclear communication by Docherty et al. (2012) and Moe et al. (2018), and a composite score (a composite of theory of mind and emotion processing - Buck et al. (2015)). Adamczyk et al. (2016) identified employment status as associated with executive functioning, as assessed by the Tower of Hanoi task. Buck et al. (2015) directly addressed social cognition, finding specific LIWC categories (pronouns, second-person pronouns, articles, and prepositions) to be correlated (pronouns negatively and articles/prepositions positively) with social cognition composite metrics at a significant level. Horton and Silverstein (2008) identified the psycholinguistic property of word memory as a predictor of affective (facial affect processing) and mentalizing (theory of mind) deficits. Lastly, Grand et al. (1975) identified verbally articulated anxiety as a predictor of isolation in people with schizophrenia.

2.9 Synthesis of Results

Decreased metaphorical language and humor deficits were related to improved social functioning and coping by Wyszomirska et al. (Wyszomirska et al., 2020). Upon reviewing articles that linked humor with theory of mind, these authors concluded that *concretism* (a lack of linguistic manifestations of metaphorical language and humor) is inversely correlated with social

functioning in schizophrenia (Wyszomirska et al., 2020). Furthermore, two studies found linguistic indicators of communication failure defined by Andreasen's Thought, Language, and Communication Scale or the Right Hemisphere Language Battery to be related to social functioning and employment (Andreasen, 1986) (Bryan, 1995). Thomas et al. conducted work on speech complexity and its relationship to social class in people with schizophrenia (*n*=38), mania (*n*=11), and controls (*n*=16). These authors found that lower speech complexity was associated with lower social class (Thomas et al., 1996). Moe and colleagues linked language complexity as measured by the Computerized Propositional Idea Density Rater (CPIDR; Covington, 2010) to social problem solving as measured by an Assessment of Interpersonal Problem-Solving Skills **(AIPSS),** which involves a role-playing scenario (Moe et al., 2018). Here, only the Sending Social Cues scale was associated with increased sentence complexity. Grand et al. linked verbally articulated anxiety (as measured with the Manifest Anxiety Scale, a component of the Signal Function Anxiety Score) to patients with schizophrenia experiencing more isolation in a study that involved interviewing eight non-isolated people with schizophrenia and eight isolated people with schizophrenia (Grand et al., 1975). The linguistic responses revealed that verbally articulated anxiety positively correlates with isolation in this context. Horton and Silverstein found associations between word memory and mentalizing in people with schizophrenia that were deaf (n=34) and hearing (n=31) (Horton & Silverstein, 2008). Here, mentalizing was measured with the Hinting Test, which measures the ability to infer beliefs and intentions (R. Corcoran et al., 1995). In another study included in this chapter, Horton found a positive association between the Hinting Test and language acquisition in deaf people with schizophrenia (Horton, 2010). Buck et al. derived a composite social cognition score, and found that it inversely correlated significantly with the frequency of pronoun use. Buck and colleagues also

found the frequency of prepositions and articles to be positively correlated with this composite social cognition score. For measurement, they used the LIWC software package (Pennebaker, 2001), a system that counts the proportion of words in a text that fall into manually-defined categories.

2.10 Discussion

2.10.1 Landscape of the Literature

To develop this scoping review, I found ten articles that associated a social measure with a linguistic measure in the context of schizophrenia. The first article the review describing relationships between linguistics and social cognition was published in 1975. Further articles were published sparsely until 2008, when research in this area accelerated. Identifying linguistic measures of social cognition appears to be a growing area of research. The journals publishing five papers (n = 5, 50%) identified were related to schizophrenia specifically: Schizophrenia Research and Schizophrenia Bulletin. Papers from other journals were identified once in the search, and these journals are all related to psychiatric research. Journals of this nature comprised five of those from which my final dataset was drawn. Schizophrenia Research and Schizophrenia Bulletin are the primary journals publishing papers concerning linguistic indicators of social cognitive ability in people with schizophrenia. However, there is no clear indicator of a primary research venue for such work. The analysis of the publications by year and journals indicates that research on linguistic indicators of social cognitive ability may still be in its early phases and has not yet matured.

2.10.2 Psycholinguistic Indicators of Social Cognition

Further supporting the assertion that research on psycholinguistic indicators of social cognition in schizophrenia is not mature, the number of linguistic indicators discussed in a single paper (n=1) only indicates little cohesion among researchers in this area. Communication (n=2) as measured via unclear language and language complexity (n=2) were the most common indicators. The frequency with which communication was modeled may be due to attributional bias or systematic errors in assigning meaning, a known social cognitive impairment in people with schizophrenia. Alternatively, this could be due to other known neurocognitive deficiencies such as attention deficits. Similarly, differences in language complexity may be due to lower educational attainment deficits or cognitive deficiencies. Experimental protocols and tools applied also varied across studies. Buck et al. (2015) leveraged the tool LIWC to identify descriptive statistics of language associated with a composite social cognition score. Horton and Silverstein (2008) used a task that measures language acquisition. Language complexity and metaphorical language were also identified.

Amidst this lack of methodological cohesion, this chapter provides a central venue for computational linguistics, natural language processing, data science, and machine learning researchers to learn about existing linguistic indicators for diagnosis, monitoring, and treatment. Furthermore, this chapter can inform medical researchers of the variety of linguistic properties that can be measured when studying social cognition in schizophrenia. With respect to the current work, the review described in this chapter revealed linguistic measures of social cognition from Buck and colleagues' work, which provide points of comparison for the novel methods I have developed. Not only does Buck and colleagues' work’ provide an indication that linguistic indicators of social cognition exist and can be measured, but it indicates how much

variation exists in defining normative social cognitive abilities. Several methods were used to measure multiple social cognitive domains. The following chapters describe the development of a novel measure of the normativity of a social response with a new paradigm that is informed by this review.

2.10.3 Social Cognition Related Topics

As is evident from **Table 6**, a wide variety of measures related to social cognition or social functioning were used. Social functioning was the primary outcome and was measured in various ways. Social cognition was only directly measured via computational methods once by Buck et al. (2015). The lack of direct social cognitive psycholinguistic studies in schizophrenia research indicates the lack of cohesion around social cognition in schizophrenia research. Social cognition is a broad topic, and authors have focused on different aspects of this complex construct. Horton and Silverstein (2008) addressed affect recognition and mentalizing in patients with schizophrenia, which are two of the more popular psychosocial constructs to model. Several other variables could be related to manifestations of social cognitive ability in how these deficits manifest in functional outcomes: for example, isolation, coping, or employment. Generally, linguistic measures were associated or correlated with social cognitive measures. This suggests linguistic characteristics have some relationships to social cognition that could justify the use of linguistic methods to assess social cognition.

2.10.4 Advantages and Disadvantages (Context)

An advantage of computational linguistics is that the methods that derive from this field are often both scalable and objective. Speech is a vital source of clinical information that psychiatrists and

psychologists use (Low et al., 2020). Text-based sources of information are increasingly prominent in the growing field of telemedicine, and are hence available for analysis using automated methods. Real-time assessments of social cognition from linguistic manifestations of schizophrenia and other psychiatric disorders can be accomplished if research on linguistic indicators of social cognition in schizophrenia continues to mature. With further methodological developments, such as those described in the following chapters, linguistic properties associated with social functioning may result in robust outcome measures for the emerging field of precision psychiatry, with patients selected for treatments based on their linguistic ”fingerprints.” Through the work described in this dissertation, I aim to advance this area of research to realize this vision for precision psychiatry. The chapters that follow define the alignment paradigm to measure social cognitive ability, and describe the creation of four measures of alignment that map back to known deficits.

These linguistic indicators of social cognitive ability may be used to assess the effectiveness of treatments for social functioning deficits objectively, because they are automated analysis methods. The acquisition of information about the effectiveness of interventions on negative symptoms, and the precise relationship between negative symptoms and social functioning, could be mediated by continued measurements using linguistically-informed methods. These methods are potentially advantageous in the context of the large amounts of speech that can be acquired via technology in the psychiatric setting, the psychological setting, and via cellular phones. Methods from the fields of data science and statistics may also be leveraged for downstream analysis of automated measures derived from these data.

Management of symptoms may not be the future primary goal of psychiatry and psychology. Patients who live with symptoms and maintain high levels of functioning may not require further

intervention. Instead, clinical interventions may focus on remediating domains associated with functioning broadly. This could partially account for the increase in interest in approaches to model social cognition, as reviewed in this chapter. Psycholinguistic indicators could be useful as a way to monitor remediation, particularly as methodological research related to social cognition continues to develop. Of note, this review identifies a high degree of heterogeneity in social cognitive domains and social cognitive measures used in research in this area. These proxy measures of social cognition may be used in part because the nature of normativity in this domain is difficult to define precisely. One goal of the work presented in subsequent chapters of this dissertation is to quantify normativity, without the need for a proxy measure that may be an imperfect fit for assessment of a particular social cognitive domain.

Despite the potential that measurements of linguistic associations with social cognition offer, several limitations have been identified in using linguistic models in precision psychiatry and pharmaceutical research. The main disadvantage of using linguistic methods is
that psycholinguistic methods for this purpose are not yet mature (in the sense that they are rapidly developing and standard approaches have yet to be established - as compared with machine learning, where an extensive suite of widely-used and well-validated methods is available). It is not only challenging to build algorithms, but it is also challenging to assess what an algorithm is measuring due to a dearth of annotated data from people with schizophrenia. Similarly, the populations studied in the articles above are too small to represent the heterogeneity of schizophrenia (Joyce and Roiser, 2007).

With real-time data retrieved from recording devices such as smartphones, computational linguistics methods have become much faster to implement than traditional assessment methods that depend on human evaluators. Computational linguistics methods related to social cognition

provide potentially scalable approaches to support and even to alter assessment in psychiatry and psychology. However, the development of computational linguistic methods that assess the full range of symptoms people with schizophrenia suffer from has not yet occurred. Overall, the synthesis presented in this chapter suggests that computational linguistics methods can be applied to social cognition in schizophrenia.

2.10.5 Limitations

The primary limitation of this scoping review is the inclusion of articles. This search only included papers with the disorder of schizophrenia mentioned in the abstract or title. It may not retrieve all studies of measures of social cognition in schizophrenia, or in the study of other disorders. Initially, there were only 43 unique articles retrieved, which were then reduced to 10 relevant articles that contained both a psycholinguistic indicator and a social cognitive construct. Furthermore, the search terms used for this review did not include "natural language processing," which is related to the term "linguistics."

This chapter informs the work described in the following chapters by providing an overview of the current status of the field of social cognition measurement with computational linguistics measures, providing points of comparison for the novel measures developed in this dissertation.

2.11 Conclusions

Over the past decade, several automated linguistic methods applied to text-based data sources in schizophrenia have been created, and in some cases shown to be related to aspects of social cognition. The results of this scoping review indicate that this is an emerging field. The articles

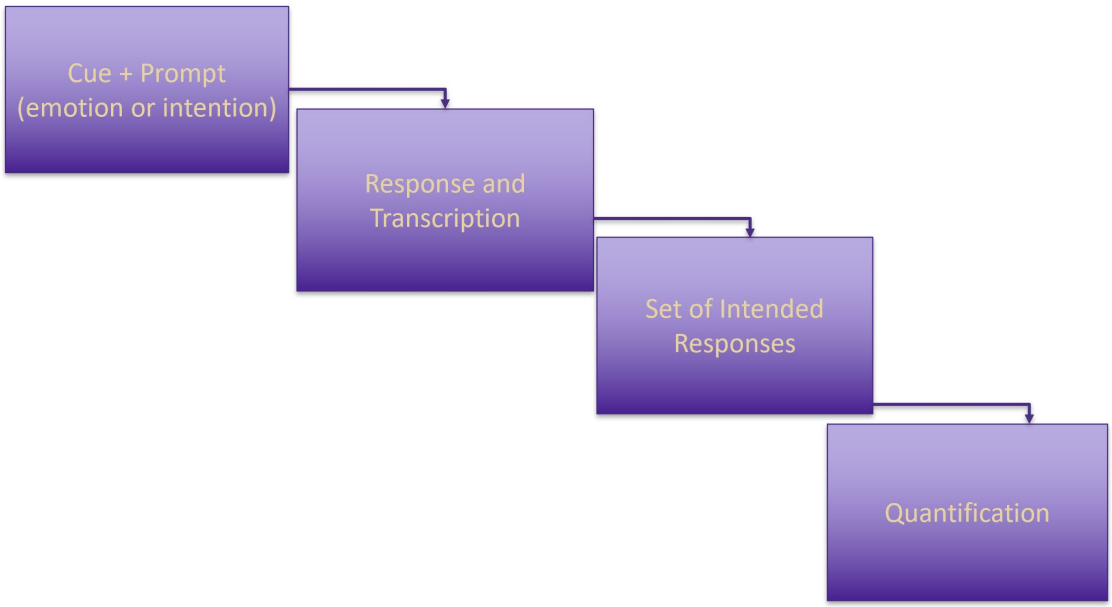
covered a wide variety of methods, experiments, and potential biomarkers. Accurate computational linguistics methods to measure social cognition in people with schizophrenia could serve as outcome measures in clinical trials to identify potential treatments for social cognitive deficits in this population. As artificial intelligence, mobile technology, and social media use increase, employing natural language processing models for mental health assessment could play a complementary role to biological psychiatry in future management of mental health.

Chapter 3: The Alignment Paradigm

3.1 Overview of Alignment

In this work, I define alignment as an agreement with either the observed responses of, or the labels assigned by, a neurotypical (or healthy control) population. The Alignment Paradigm is a framework for quantifying the appropriateness of a response.

Figure 2: The 4-step experimental process that exemplifies the Alignment Paradigm.



The key components of the Alignment Paradigm, shown in **Figure 2**, are a stimulus (e.g. an evocative video), a prompt (e.g. "How did this make you feel?"), a response to the prompt, a set of expected responses (i.e. "positive" and "negative") for a set of stimuli (i.e., the "reference" labels), and a method for the quantification of the extent to which the observed and expected responses relate to one another. In this work, these measurements are derived from transcribed

recordings, using Natural Language Processing (NLP) techniques. However, the paradigm is general in nature can accommodate the application of other methods to different types of data.

In Step 1, participants observe a stimulus and respond to a prompt such as “How did the video make you feel?”. In Step 2, the response is recorded and transcribed. In Step 3, the expected responses for the set of stimuli (e.g., a set of emotions or intentions) are used to develop vector representations of the transcribed responses (each vector can be viewed as the equivalent of a numerical row in a data frame or matrix containing continuous values only). Different methods from NLP research can be applied to the data set to relate participant responses to expected labels (e.g. “positive” or “negative”). However, each component of the resulting vectors will ultimately relate to one of the expected responses in the set. These expected responses can be derived from labels assigned to the stimuli by raters or investigators, or from the remaining neurotypical participants in the data set. These vectors can be used to evaluate how similar a given response is to each “reference” label or response. In Step 4, this is accomplished by one of two different methods: (1) summing the diagonal of the matrix (the cells containing estimates of the relatedness between a response at the "correct" label for a transcript), or (2) taking the similarity between the vector representing a participant's response to a particular stimulus, and the vector average of the responses to this stimulus from healthy controls. Henceforth, I will refer to these methods as (1) Alignment with Assigned Labels (AAL) and (2) Alignment with Neurotypical Controls (ANC).

3.2 Alignment with Assigned Labels (AAL)

Table 7 below represents an instantiation of AAL: Emotional Alignment with Assigned Labels (EmoAAL). The values in the table represent the normalized NLP-derived predicted probabilities

of the valence expected for responses to the positive and negative video stimuli, used here as the predefined expected responses or assigned labels. AAL involves (1) constructing a matrix of participant responses versus labels assigned by neurotypical people (in the example below, neurotypical raters rated the stimuli for valence); and (2) taking the trace of this matrix (i.e. summing the diagonal – where participant responses and assigned labels are matched). For AAL, one needs to have both assigned and observed labels for each stimulus/response pair. The individual's response is represented by the row label of the matrix. The "reference" label can be considered to be represented by the column label of the matrix. AAL thus quantifies the similarity between the participant response and the expected response.

Table 7: Example of AAL method. The rows correspond to participant responses. The columns correspond to expected responses.

	Positive (expected label)	Negative (expected label)
Participant response 1 (positive stimulus)	.9	.1
Participant response 2 (negative stimulus)	.4	.6

The construction of a 2x2 matrix to calculate AAL for a simple example with a set of two expected responses to two stimuli is shown in **Table 7**. The row label indicates the observation (measurements derived from the responses from the person involved using NLP). The column represents the assigned label of positive or negative (indicating the response the designer of a

stimulus expected it to evoke). Measurements are normalized at the row level to sum to one.[1] Then the trace (the sum of the diagonal) of this matrix is calculated and divided by the number of rows (or columns) to acquire an alignment score between 0 and 1. Two important features are (1) that the cells contain NLP-based measures of the extent to which a participant response relates to each of the possible assigned labels; and (2) that these values are normalized, so the diagonal indicates the *proportion* of measured relatedness that concerns the "correct" or "aligned" label. In this simple example, 90% of the observed positive response was to positively labeled stimuli, and 60% of the observed negative response was to negatively labeled stimuli. The resulting AAL score would be 1.5/2=0.75. Notice that for AAL, the set of stimuli must have multiple expected responses (e.g., positive and negative).

AAL answers the question, 'to what extent did the person respond to the evocative stimulus correctly (i.e., in accordance with an assigned "reference" label)?' However, this approach is contingent upon the availability of assigned labels that accurately reflect neurotypical responses to the stimuli concerned. The second application of the paradigm, alignment with neurotypical controls (ANC), provides a more direct approach to this problem.

3.3 Alignment with neurotypical controls (ANC)

ANC differs from AAL because it does not require a pre-existing "expected" label for a particular stimulus. Instead, it relies on a "reference" response generated from the collected data. ANC estimates the alignment of a particular response with a vector representing a prototypical neurotypical response.

[1] In this example there is only one stimulus with each label, but in the event that multiple stimuli with the same label are applied, the measurements would be averaged to make up this row

To calculate this alignment, one first quantifies the responses to the transcript as numerical vectors using Natural Language Processing methods to acquire a vector representing the extent to which each of a set of predefined categories is expressed in the transcript (e.g., % positive and % negative). Then, a vector representation of each participant from the data set is calculated (Equation 1):

$$v_{\text{participant}} = v_{\text{NLPMeasurementsForTranscript}} \quad (1)$$

For example, the vector for the first stimulus in **Table 7** would be: [0.9, 01].

Then, a vector summarizing neurotypical responses to the stimulus concerned is constructed from the remaining *n* neurotypical participants (Equation 2).

$$v_{\text{neurotypical}} = 1/\text{n} \sum_{i=1}^{n} v_{\text{participant [i]}} \quad (2)$$

The average vector can be estimated with many aggregation methods; however, for the project and analyses to follow, I use the arithmetic mean.

Once there is an average vector for neurotypical controls ($v_{\text{neurotypical}}$), this is compared to each $v_{\text{participant}}$ using the cosine similarity metric. Note that when the participant concerned is themselves neurotypical, their data are not used to construct $v_{\text{neurotypical}}$. Rather, their deviation from the responses of *other* neurotypical controls is measured.

Consider the previous example (**Table 7**). To calculate alignment in this case, the first step involves constructing the observation vector for a particular participant and stimulus from the matrix (row 1) to represent the individual. The next step involves aggregating the positive and negative measurements from the transcripts of neurotypical participants to represent normativity.

Finally, a similarity function (i.e. cosine similarity) is used to compare the two vectors and derive a similarity score, indicating the ANC for this participant in response to this stimulus.

	Positive	Negative
Observed participant	.9	.1
Average of neurotypical controls, excluding this participant	.6	.4

Table 8: Calculation of the ANC for the first stimulus provided to the participant in the example in **Table 7**.

It should be noted that the cosine similarity is not the only option to measure relatedness between vectors. For example, another commonly used operation is the scalar or "dot" product. I have selected cosine similarity because it adjusts for magnitude when the vectors concerned do not sum to one (that is, it compares the directions of the vectors without considering their length). Therefore, it innately measures the similarity between the *proportional* assignment of relatedness to each of the assigned labels (to assess whether or not participants have *focused* on the same assigned labels as neurotypical controls). As such, the resulting alignment score is designed to fit circumstances in which all assigned values may be expressed, and assigned values are expressed at different levels.

Unlike ALA, ANC does not require each stimulus to have a preassigned "reference" label assignment representing an expected response. Instead, the average response of neurotypical controls is used as a point of comparison. However, ANC still requires a set of assigned, or

"reference" labels that can apply to all examples to generate vector representations for each participant/stimulus pair; these provide the set of attributes that define what variables are used to represent normativity. However, these labels do not need to align with each stimulus for ANC to be instantiated.

3.4 Emotional and Intentional Applications of Alignment

The following chapters describe studies of applications of the Alignment Paradigm to responses to video stimuli drawn from archival data. For the first study (Chapter 4), the stimuli are emotionally evocative videos. Here, I assigned probabilities for emotions to transcripts with an NLP tool (a deep neural network) to identify emotions expressed in text. Then I applied AAL and ANC to the resulting emotion measurements (to develop EmoAAL and EmoANC). In the second study, I used the Alignment Paradigm to characterize transcriptions of participants' responses to animated shape videos (Chapter 5), which are assigned particular intentions (e.g., "mocking," "chasing") to quantify participants' understanding of intentionality using each method (intAAL and intANC). I first "vectorized" responses to each video, using neural word embeddings to draw connections between the transcribed responses and the set of intentions assigned to these video stimuli, using assigned labels from the titles of the Frith-Happe animations in the videos concerned. I applied AAL and ANC to develop Intentional Alignment scores in this case (IntAAL and IntANC).

These initial studies represent a targeted application of the AAL and ANC methods, which can be extended to other areas where agreement with preassigned responses or average normative responses is relevant.

Chapter 4: Emotional Alignment

4.1. Introduction

Social cognition concerns the mental operations that underlie the ability to identify emotions, feel connected to others, infer people's thoughts, and react emotionally to others (Green et al., 2015). People with schizophrenia, and mental illness in general, can possess social cognitive deficits and these deficits have been estimated to impact approximately 90% of people with schizophrenia (Fiszdon et al., 2013). These deficits are stable over time, nonresponsive to medication, and more robust predictors of functional outcomes than either psychotic symptoms or nonsocial cognitive deficits (Couture et al., 2006; Fett et al., 2011; Mancuso et al., 2011; Green, 2016). One such deficit identified in people with schizophrenia is a deficit in emotion processing (Green et al., 2019). Emotion processing has been defined as the ability to identify emotions accurately (e.g., facial expression) in others and to manage in a healthy manner one's own emotions (Green et al., 2019).

Emotion processing is one of the core domains of social cognitive deficits people with schizophrenia possess and is challenging to measure (Green et al., 2019). The Social Cognition Psychometric Evaluation (SCOPE) study, introduced in Chapter 2, highlighted BLERT and ER-40's strong psychometric properties for measuring emotion processing (Pinkham 2014, 2018). These strengths notwithstanding, BLERT and ER-40 focus on emotion perception only (detecting emotions accurately in the face of brief interactions), and lack the resolution to detect subtle impairments (Pinkham et al., 2018; Vaskinn & Horan, 2020). Precise and robust

measurement of emotion processing is needed to identify individuals with deficits and track changes in response to treatment.

There is evidence that natural language processing (NLP) can help to address this need. First, several studies have demonstrated that counts of words in particular language categories are associated with relevant domains in schizophrenia. Linguistic Inquiry and Word Count, or LIWC, is a software package that counts words in predefined categories to measure the frequency of words within these categories in text (Pennebaker et al., 2015). The frequencies of words in LIWC's "anger", "anxious", and "emotion" categories have been shown to correlate with measures of anhedonia, flat affect, hope, and pleasure (Bonfils et al., 2016; Buck et al., 2015a; Cohen & Minor, 2010). Also, the frequency of words in LIWC's emotion category has been shown to correlate with measures of symptom severity, functioning, and Importantly, social cognition (Buck & Penn, 2015; Minor et al., 2015). In particular, the frequency of use of words in LIWC's pronoun category has been found to be associated with a social cognition composite score (Buck et al., 2015b).

Second, several studies of language in schizophrenia-spectrum disorders have focused on measuring Formal Thought Disorder (FTD) (Corcoran et al., 2020; Xu et al., 2022). This work is motivated by prior research characterizing linguistic anomalies in schizophrenia (Andreasen, 1986). Techniques such as Latent Semantic Analysis, Speech Graphs, and Semantic Density have been shown to detect nuanced linguistic differences between individuals with schizophrenia and healthy controls (Elvevåg et al., 2007; Mota et al., 2012; Rezaii et al., 2019). These techniques have also been used to generate features for models that have predicted conversion from clinical high-risk states to psychosis (Bedi et al., 2015; Corcoran et al., 2018; Elvevåg et al.,

2007). This work provides evidence that linguistic anomalies prevalent in schizophrenia-spectrum disorders can be detected using NLP.

While this work has demonstrated promise in these other domains, the use of NLP to detect social cognitive deficits in schizophrenia is understudied. Furthering applications of NLP to use more granular methods based on neural networks, which have been shown to improve performance on benchmark NLP tasks (Devlin et al., 2018; Camacho-Collados and Pilehvar, 2018), presents a potential next step for this research direction. Applying the Alignment Paradigm (introduced in Chapter 3) to create valid measures of social cognition provides one particular path forward. Using a neural network trained to relate text to expected labels in accordance with the Alignment Paradigm could create an outcome measure to measure the effects of treatments for social cognitive deficits in people with schizophrenia. The goal of this dissertation is to extend applications of NLP to address this gap by developing NLP measures of social cognitive deficits common in schizophrenia-spectrum disorders. In doing so, this work could provide a scalable and affordable option for deployment in clinical trials.

This chapter focuses on using the Alignment Paradigm to assess emotion processing. Specifically, it presents work in which I created two automated measures of emotion processing, and evaluated them in the context of a sample of men with schizophrenia and healthy controls. My hypotheses for the work described in this chapter are (H1) these two measures will differentiate patients from controls, (H2) these measures will have psychometric validity via appropriate correlations, and (H3) these measures will be sensitive to effects of oxytocin, which I hypothesize because previous studies have demonstrated that oxytocin may improve emotion processing (Romero-Martínez et al., 2021).

I approached creating the measures in two ways, leveraging data from individual responses to a set of emotionally evocative video stimuli from YouTube clips selected by the UCSF research team. First, I created an Alignment with Assigned Labels (AAL) measure for emotion processing. Neurotypical raters annotated video stimuli with labels indicating whether they evoked a negative or positive response. Following the general approach to AAL described in Chapter 3, I created a matrix with the observed response (a measurement derived using NLP) on one axis and the expected response, as identified by assigned labels of sentiment from neurotypical raters, on the other. Each row was normalized to sum to one after aggregating by the stimulus type (i.e., positive or negative). **Table 8** presents the illustrative matrix from Chapter 3 below as an example of this.

Table 8: Matrix Illustrating AAL

	Positive (intended label)	Negative (intended label)
Observed response 1 (positive stimulus)	.9	.1
Observed response 2 (negative stimulus)	.4	.6

I then took the trace (or diagonal) of that matrix to find the sum of the similar aspects (i.e., positive expression to positive stimulus) and then divided by 2. In the example above, the measure returns an alignment score of 0.75. Second, I used a measure of Alignment with Neurotypical Controls (ANC – as described in Chapter 3) for emotion processing, which

compares each participant's emotional response (as measured via NLP) to the mean emotion vector of the healthy controls for a particular stimulus. These processes allowed me to answer the question, "*how aligned is this person's emotional response with typical emotional responses to the same stimuli?"*, with the expectation that participants with schizophrenia would have lower AAL and ANC scores than neurotypical ones, providing the means to distinguish cases from controls (H1). Further details on how AAL and ANC were implemented to measure emotion processing are provided in section 4.2 below.

My secondary goal was to evaluate AAL or ANC with respect to psychometric validity criteria (H2). I used the Hinting Task to assess convergent validity with a measure related to relevant social cognitive domains (Corcoran et al., 1995). The SCOPE studies identified the Hinting Task as a psychometrically valid theory of mind measure (Pinkham et al., 2018). Theory of mind, also known as mentalizing, is the ability to infer other people's intentions and emotions (Green et al., 2019). Theory of mind is a high-level social cognitive process (Green & Horan, 2010) that helps to drive critical interpersonal skills, such as empathy, and support fluid communication (Cotter et al., 2017). A meta-analysis illustrates that many studies have shown significant impairments in mentalizing among people with schizophrenia (Cohen's d=0.96, Salva, 2013). Mentalizing and emotion processing load on a common factor (Browne et al., 2016), and other studies have demonstrated high correlations between measures of emotion perception or processing and theory of mind. The Hinting Task was administered to all participants in this study, providing the means to examine the relationship between performance on a validated assessment of mentalizing and our automatically generated measure of emotional alignment.

To validate the measures as a potential tool for clinical trials, I explored these measures as potential tools for trials by using them to quantify oxytocin-induced changes in social cognition

(H3) by considering the effects of a single dose of oxytocin. Oxytocin is an evolutionarily conserved neuropeptide that plays a crucial role in social behavior across species (MacDonald, 2010). Oxytocin plays this role by fine-tuning sensory systems (Oettl et al., 2016), modulating stress responses (Quintana, 2016), and influencing basic computations of social value representation (Liu et al., 2019). In schizophrenia, low endogenous oxytocin levels are associated with impaired social cognition (Strauss et al., 2019), and intranasal oxytocin has been shown to improve social cognition in people with schizophrenia (Woolley et al., 2014). It is hypothesized that oxytocin improves performance through increasing the salience of social information (Bradley et al., 2020) and therefore improving mentalizing (Woolley et al., 2014). There is also evidence that oxytocin positively affects facial emotion processing (Romero-Martínez et al., 2021).

As points of comparison for these new metrics, I draw on the work of Buck et al. (2015b), Mota et al. (2017), and Elvevåg et al. (2007), including extensions by Xu et al. (2020). Buck and colleagues established that LIWC-derived estimates of pronoun use correlate with a composite of scale-based measures of social cognition (Buck et al., 2015b). Therefore, LIWC's "pronoun" variable was chosen as the first point of comparison. Measurements of speech organization emerging from the work of Mota et al. and Elvevåg et al. have been validated as indicators of Formal Thought Disorder (Elvevåg et al., 2007; Mota et al., 2012). I picked the best performing measures from these previous studies to establish comparison points to represent the standard of the field of NLP in social cognition and in assessment of schizophrenia.

4.2. Methods

4.2.1 Participants

This study involves a secondary analysis using data collected by our colleagues at the University of California, San Francisco (UCSF) led by Dr. Ellen Bradley. This study recruited 31 men diagnosed with schizophrenia or schizoaffective disorder (referred to here as "patients") from local outpatient clinics, and 51 healthy controls via online advertisements in the San Francisco Bay Area (**Table 9**). This sample was limited to men on account of evidence that oxytocin administration has sexually dimorphic effects on neural responses (Lieberz, 2020) and behavior (Gao, 2016). This decision was further motivated by work showing that epigenetic alterations in the oxytocin receptor gene (Bang, 2019) and plasma oxytocin levels (Rubin et al., 2018) have sex-dependent relationships with symptoms of schizophrenia. Patients were clinically stable, defined as having no medication changes or hospitalizations within the month preceding data collection. Controls had no psychiatric disorder within the preceding year, no lifetime history of a psychotic disorder, and no history of a psychotic disorder in a first-degree relative. All participants had no history of a neurological or substance use disorder within the preceding six months and a negative urine toxicology screen on each testing day. Participants provided written informed consent, and study protocols were approved by the Institutional Review Board at UCSF.

4.2.2 Procedures

First, all participants (patients with schizophrenia and healthy controls) completed assessments of baseline clinical characteristics (described in section 4.2.3 below). Patients then completed two testing days, separated by at least one week, randomized at each visit either to receive intranasal oxytocin or saline placebo (Wellspring Pharmacy, Berkeley, CA) per a standardized protocol (Guastella 2013). Patients completed the video response task (VRT), beginning ~30 minutes and concluding ~45 minutes following drug administration. Dr. Bradley and colleagues

selected this dosage and timing based on previous findings regarding oxytocin's behavioral and neural effects in men with schizophrenia (De Coster et al., 2019). Control participants completed the VRT only once and did not receive any drug.

To complete the VRT, participants viewed a series of 7 brief videos (30-90 seconds) on a monitor in a private room in the laboratory. Videos had been pre-selected by the investigators from YouTube for being likely to evoke a range of emotional responses and rated on valence and intensity by twelve neurotypical people in the laboratory. Two matched forms were created, so those in the patient group did not see the same videos twice; the order was randomized between participants. After watching each video, participants were prompted by on-screen instructions to respond to the question, "*How did the video make you feel?*" They were then given 30 seconds to answer aloud after the video. Participants were recorded throughout the task, and their audio responses were transcribed.

4.2.3 Measures

4.2.3.1 Baseline clinical characteristics

An adapted (positive symptom only version) semi-structured version of the Positive and Negative Syndrome Scale (PANSS; Kay et al. 1987) was used to quantify positive symptoms among patients. They quantified negative symptoms among all participants using the Clinical Assessment Interview for Negative Symptoms (CAINS; Kring et al., 2013). The CAINS is divided into motivation and pleasure (MAP) and expression (EXP) subscales, unlike other negative symptom measures. I also computed daily Chlorpromazine (CPZ) equivalents to estimate patients' anti-dopaminergic medication exposure using standardized conversion tables (Andreasen et al., 2010, Leucht, 2014). All participants also completed the Hinting Task to

assess mentalizing ability, the American National Adult Reading Test (AmNART) to estimate premorbid verbal IQ (Nelson and Willison, 1991), the Letter-Number Sequence (LNS) task to assess working memory (Valentine et al., 2020), the category fluency task to assess verbal fluency, and the Role Functioning Scale (RFS; Goodman et al., 1993) to assess real-world functioning.

4.2.3.2 Emotional alignment

Using transcribed speech from the VRT, I derived a measure of emotional alignment that reflects the similarity between the emotional content of a participant's response to a given video stimulus and the emotional content of a normative response or an assigned label. To do this, I applied a deep learning framework trained to identify the expression of emotion in text (Demszky et al., 2020). This process is summarized in section 4.2.3.2.1 below and illustrated in **Figure 3**.

4.2.3.2.1 Quantifying emotional content of participants' responses

First, I applied a pre-trained neural network developed by fine-tuning the widely used Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin et al., 2019). This network was initially trained to predict masked words, sentences, and sentence sequences in a large corpus of unannotated text. It was subsequently fine-tuned to identify the expression of emotions in text using the GoEmotions corpus (Demszky et al., 2020), which is composed of 58,000 Reddit comments manually annotated with 27 distinct emotion labels (e.g., amusement, desire, disappointment, embarrassment, etc.; see Cowen et al., 2017). These labels were found to be independent across multiple data sets when analyzed with Preserved Principal Component Analysis (PPCA; Cowen et al., 2019), which is a form of Principal Component Analysis conducted over more than one data set. The 27 emotion labels have been organized into

broad valence categories (positive, negative, ambiguous) using PPCA. I will henceforth refer to the resulting fine-tuned neural network as the *GoEmotions* model. For each transcript (i.e., a given participant's response to one video stimulus), I used the *GoEmotions* model to generate probabilities for each of the 27 emotions plus a "neutral" label reflecting the absence of any emotion. **Table 10** below is an example of the predictions of the neural network in response to an excerpt from a transcript.

Excerpt:

"I didn't like this video, . I'm not a fan of scary movies. And I feel like this was a, like a flash horror kind of thing. Um, I also think I've seen it before, and so I knew what was coming. But, uh, yeah, it didn't make me feel good per se. Um, it made me feel a little bit anxious and, uh, the suspense was killing me."

Table 10: GoEmotions Predictions. GoEmotions predictions with predicted probabilities over 0.02 (7 of the 28 categories)

disappointment	disapproval	excitement	fear	nervousness	realization	relief
0.017	0.020	0.201	0.094	0.470	0.589	0.026

4.2.3.2.2 Estimating emotional alignment with neurotypical controls (emoANC) for patients

Second, I constructed a 28-dimensional vector representation of these probabilities (27 emotion labels plus the neutral label) by concatenating vectors of probabilities assigned by the *GoEmotions* model for each label:

$$\vec{emotion}_{i,j} = [\, p(confusion),\ p(curiosity),\ \ldots\ ,\ p(sadness),\ p(relief)\,]$$

Where emotion(i,j) is the vector representing participant *i*'s response to stimulus *j*.

Third, I constructed another vector representing the normative response across all control participants to stimulus *j* as the average (centroid) of the vectors from set *C,* our sample of control participants.

$$\vec{normativeresponse}_{j} = \frac{1}{N_C} \sum_{i \in C} \vec{emotion}_{i,j}$$

Where N_C is the number of control participants.

Fourth, using cosine similarity, I quantified a given patient participant i's emotional alignment with the normative response to stimulus j. That is, a patient's emotional alignment is the cosine of the angle between their emotion vector and the centroid vector, varying between 0 for orthogonality[2] and 1 for perfect alignment:

$$emoANC = \frac{\vec{emotion}_{i,j} \bullet \vec{norm}_{j}}{|\vec{emotion}_{i,j}|\,|\vec{norm}_{j}|} = cos(\theta_{i,j})$$

Note that cosine similarity disregards the length of the vectors and focuses exclusively on their directionality. This choice of metric is appropriate because I am most interested in the distribution rather than the quantity of emotional expression. Consequently, this metric will judge two responses to be more similar if they focus on the same subset of emotions. Expressing different emotions beyond this subset will decrease similarity. Finally, I computed an average

[2] Cosine similarity varies between -1 and 1, but I did not encounter negative values in this work.

emotional alignment score for each patient (*i)* from their responses to the full set of video stimuli viewed that day, where M is the total quantity of stimuli:

$$averageAlignment_{i} = \frac{1}{M}\sum_{j=1}^{M} emoANC(i,j)$$

Each patient completed the VRT twice and thus had two scores, corresponding to the placebo and oxytocin administration days.

4.2.3.2.3 Estimating emoANC for control participants

I estimated emotional alignment with neurotypical controls (emoANC) for each control participant via leave-one-out cross-validation. Specifically, I created a 28-dimensional vector for each transcript (as for the patient group), removed the control participant from the dataset, and generated a new vector representing the normative response from the remaining controls (the centroid). I computed the cosine of the angle between the control participant's vector and the centroid, performing this procedure for each held-out control participant. As for the patients, I computed a mean score for all stimuli viewed on the testing day; each control participant completed the VRT once and thus had one score.

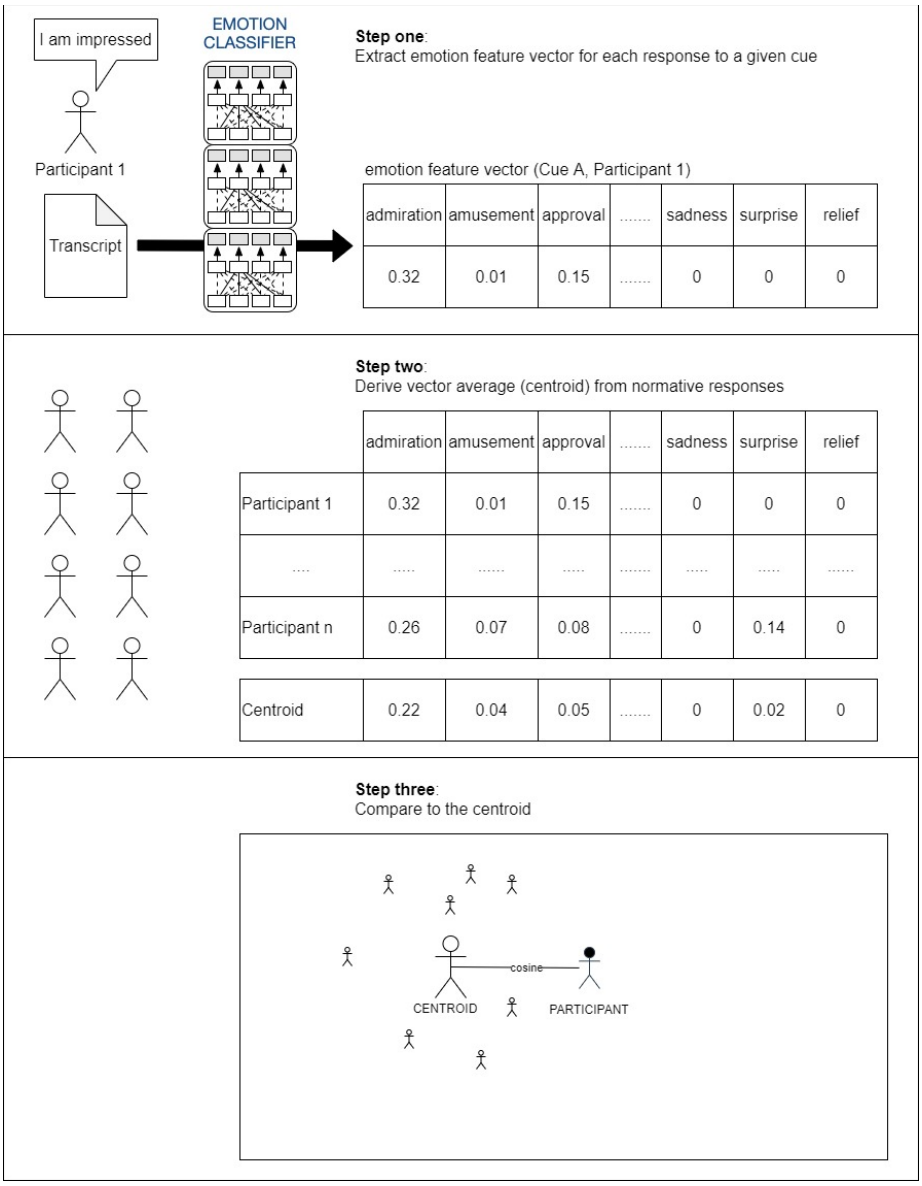


Figure 3. emoANC estimation procedure. emoANC is the cosine similarity between the emotion vector for the participant and the normative vector, which is the mean of the emotion vectors in the remaining healthy controls.

4.2.3.1 Emotional Alignment with Assigned Labels (emoAAL)

I computed a measure of emotional alignment with assigned labels (emoAAL) by considering the extent to which emotion scores for a transcript (as provided by a neural network trained to

identify expression of emotion in text) align with the emotional valence of the video being responded to (as assigned by human annotators). I applied my methods to participant responses to the question, 'how does this video make you feel?' (the same responses as with emoANC), with video stimuli categorized as ‘positive’ or ‘negative’. **Figure 3** illustrates the estimation procedure for positive video stimuli. For each transcript for an individual concerning a stimulus of this nature, the neural network provides a probabilistic estimate for groups of positive and negative emotions, ambiguous emotions, and a neutral category, with the latter two discarded. The remaining two values are then normalized to sum to one. **Figure 3** shows how these values are used to generate the ‘positive video’ row of the emotion alignment matrix (bottom row, cells labeled ‘a’ and ‘b’). Once this procedure is repeated for the ‘negative’ video stimuli, I estimate emoAAL by summing along the diagonal of this matrix (the trace), which quantifies the negative emotions extracted in response to 'negative' videos and positive emotions in response to ‘positive’ videos. I will proceed to describe each step in this procedure in detail.’

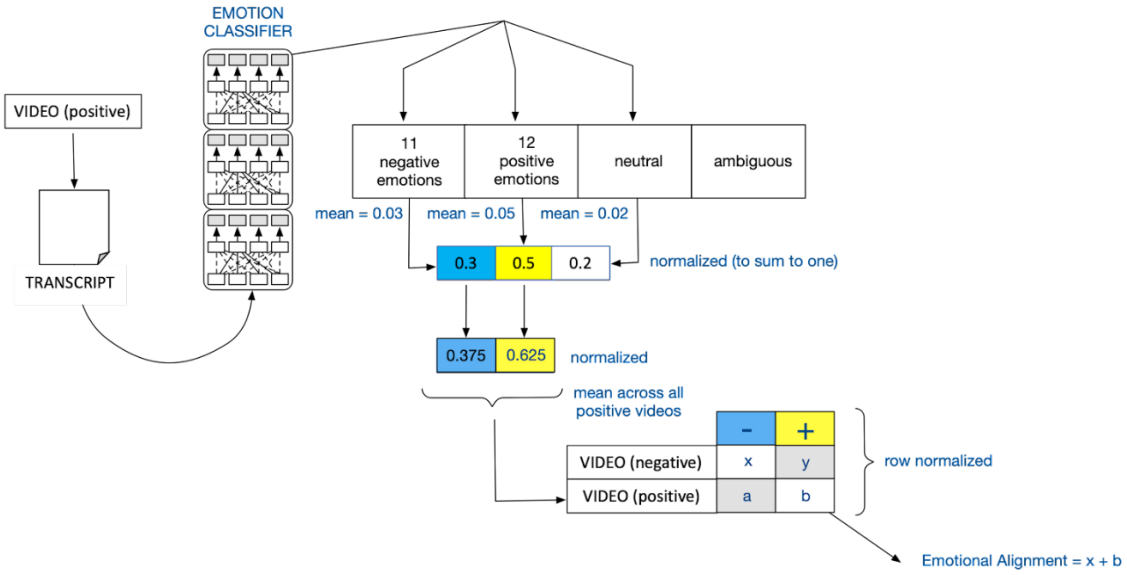


Figure 4: Estimation of emoAAL for positive video stimulus

4.2.3.1.2 Estimating emotional alignment with assigned labels

To estimate emoAAL from each run of the experimental task (i.e. across all transcripts for a particular individual/day), for each transcript in the set of responses, I added the predicted probabilities for emotions composed of the positive, negative, and neutral categories. I divided them by the total number of emotions in each category to acquire equally weighted positive, negative, and neutral emotion categories for each transcript.

Second, I created a 2x2 matrix for each participant that contained the video assigned label type as the row and the expressed emotion type (e.g. normalized sum across all "positive" emotions) as the column for each participant across all their transcripts. I further normalized the positive and negative emotion categories so that the rows (or stimuli) of the 2x2 matrix equal 1, so that the final score would range between 0 and 1. Third, I obtained the emotional alignment metric, ranging from 0 to 1, by taking the trace (the sum of the diagonal) of the 2x2 matrix and dividing this by two. Thus, I have a score between 0 and 1, representing the degree to which a participant responded with the expected emotional valence to the positive and negative videos.

4.2.3.3 Semantic coherence

To quantify the semantic coherence of each transcript, I used the Comprehensive Coherence Calculator (https://github.com/LinguisticAnomalies/Coherence) package (Xu et al., 2022). This package computes 14 different coherence measures using neural word embeddings, i.e., distributed representations of words obtained from a neural network architecture trained to predict a nearby word (or set of words) given a specific context (e.g. the surrounding words). I refer the interested reader to the work of Xu and colleagues (Xu et al., 2020, 2022). To select a specific measure that was clinically meaningful, I compared each of the measures to gold standard clinical ratings of the patient's symptoms. Specifically, I added scores for the

"Incoherent Speech" and "Conceptual Disorganization" items of the PANSS interviews. This aggregation gave me a composite clinical rating. I use this composite as a proxy for disorganized speech (an indicator of Formal Thought Disorder), in line with prior work in schizophrenia (for a review, see Corcoran et al., 2020). I assessed the correlation between this aggregate and each of the fourteen measures. A sentence-level measure that reflects the similarity between vector representations of sequential sentences generated as sums of word vectors, with the contributions of individual words weighted using inverse document frequency (Jones, 1973), had the highest absolute correlation (coherence estimates decrease as ratings increase) with clinical ratings (measure: "sentidfseq"; Spearman rho(32) =-0.419, p=0.008). I computed the minimum coherence for each transcript across all pairs of sequential sentences, the predominant aggregation strategy to generate transcript-level scores in prior work (see e.g. Corcoran et al., 2018). I then averaged these across all transcripts to obtain a single coherence score for each participant for each testing day. I used the Comprehensive Coherence Calculator package because it conveniently outputs scores from both recent state-of-the-art and older more established methods to estimate semantic coherence. The PANSS variables selected were those that related most closely to disorganized thinking and incoherent speech.

4.2.3.4 Speech graph connectedness

Mota and colleagues' Speech Graph method provides an alternative approach to modeling disorganized speech, and the SpeechGraphs software package developed by this group is freely available for researchers to use (https://www.neuro.ufrn.br/softwares/speechgraphs, see Mota et al., 2012, Mota et al., 2014). I converted each transcript into a graph using this software package. In this approach, each word corresponds to a node in a graph, and "temporal" links between consecutive words are represented by directed edges. I followed the approach described in

Spencer et al. (2021), averaging connectedness measures calculated across sequences within a transcript (bins) using a bin size of 30 and a step-size of 15 (that is, connectedness was calculated within 30-word windows that overlapped by 15 words, and these scores were averaged across all windows in the transcript). Multiple measures can be computed from each speech graph representation to quantify connectedness. Of these, I selected the Largest Strongly Connected Component (LSC) empirically, using the same method as for semantic coherence. I measured the correlation between the disorganized speech proxy and all of the individual speech graph measures and selected the measure with the strongest correlation (unlike with coherence metrics, the strongest correlation was not statistically significant ($r=0.141$, $p=0.390$)). LSC estimates the total number of nodes where any two are linked by a directed path, reachable from one another in both directions (Mota et al., 2014, 2017, Spencer et al., 2021). I calculated an average LSC value for each participant for each testing day.

4.2.3.5 Pronoun frequency

Linguistic Inquiry and Word Count (Pennebaker et al., 2015) is an automated lexical analysis program that computes the frequency of words in a given text that reflect different emotions, thinking styles, social concerns, and parts of speech, amongst other categories. LIWC has been validated as a measure of verbal expression of emotion (Kahn et al., 2007) and has previously been used to characterize abnormalities in spoken language in SSD (Buck et al., 2015b; Cohen et al., 2008; Cohen & Minor, 2010; Minor et al., 2015). In a prior study examining lexical indicators of social cognitive deficits in a sample of people with schizophrenia (Buck et al., 2015b), the use of certain parts of speech was associated with performance on a composite measure of social cognition. The measure that had the strongest relationship with social cognition in that study was pronoun use — the sum of the frequency of all pronouns was

inversely correlated with patients' social cognition scores. I analyzed each transcript using LIWC, which compares words in a transcript to a dictionary of over 4500 words and word stems. LIWC returns the proportion of words in a transcript that match the specific category of interest, “pronoun”. I computed a mean pronoun frequency score for each participant for each testing day.

4.2.4 Statistical Analyses

I conducted analyses in R (version 4.0.3; R Core Team 2022) and Python (version 3.6, Van Rossum, G., & Drake, F. L., 2009). VRT responses with <30 words (29 responses) were excluded, given that these are not compatible with speech graph analyses, and responses with <2 sentences (5 responses) were excluded, given that coherence analysis is not feasible in these cases. Furthermore, any observation that did not have a response to any positive or negative stimuli was excluded because we compared emoANC and emoAAL, which brings the data set down to 31 patients. I used unpaired t-tests to evaluate group differences in emotional alignment and the existing NLP measures. I also compared each NLP measure's performance in distinguishing patients and controls using the area under the receiver operating characteristic (AUROC) curves estimated via leave-one-out cross-validation. I used Pearson correlations to assess relationships between emotional alignment, clinical characteristics, and the existing NLP measures. To assess oxytocin's effects on performance, I used paired t-tests and explored potential moderators using correlations.

4.3. Results

4.3.1 Sample characteristics

On average, patients were older than controls and had fewer years of education (**Table 11**). Note that groups were not matched on education, given that decreased educational attainment has been identified as associated with a diagnosis of schizophrenia (Resnick, 1992). Most clinical variables had quite large differences between patients and controls.

Table 11. Sample clinical characteristics. PANSS-positive = Positive and Negative Syndrome Scale items reflect positive symptoms. CAINS-EXP and CAINS-MAP = Clinical Assessment Interview for Negative Symptoms Expressivity Subscale and Motivation and Pleasure Subscale, respectively. CPZ = Chlorpromazine. AmNART = American National Adult Reading Test. RFS = Role Functioning Scale.

	Patients (n=33)	**Controls (n=51)**	**Patients vs. Controls**
	Mean (SD)	Mean (SD)	
Age	33.42 (12.45)	29.08 (8.69)	*p=0.043, d=0.500*
Education years	14.32 (1.73)	15.33 (1.50)	*p=0.008, d=0.630*
PANSS-positive	9.82 (3.85)	-	-
CAINS	15.97 (8.43)	5.88 (3.70)	*p<0.001, d=1.680*
CAINS-EXP	2.52 (2.82)	0.373 (0.80)	*p<0.001, d=1.147*
CAINS-MAP	12.21 (6.48)	5.412 (3.47)	*p<0.001, d=1.396*
CPZ equivalents	187.90 (189.65)	-	-
AmNART	30.73 (7.30)	34.61 (5.38)	*p=0.011, d=0.626*
Hinting Task	11.57 (4.65)	14.88 (2.80)	*p<0.001, d=0.903*
RFS	20.06 (5.21)	25.51 (2.13)	*p<0.001, d=1.490*

4.3.2 Group differences in emotional alignment

Consistent with my first hypothesis (H1), I found that patients had impaired emoANC (M=0.443, SD= 0.157) relative to controls (M=0.531, SD=0.142), t(63.656)=2.630, p=0.011; Cohen's d=0.664). emoAAL results were also consistent with hypothesis H1, in that patients had

impaired emoAAL (M=0.699, SD=0.143) relative to controls (M=0.764, SD=0.127) , t(62.502)=-2.11, p=0.039, d=0.4842.

As expected based on prior literature, patients were also significantly less coherent (M=0.422, SD=0.107) than controls (M=0.507, SD=0.129), t(77.139)=3.267, p=0.002, *d*=0.700) and used pronouns more frequently (patients: M=21.09%, SD=3.052; controls: M=19.48%, SD=3.252), t(71.613)=-2.294, p=0.024; *d*=0.505. For the speech graph measure (LSC), I did not find significantly less connected speech among patients (M=18.425, SD=1.50) relative to controls (M=18.631, SD=1.302), t(61.479)=0.648, p=0.520; d=0.521. See **Figure 6**.

4.3.3. Relationships between emotional alignment and clinical characteristics

I also examined correlations between emotional alignment and patients' baseline clinical characteristics (see **Table 12**). Consistent with this hypothesis (H2), results revealed a significant association with Hinting Task performance for emoANC (r=0.461, p=0.007). However, there was not a significant association for emoAAL (r=0.308, p=0.081). Coherence (r=-0.003, p=0.986), connectedness (r=0.076, p=0.673), and pronoun use (r=-0.035, p=0.845) were not significantly correlated with Hinting Task scores.

Given that emotional alignment correlated with two measures of neurocognition (AmNART and LNS performance), I also examined partial correlations while controlling for AmNART and LNS scores; see **Table 13** and **Table 14**. The association between emotional alignment and Hinting Task performance persisted but was no longer statistically significant for emoANC when controlling for AMNART (r=0.325, p=0.069) or LNS (r=0.329, p=0.066).

emoANC was significantly negatively correlated with negative symptoms' motivation and pleasure domain (r=-0.388, p=0.026) but not with the expressivity subscale (r=-0.021, p=0.909). emoAAL was not associated with the negative symptom subscales related to motivation (r=-

0.236, p=0.187) or expressivity (r=0.075, p=0.678). Surprisingly, functioning was not associated with emoANC (r=0.229, p=0.201) nor with emoAAL (r=0.117, p=0.518). Coherence (r=-0.131, p=0.467), connectedness (r=0.192, p=0.284), and pronoun use (r=0.065, p=0.719) were also not significantly correlated with functioning.

Table 12. Correlations with emoANC and emoAAL in the patient group. **Boldface** indicates statistical significance.

	emoANC		**emoAAL**	
	Pearson's *r*	***p*-value**	**Pearson's *r***	***p*-value**
Age	-0.093	0.601	-0.192	0.285
Education years	0.165	0.358	0.243	0.173
Hinting Task	**0.464**	**0.007**	0.308	0.081
PANSS-E	-0.159	0.377	-0.022	0.903
CAINS	-0.290	0.102	-0.125	0.489
CAINS-EXP	-0.021	0.909	0.075	0.678
CAINS-MAP	**-0.388**	**0.026**	-0.236	0.187
CPZ equivalents	0.011	0.950	-0.023	0.901
AmNART	**0.470**	**0.006**	**0.475**	**0.005**
LNS	**0.443**	**0.010**	**0.480**	**0.005**
Category fluency	0.156	0.387	0.124	0.493
RFS	0.229	0.201	0.117	0.518

Table 13. Partial Correlations with emoANC and emoAAL in the patient group when controlling for AMNART.

	emoANC		**emoAAL**	
	Pearson's *r*	***p*-value**	**Pearson's *r***	***p*-value**
Age	0.045	0.807	-0.071	0.700
Education years	0.014	0.938	0.106	0.563
Hinting Task	0.325	0.069	0.131	0.475
PANSS-E	0.034	0.853	0.206	0.258
CAINS	-0.236	0.192	-0.045	0.806
CAINS-EXP	-0.0005	0.998	0.109	0.554
CAINS-MAP	-0.335	0.060	-0.158	0.388
CPZ equivalents	-0.024	0.900	-0.067	0.722
LNS	0.232	0.201	0.280	0.120
Category fluency	0.096	0.603	0.058	0.751
RFS	0.066	0.719	-0.073	0.693

Table 14. Partial Correlations with emoANC and emoAAL in the patient group when controlling for LNS.

	emoANC		**emoAAL**	
	Pearson's *r*	***p*-value**	**Pearson's *r***	***p*-value**
Age	0.060	0.742	-0.043	0.814
Education years	0.090	0.625	0.174	0.342
Hinting Task	0.329	0.066	0.120	0.514
PANSS-E	-0.219	0.229	-0.070	0.702
CAINS	-0.216	0.236	-0.017	0.928
CAINS-EXP	0.008	0.965	0.120	0.513
CAINS-MAP	-0.308	0.582	-0.121	0.509
CPZ equivalents	0.103	0.582	0.056	0.766
AMNART	0.289	0.109	0.270	0.134
Category fluency	0.102	0.579	0.060	0.741
RFS	0.102	0.577	-0.044	0.812

4.3.4. Relationships between emotional alignment and existing NLP measures

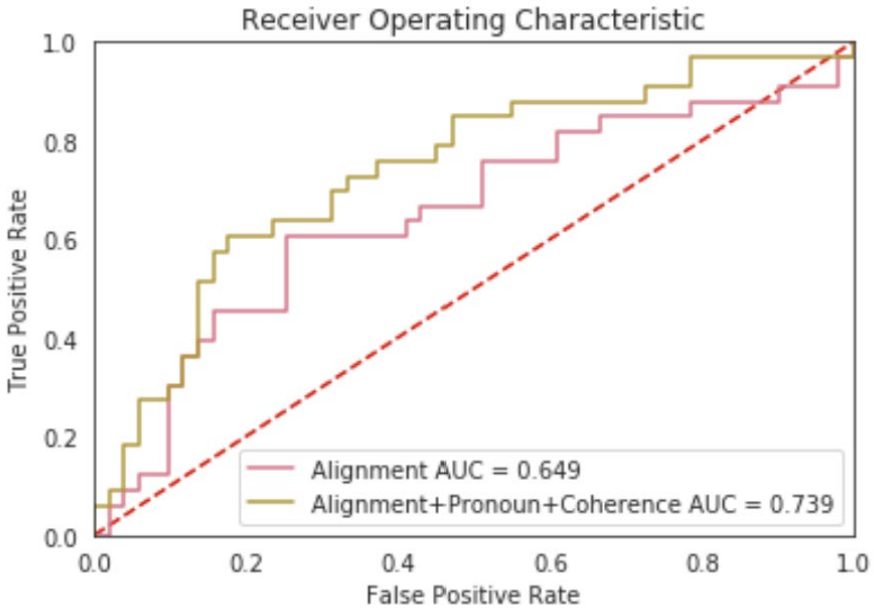
I explored whether each novel measure was associated with existing NLP measures to further characterize the emotional alignment measures. I fit a linear regression model to each pair of measures, with group status as a covariate. Across the entire sample, emoANC was not

significantly associated with coherence (beta=0.248, p=0.068), speech graph connectedness (beta=0.006, p=0.596), or pronoun use (beta=0.911, p=0.077). Examining correlations within the patient group only, these relationships shifted as follows: coherence (r=0.395, p=0.023), speech graph connectedness (r=0.211, p=0.238), and pronoun use (r=-0.028, p=0.876).

Across the entire sample in the linear regression model with case status as a covariate, emoAAL was not significantly associated with coherence (beta=0.182, p=0.136), speech graph connectedness (beta=0.004, p=0.6935), or pronoun use (beta=0.576, p=0.217). emoAAL was significantly associated with emoANC (beta=0.568, p<0.001). Within the patient group, emoAAL was also not significantly associated with existing NLP methods: coherence (r=0.297, p=0.093), speech graph connectedness (r=0.006, p=0.974), and pronoun use (r=-0.023, p=0.874).

I also explored how emotional alignment predicted participants' group status (patient vs. control) relative to the other NLP measures, motivated by previous work using coherence (Elvevåg et al., 2007), connectedness (Mota et al., 2017), and lexical characteristics (Buck et al., 2015b) to identify schizophrenia. When cross-validated (leave-one-participant-out cross-validation), emoANC as an individual predictor yielded an AUROC of 0.649 (n=85). emoAAL as an individual predictor yielded an AUROC of 0.616 when cross-validated. These results are similar to the performance of the other NLP measures as individual predictors (coherence=0.669; pronoun use=0.638). Speech graph connectedness performed much more poorly than the other measures (connectedness=0.478). Combining any two measures did not meaningfully improve prediction but combining the three best-performing measures (emoANC, coherence, pronoun use) in a logistic regression model resulted in an AUROC of 0.739. See **Figure 5**.

Figure 5. ROC curve and AUCs for group prediction using Alignment (emoANC) as an individual feature and combined with the three top-performing NLP measures in a logistic regression model.



4.3.5 Oxytocin effects on emotional alignment among patients

Consistent with hypothesis H3, oxytocin administration was associated with better emoANC among patients (M=0.522, SD=0.157) relative to the placebo group (M=0.447, SD=0.157), t(30)=3.126, p= 0.003; d=0.481); see **Figure 6**. Also consistent with H3, oxytocin administration was associated with better emoAAL among patients (M=0.522, SD=0.153) relative to placebo (M=0.447, SD=0.161), t(30)=0.976, p=0.337 ; d=0.174); however, unlike emoANC, this association was not significant. Oxytocin administration was not significantly associated with reduced pronoun use (oxytocin: M=19.944, SD=3.497; placebo: M=21.179, SD=3.031 t(30)=-1.742, p=0.092, d=-0.266). I found no significant effect of oxytocin on coherence, t(30)=-1.617, p=0.116, or on speech graph connectedness, t(30)= -0.983, p=0.334.

Figure 6. Associations between oxytocin administration and automatically computed language measures. Performance of the control group is shown for reference. A. Emotional alignment with Neurotypical Controls (emoANC). B. Coherence. C. Pronoun use.

Figure 6a. emoANC

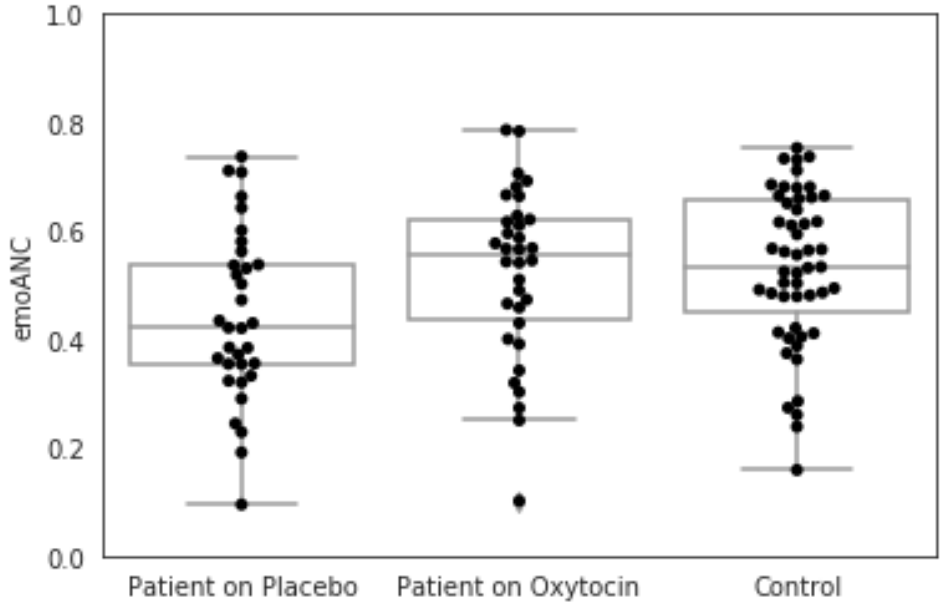


Figure 6b. Coherence

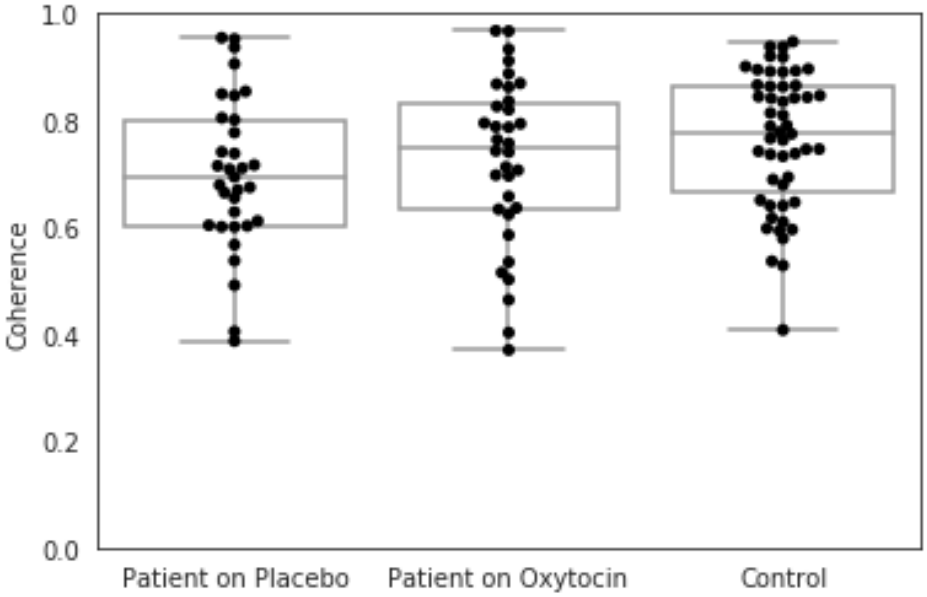
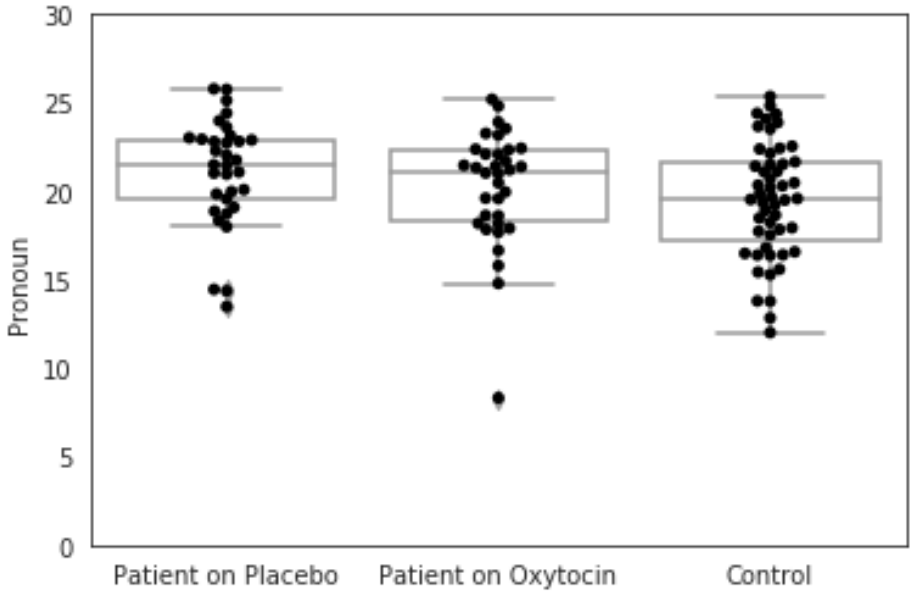


Figure 6c. Pronoun use



4.4.4 Discussion

Table 15: emoANC and emoAAL in relation to desiderata for a social cognition measure

Desideratum	**emoANC**	**emoAAL**
1. Automated	√	√
2. Scalable	√	√
3. Case/control differentiation	√	√
4. Agreement with validated measures of social cognition	√	**x**
5. Correlation with symptom/endophenotype scales	**x (0/2)**	**x (0/2)**
5a. Negative Symptoms	**X**	**x**
5b. Functioning	**X**	**x**
6. Distinct from neurocognition	**x (1/3)**	**x (1/3)**
6a. AMNART	**X**	**x**
6b. LNS	**X**	**x**
6c. Category Fluency	√	√
7. Detects oxytocin effects	√	**x**

By design, both measures satisfy the first two desiderata (**Table 15**). They are objectively derived using automated methods in a reproducible manner. They are both also scalable due to their computational nature, and in theory the technologies could run on their own and a user could self-administer the task. Potential limits to scalability include the computational resources required to conduct the analysis, and the need for transcription of video responses, though it is conceivable that advances in automated speech recognition could overcome this limit in the near future.

Beyond these, results indicate that the emoANC measure better exemplifies the seven desiderata than emoAAL. In particular, while the emoAAL measure shows group differences, it does not show convergent validity with the Hinting Task. emoANC and emoAAL show group differences between cases and controls, as anticipated on account of the known social cognitive deficits in schizophrenia. Thus, emoANC and emoAAL satisfy the third desideratum: the measure must show known group differences, and only emoANC satisfies the fourth.

In cross-validation experiments, the measures have AUROCs for predicting case/control status of 0.616 and 0.646 for emoAAL and emoANC, respectively. Interestingly, the more granular emoANC method, which maps to 28 emotion-related categories, outperformed the emoAAL method, which reduces these features to a binary (positive/negative) label. This performance boost could be due to the emoANC method better capturing aspects of social cognition such as emotion processing or theory of mind that differentiate patients from healthy controls. Perhaps this is because emoANC is more granular than emoAAL, and better captures the full range of the Cowan and Keltner emotions. Another explanation may be that emoANC is innately a better method for capturing a construct as complex as social cognition, due to it taking into account the range of variation in emotions in neurotypical responses, and not just adherence to a fixed label.

The fourth desideratum concerns the correlation with validated social cognition tasks, or in this case, the Hinting Task. Only emoANC significantly correlates with the Hinting Task, a validated measure for social cognition, in the patient population. emoAAL does not significantly correlate; however, the relationship of emoAAL to the Hinting Task is in the expected direction and it trends toward significance ($r=0.308$, $p=0.081$). Further, when combining this analysis across patients and controls it significantly correlates with this task ($r=0.311$, $p=0.005$). Nonetheless, a more conservative interpretation based on the patient-only analysis is preferable on account of the ceiling effects of the task. Thus, while emoAAL may show promise for identifying social cognition (and may do so in a larger sample) only emoANC exemplifies the fourth desideratum.

With respect to the fifth desideratum, emoAAL and emoANC are not significantly associated with relevant illness criteria (negative symptoms and functioning) in the patient group. However, in the full sample – i.e. combining patients and controls – emoANC and emoAAL are each significantly associated with functioning and negative symptoms. In the patient sample, association with negative symptoms was stronger than with positive symptoms, which fits with the known correlation between social cognition and negative symptoms, such as that the negative symptom anhedonia is correlated with emotion management (Yolland et al., 2021). Also, emoANC and emoAAL trended towards significance as predictors of functioning with a positive association, such that higher functioning is related to higher emotion processing ability as measured by these metrics. These results align with expectations because social cognition is associated with functional outcomes (Couture et al., 2011; Fett et al., 2011). Nonetheless, by the standard of statistical significance the measures do not associate with symptom and

endophenotype scales related to schizophrenia - the illness in question – and, as such, do not adhere to the fifth desideratum.

With respect to the sixth desideratum, the results fail to discriminate from measures of neurocognition in a few ways. First, the correlation with the Hinting Task is non-significant when controlling for AMNART or LNS values, for emoANC, though these relationships still trend toward significance ($p < .06$). Results are somewhat equivocal; given trend relationships, it seems emoANC captures an aspect of social cognition that is not solely a result of neurocognition, but at the same time, it is not clearly distinct either. Further, both measures are significantly correlated with two of three measures of cognition (AMNART and LNS for both measures). As a result of this, neither measure adheres to the sixth desideratum.

While both measures detected some positive social cognitive change from oxytocin, a potential treatment for social cognition, only emoANC showed significant placebo/oxytocin differences. Thus, emoANC alone satisfies the seventh desideratum.

In summary, emoANC exemplifies most of the desired properties of an automated measure, with the exception that it captures some aspects of neurocognition and does not capture relevant symptoms and endophenotypes in the patient population. emoAAL also exemplifies several of the desired properties, but among other things, it does not converge with the Hinting Task in the patient population and this analysis suggests emoANC is to be preferred.

Other automated measures of social cognition have not been considered in relation to these desiderata. The majority of prior work with LIWC does not evaluate correlation with established measures of social cognition. The work of Buck and colleagues is an exception (Buck et al., 2015b), though that study did not evaluate other psychometric characteristics. The current work is one of the first attempts to develop an automated measure of social cognition in schizophrenia.

Other prior studies also did not evaluate extant assessments according to psychometric desiderata as presented here. For example, there is evidence that the scalable lexical characteristics correlate with anger and functioning (Minor et al., 2015), however, none evaluate LIWC scales for the extent to which they assess a construct distinct from neurocognition (Buck et al., 2015b; Buck & Penn, 2015; Cohen & Minor, 2010; Minor et al., 2015).

Studies of automated estimation of semantic coherence to quantify FTD also primarily focus on group differences (Bedi et al., 2015; Corcoran et al., 2018; Mota et al., 2012). While previous studies showed that both semantic coherence (measured via LSA) and speech graph connectivity are predictive of case status, the general predictive performance of the FTD measures on the data sets used in the studies where the measures were created does not translate fully to the data used in the current experiment. In this context, semantic coherence performed similarly (in terms of group prediction) to the newly-developed social cognitive measures emoAAL and emoANC. However, the speech-graph-based LSC measure was not predictive of case-control status. LSC's lack of predictive power is a surprising finding in that FTD measures should arguably be expected to be more predictive of illness because FTD is a cardinal sign of schizophrenia. The focused, short, and emotional nature of the emotionally evocative task may be relatively poorly suited to elicit incoherence or unconnected speech compared with the open-ended interview questions used in prior FTD work (Elvevåg et al., 2007). Nonetheless, the coherence measure was associated with PANSS's positive symptoms subscale, a validated measure of positive symptom severity. Just as emotional alignment relating to negative symptoms further supports its utility as a measure of cognition, this relationship between a coherence measure and positive symptoms provides further support for its validity as a measure of FTD.

BLERT and ER-40 are validated measurements with psychometric properties that make them suitable for clinical trials (Pinkham et al., 2018). However, BLERT and ER-40 focus on emotion perception only and lack resolution to detect subtle impairments (Pinkham et al., 2018; Vaskinn & Horan, 2020). The broader and more granular nature of emotional alignment indicates that the measures resulting from this work constitute an important first step toward creating a measure of emotion processing deficits with clinical utility.

The emotional alignment measures presented in this chapter can be viewed as a proof-of-concept for a linguistically based, automated, objective, scalable measure for emotion processing deficits. Unlike the studies using computational linguistic measures of FTD, this work is focused on measures of social cognition. Relative to BLERT and ER-40, these measures' strengths are that they have no ceiling effects and capture aspects of both emotional processing and expression.

One limitation of this work is that the findings may not generalize to other tasks. The Alignment Paradigm has thus far been evaluated in the context of studies using emotionally evocative stimuli. It is not necessarily the case that tasks evoking other things (such as intentions) will result in an equally strong alignment measure. Another limitation pertains to the selection of coherence and connectedness variables; the specific measures within these categories used in the studies were selected based on their correlation with an aggregate measure derived from clinician ratings. However, a composite measure combining a range of coherence and connectivity metrics may have performed better. The ratings applied by human raters to categorize the valence of the video stimuli was much less granular than the full range of emotions captured by the GoEmotions network, and with some videos there were disagreements amongst annotators regarding this valence. This lack of inter-rater reliability may adversely affect the emoAAL

method, though it would not present problems for the emoANC method, which does not require ground truth labels.

Furthermore, there are limitations to the computational approaches that underlie emotional alignment. The neural network used to extract emotions was trained on the GoEmotions dataset, and the emotion definitions provided in the publication describing this work are not specified in the detail typically expected from an annotation guideline, consisting of single sentence definitions at times accompanied by an emoji. In validation experiments (See Appendix A) on an independently labeled subset of our dataset, our macro-average F-score was somewhat lower (F1=0.37) than that reported in the paper (F=0.46), indicating that some model accuracy may be lost in the transition between datasets (Demszky et al., 2020), though the weighted macro-average F1 score of approximately 0.5 does suggest performance on our dataset is comparable to that reported. The GoEmotions dataset was derived from the social media site Reddit. Users of this site are primarily White and from the United States; as a result models trained on this corpus may not translate across languages, ethnicities, or geographies. Last, the selection of a Hinting Task for convergent validity analyses may be suboptimal; this measure assesses theory of mind rather than emotion processing. Because it requires empathizing with individuals in videos, the emotional alignment paradigm appears to assess components of both theory of mind and emotion perception; however, is primarily a measure of emotion perception. However, this claim has not been validated by direct comparison to a validated measure of emotion processing. In addition, the ceiling effects of the scales used in the alignment-to-scale comparisons limit the range of performance available as points of comparison.

Future work will further examine the best-performing measure (emoANC) and explore the measure's potential applications. Larger samples are necessary to establish further the

significance of the correlations of the measures with relevant criteria of the illness. Furthermore, validating this measure as an emotion processing measure would require using emotion processing measures (e.g. BLERT or ER-40) and other nonsocial cognitive measures. Further studies involving a more demographically diverse sample would help establish these findings' generalizability as well.

This approach aims to develop a paradigm for automated measures for potential use in clinical trials or clinical monitoring. If validated in future work, these measures have potential to evaluate existing (e.g. social skills training and cognitive remediation; Reeder et al., 2006; Torres et al., 2002) or novel interventions, like intranasal oxytocin, as was examined in this study. With its potential nuanced insights about alignment with normativity of social cognition, these measures could provide several benefits, including evaluation of new treatments, ongoing monitoring in clinical settings, or personalizing treatment plans. Because it can be self-administered, this tool could also help patients track and improve their symptoms over time through social cognitive remediation. Perhaps these quantitative snapshots of alignment could provide direct and helpful feedback that supports recovery even without involvement of clinicians.

In summary, the work presented in this chapter has resulted in two methods of emotional alignment. Of these, the more granular emoANC measure accurately measures social cognitive deficits in schizophrenia. The emoAAL and emoANC measures are amongst the first automated measures of social cognitive ability. Furthermore, emoANC satisfies the majority of the desiderata for an automated measure of social cognition described in Chapter 1; however, it does not meet desideratum 5 (criterion validity) and desideratum 6 (discriminant validity). While further validation is necessary, these measures demonstrate the potential of NLP for assessing

emotion processing deficits. The extension of automated measures to this domain could broadly impact treating, monitoring, and rehabilitating patients with schizophrenia and other mental illnesses characterized by emotional deficits.

Chapter 5: Intentional Alignment

5.1 Introduction

As established in the previous chapters, people with schizophrenia suffer from deficits in social cognition, which are pervasive and have significant consequences. One aspect of social cognition known to be impaired in people with schizophrenia is mentalizing or theory of mind. Mentalizing is the capacity to understand other people's mental states and make judgments about their intentions (Green et al., 2019). Multiple studies have aimed to validate measures of theory of mind, with the most comprehensive arguably being the Social Cognition Psychometric Evaluation (SCOPE) study (Pinkham et al., 2018). The SCOPE study found the Hinting Task to have acceptable psychometric properties as a measure of mentalizing (Pinkham et al., 2018). However, similarly to past tools such as the Sally-Anne test, the Hinting Task requires additional staffing, limits the ceiling on scores for higher-achieving participants, and is not repeatable (Vaskinn & Horan, 2020). Accessible measurement techniques that provide precise, rich quantification of mentalizing behavior and response are needed to enhance our understanding of mentalizing deficits in schizophrenia and characterize the effects of potential treatments.

Prior work in computational linguistics and cognitive psychology has focused on two aspects that relate to mentalizing. Several of the studies discussed in Chapter 2 found an association between a psycholinguistic property and a mentalizing deficit in schizophrenia. In a study examining humoristic content of speech of people with schizophrenia, decreases in metaphorical language and humor were associated with decreases in mentalization (Wyszomirska et al., 2020). Furthermore, Horton and Silverstein found that word memory measures were associated with improvements in mentalizing (Horton & Silverstein, 2008). In another study, Horton found that

sign language acquisition was associated with mentalizing, with quicker acquisition associated with higher estimates of mentalizing performance (Horton, 2010). Buck and colleagues used the Hinting Task as part of a social cognition composite and identified increased prepositions and articles and decreased pronoun use associated with improved social cognition as measured by a cognitive battery that included a measure of mentalization (Buck et al., 2015b). While these findings indicate linguistic manifestations of changes in mentalizing exist, no mentalizing measure based on computational linguistics has yet been psychometrically validated. While each of these studies has noted relationships between language and mentalizing, none have tried to use language specifically to build a measure of mentalizing.

As discussed in Chapters 2 and 4, methods to estimate semantic coherence as an indicator of thought disorder have been the main focus of computational linguistics work related to schizophrenia. However, the assessment of mentalizing using methods of computational linguistics is relatively unexplored. In the current work, I address this gap in the literature by applying the Alignment Paradigm to measure deficits in mentalizing.

I do so by creating two automated measures of mentalizing, using data from a sample of men with schizophrenia and healthy controls, by analyzing how similar an observed intentional response is to an expected response, represented by the topic of the video in question. Here, word embeddings are used to represent these topics, and measure the extent to which they are represented in a transcript. The hypotheses for the work described in this chapter are (H1) these measures will differentiate cases from controls, (H2) these measures will have convergent, criterion, and discriminant validity (as demonstrated through correlation with validated measures of mentalization), and (H3) these measures will detect treatment-induced changes in mentalizing

when oxytocin is administered. This study was motivated by the need for an economical and scalable measure of mentalizing deficits in people with schizophrenia for use in clinical trials. As with emotional alignment, I created two measures. First, I created an intentional alignment with assigned labels (intAAL) measure for mentalizing, as it pertains to the ability to recognize intentions (known as mentalizing). The labels for the Frith-Happe animations were provided to the UCSF research team. The Frith-Happe animations are videos from a task that show animated shapes moving around a screen in a manner that may indicate particular intentions. As such, this task employed a set of stimuli that have previously been used to assess mentalizing (but without the use of automated approaches to evaluate responses). Natural Language Processing (NLP) methods, described in detail in section 2.4, were used to measure the relatedness between transcribed responses to video stimuli and each of the set of assigned labels for the set of video stimuli. I then created a matrix containing the resulting measures of relatedness, such that each row contained the NLP-derived measures of relatedness for a particular video stimulus, and each column indicated the assigned label for that stimulus. I then took the trace (or diagonal) of that matrix to find the sum of the similar aspects (i.e., positive expression to positive stimulus). The main differences from the emoAAL measure described in Chapter 4 are: (1) the video stimuli were designed to indicate intentions (as described above) (2) there are more assigned labels to consider (n=12 rather than n=2); (3) the measures of relatedness apply to the assigned labels directly (so there is no need to collapse the NLP-derived assignments into “positive” and “negative” categories); (4) Responses were recorded as the videos were being viewed; (5) The amount of time given to record a response to the video was 26-48 seconds - the lengths of the videos concerned.

Second, I developed a measure of intentional alignment with neurotypical controls (intANC) for mentalizing. As with emoANC, this measure compares the intentions indicated in a participant's responses (as measured via NLP) to the mean "intention vector" (representing the measured response-to-label relatedness for the set of assigned labels) for responses of healthy controls to a particular stimulus. The resulting measures of similarity are averaged across all transcripts for a particular participant to produce a participant-level measure of alignment. Both intAAL and intANC estimate an alignment score for each individual. These processes provide an answer to the question, "*how aligned is this person's intentional response with typical intentional responses to the same stimuli?".* As with the measures of emotional alignment described in Chapter 4, I hypothesized that participants with schizophrenia would have lower alignment scores, providing the means to distinguish them from neurotypical controls (H1). Further details on the implementation of intAAL and intANC are provided in section 2.3.2 below.

My secondary goal was to evaluate whether intAAL or intANC for *convergent validity*, which I accomplished by assessing the extent to which they align with validated measures of mentalizing. Members of the research team at the University of California, San Francisco (UCSF) manually rated the transcripts according to the coding procedures outlined by Abell, the developer of the video stimuli concerned (Abell et al., 2000). The raters provided *accuracy* and *intentionality* scores for mentalizing for each response to a video stimulus. A description of the scoring for accuracy is shown below in **Table 16**. A description of the scoring for intentionality is shown in **Table 17**.

Table 16: Description of Accuracy Scoring. Excerpt from Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1-16.

0	Bizarre descriptions, plainly wrong descriptions, and responses that focus solely on a minor unimportant aspect of the sequence
1	Partial description of the sequence; description is related to the sequence, but imprecise or incomplete
2	Spot-on description of the story or the actions represented; can be concise just capturing gist, or can be discursive

Table 17: Description of Intentionality Scoring. Excerpt from Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1-16.

0 = no intentionality/action only	Any response comprising a simple action statement with no explicit mention of interaction between the triangles, or mental state/psychological language (e.g. bouncing)
1 = goal-directed, interacting, but not mentalizing	Any response that explicitly mentioned interaction between the triangles, without reference to mental state/psychological language
2 = mentalizing	Descriptions that included explicit psychological or mental state terms (e.g. tricking).

The animations' human ratings provided two mentalizing scores (accuracy and intentionality) for each participant by aggregating all the scores from the non-treatment day. I also used the Hinting Task to assess *convergent validity* (the extent of relatedness with other measures of the same construct). As discussed in Chapter 4, the UCSF team administered the Hinting Task to all participants in this study to examine the relationship between performance on a validated assessment of mentalizing and our automatically generated measure of intentional alignment. In

addition to assessing the correlation between measures of intentional alignment and these established measures of mentalizing, I compared my new metrics to the set of established NLP approaches described in detail in Chapter 4: pronoun use (Buck et al., 2015b), semantic coherence (Elvevåg et al., 2007) and speech graph connectedness (Mota et al., 2017). Amongst the many variants of semantic coherence and speech graph connectedness, I picked the measures that best correlated with clinical assessments of disorganized thinking for participants with schizophrenia represented in our data set. Finally, I assessed whether either of the measures detects known treatment changes. I again analyzed differences in alignment in the context of a one-time dose of oxytocin, which has been found to improve mentalizing in patients with schizophrenia (Woolley et al., 2014).

5.2. Methods

5.2.1 Participants

For these studies, we used the data set described in Chapter 4: 31 men diagnosed with schizophrenia or schizoaffective disorder from outpatient clinics and 51 healthy controls (for further details, see **Table 18**).

5.2.2 Procedures

General procedures for the assessment of participants are described in Chapter 4, section 2.2. In brief, participants (patients with schizophrenia and healthy controls) completed assessments of baseline clinical characteristics. Patients then completed two testing days, separated by at least one week, with drug order (receiving oxytocin or placebo first) randomized between participants. On each testing day, 40 IU oxytocin or saline placebo (Wellspring Pharmacy, Berkeley, CA) was administered intranasally by staff trained on a standard protocol (Guastella et al., 2013). Patients completed the video response task (VRT), beginning ~30 minutes and concluding ~45 minutes

following drug administration. In this case, the VRT was designed to elicit mentalizing. Participants viewed a series of 6 brief videos (26-48 seconds) on a monitor in a private room in the laboratory. These videos show the Frith-Happe Animations (Happe, 1994), which are known to evoke a range of intentional responses. Responses were rated for accuracy and intentionality by two research team members, according to standard protocols (Abell et al., 2000). Two matched forms were created from these ratings such that each form represented the same number of stimuli of each type (i.e., random, goal, mentalizing). Those in the patient group did not see the same videos twice, and the order was randomized between participants. After watching each video, participants were prompted by on-screen instructions to respond to the question, "*What happened in the cartoon?*" Participants were recorded throughout the task, and their audio responses were transcribed. The Frith-Happe animation task was part of a larger testing battery completed on each testing day, including the emotionally evocative videos used to develop emotional alignment metrics.

5.2.3 Measures

5.2.3.1 Baseline clinical characteristics

The baseline clinical characteristics are also presented in Chapter 4. and include the: Positive and Negative Syndrome Scale (PANSS; Kay et al. 1987), the Clinical Assessment Interview for Negative Symptoms (CAINS; Kring et al., 1993), the Hinting Task to assess mentalizing ability, the American National Adult Reading Test (AmNART) to estimate premorbid verbal IQ (Nelson and Willison, 1991), the Letter-Number Sequence (LNS) task to assess working memory (Valentine et al., 2020), the category fluency task to assess verbal fluency, and the Role Functioning Scale (RFS) to assess real-world functioning (RFS; Goodman et al., 1993).

Participants' responses were rated according to the scales for accuracy and intentionality as defined by Abell et al. (2000). This provided a second mentalizing score assessed at two-time points.

5.2.3.2 Intentional alignment

Using transcribed speech from the VRT, I derived two measures of intentional alignment that reflect the similarity between the intentional content of a participant's response to a given video stimulus and the intentional content of a normative response. To do this, I applied neural word embeddings to measure the semantic relatedness between each transcript and a set of intentional labels indicating the responses they are intended to evoke: {"mocking", "coaxing", "chasing", "dancing", "surprising", "seducing", "fighting", "leading"} as well as a set of labels that do not evoke intentions: {"tennis", "star", "billiards", "drifting"}. The process is summarized below and illustrated in **Figure 7** and **Figure 8** for each method.

5.2.3.2.1 Quantifying intentional content of participants' responses

First, to measure semantic relatedness, I used a set of publicly available neural word embeddings, which were trained using the FastText software package crawl-300d-2M.vec.zip. This package provides an implementation of the skipgram-with-negative-sampling algorithm, originally developed by Mikolov and his colleagues (Mikolov et al., 2013), which was used to train the word vectors. The package also includes capabilities for embedding "subwords" (sequences of characters within words). However, these were not used to generate the vectors employed in my experiments. These vectors are the input embeddings of a neural network that was trained to predict words that occur in proximity to an observed word in a large corpus of unannotated text from Wikipedia and the Common Crawl corpus. We will refer to the resulting

neural word embeddings as semantic vectors because proximity between these vectors reflects semantic relatedness. For each transcript (i.e., a given participant's response to one video stimulus), I used the semantic vectors to generate document vectors (one vector representing each response) for each response to a stimulus. Document vectors were constructed using the semvecpy Python package as the L2-normalized vector sum of the semantic vectors for each word in a document, weighted by the inverse document frequency (Jones, 1973) of the word concerned. I then generated a measure of relatedness by taking the document vector's cosine similarity to the stimulus's word vector. I used the resulting matrix of similarities (one row per transcript, one column per label) as the basis for my ANC and AAL methods.

5.2.3.2.2 Estimating intentional alignment with neurotypical controls (intANC) for patients

Second, I constructed a high-dimensional vector representation of these similarities (12 intentions) by concatenating vectors of relatedness measures derived from the semantic vectors for each label:

$$\vec{intention}_{i,j} = [\, cosine(document,\ mocking), cosine(document,\ coaxing),\ \dots\ , cosine(document,\ surprising)\,]$$

Where intention(i,j) is the vector representing the intentions indicated by participant *i*'s response to stimulus *j*.

Third, I constructed another vector representing the normative response across all control participants to stimulus *j* as the average (centroid) of the vectors from set *C,* the sample of control participants.

$$\vec{normativeresponse}_j = \frac{1}{N_C} \sum_{i \in C} \vec{intention}_{i,j}$$

Where N_C is the number of control participants.

Fourth, I quantified a given patient participant *i*'s intentional alignment with the normative response to stimulus *j* using the cosine similarity metric. That is, a patient's intentional alignment with neurotypical controls is the cosine of the angle between their document vector and the centroid vector for each stimulus, varying between 0 for complete orthogonality and 1 for complete alignment (these measurements did not result in any negative cosine values):

$$intANC_{i,j} = \frac{\vec{intention}_{i,j} \bullet \vec{normativeresponse}_j}{|\vec{intention}_{i,j}| \, |\vec{normativeresponse}_j|} = cos(\theta_{i,j})$$

As discussed in Chapter 4, this approach disregards the length of the vectors and focuses exclusively on their directionality. Here I am most interested in the distribution rather than the quantity of intentional expression. Consequently, my metric will judge two responses to be more similar if they focus on the same subset of intentions. Expressing other intentions beyond this subset will decrease similarity estimates.

Finally, I computed a mean intentional alignment score for each patient *i* from their responses to the full set of video stimuli viewed that day, where M is the total quantity of stimuli:

$$averageAlignment_i = \frac{1}{M} \sum_{j=1}^{M} intANC(i,j)$$

Each patient completed the animated shapes task twice and thus had two scores, corresponding to the placebo and oxytocin administration days.

5.2.3.2.3 Estimating intentional alignment for control participants

I estimated intentional alignment with neurotypical controls (intANC) for each control participant via leave-one-out cross-validation. Specifically, I created a 12-dimensional vector for each transcript (as for the patient group), removed the control participant from the dataset, and generated a new vector representing the normative response from the remaining controls as the vector average, or centroid, of their 12-dimensional response vectors. I computed the cosine of the angle between the control participant's vector and the centroid, performing this procedure for each held-out control participant. As for the patients, I computed a mean score for all stimuli viewed on the testing day; each control participant completed the VRT once and thus had one score.

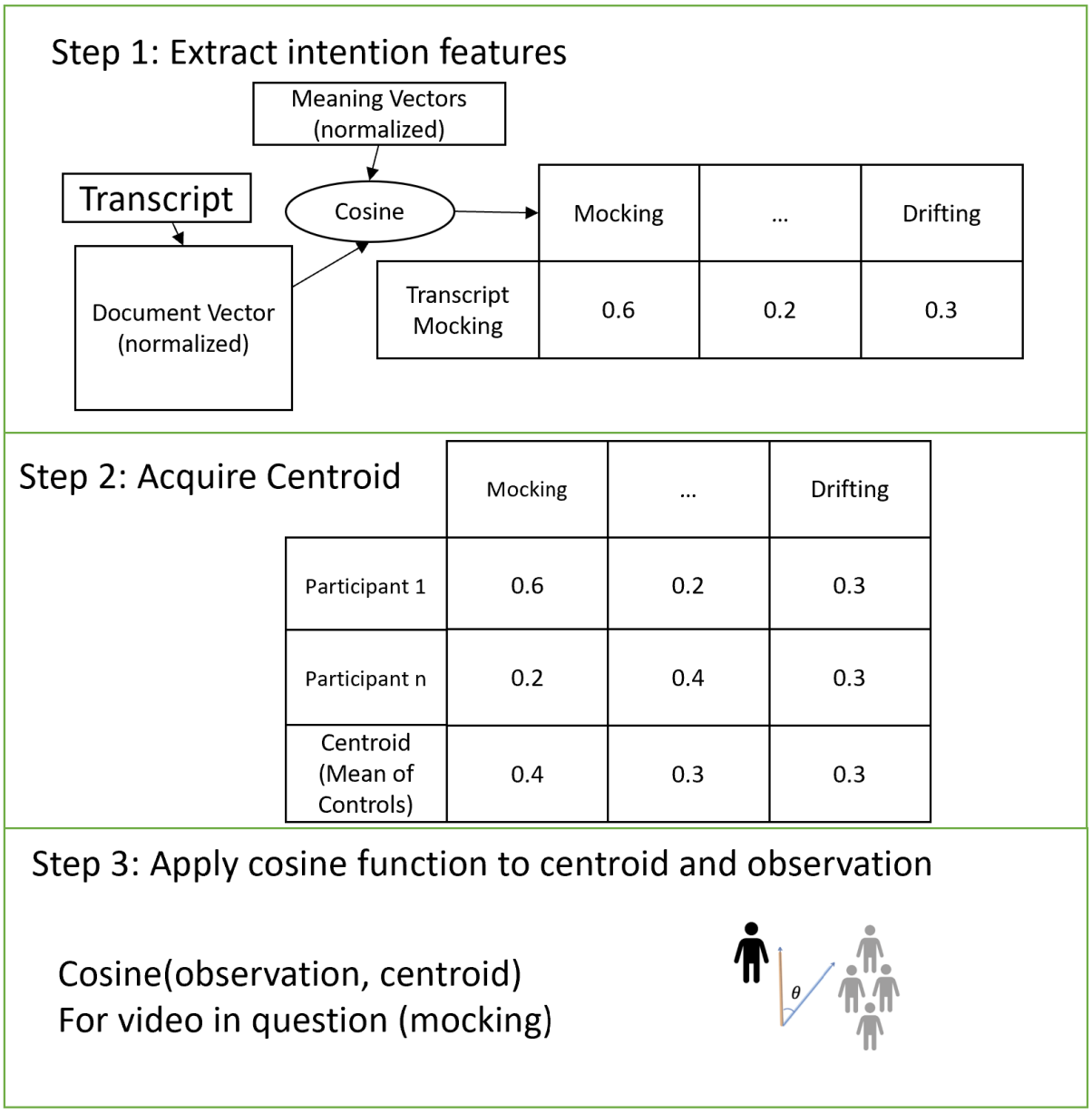
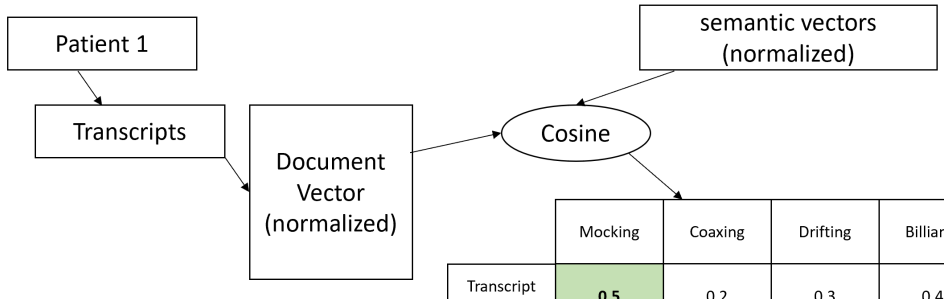


Figure 7. intANC estimation procedure. intANC is the cosine similarity between the document vector for the participant and the normative vector, which is the mean of the intention vectors in the remaining healthy controls. Below is an example with the mocking stimulus.

5.2.3.1 Intentional Alignment with Assigned Labels (intAAL)

I computed a measure of intentional alignment with assigned labels by considering the extent to which intention scores for a transcript—provided by the cosine similarity between the document vector and the ground truth word vector (e.g. the word vector for "mocking") of the video being responded to. I applied my methods to participant responses to the question 'what happened in the cartoon?' **Figure 8** illustrates the estimation procedure for the mocking video stimulus (first row of the matrix in the bottom right of the figure). For each transcript for an individual concerning this stimulus, a measure of the relatedness between each document vector and each stimulus (including mocking) was estimated using cosine similarity. Intentional alignment is estimated by summing along the diagonal of this matrix (the trace), which quantifies the intentions extracted in response to each video.



	Mocking	Coaxing	Drifting	Billiards	Dancing	Chasing
Transcript Mocking	**0.5**	0.2	0.3	0.4	0.1	0.6
Transcript Coaxing	0.2	**0.4**	0.3	0.3	0.4	0.1
Transcript Drifting	0.1	0.4	**0.6**	0.3	0.2	0.4
Transcript Billiards	0.2	0.15	0.2	**0.3**	0.4	0.5
Transcript Dancing	0.3	0.5	0.5	0.35	**0.6**	0.6
Transcript Chasing	0.4	0.3	0.4	0.3	0.5	**0.6**

intAAL =mean(0.5+0.4+0.6+0.3+0.6+0.6)=**0.583**

Figure 8: Procedure to estimate the intentional alignment matrix for patient 1

5.2.3.1.2 Estimating intentional alignment with assigned labels (intAAL)

To estimate intAAL from each run of the experimental task, I first added the measured semantic similarities for the set of intentions composed of the video categories. Second, I normalized the document vectors using the L1-norm (i.e., dividing each value by the sum of the individual values) to make the summation of the document vector for each transcript equal to 1. Third, I created a 6x6 matrix for each participant, with each video stimulus represented as a row and each intention type for the videos viewed by this participant represented as a column. This matrix was constructed for each participant, encapsulating the relatedness between each of their responses and each of the assigned labels across all their transcripts. Fourth, I obtained the intentional alignment metric, ranging from 0 to 1, by taking the trace (the sum of the diagonal) of the 6x6 matrix and then dividing this by 6.

5.2.3.3 Semantic coherence

As described in detail in Chapter 4 I measured the semantic coherence of each transcript using the Comprehensive Coherence Calculator (https://github.com/LinguisticAnomalies/Coherence) package (Xu et al. 2022). Of the 14 different coherence measures calculated by this package, I used sentidfseq. This sentence-level measure reflects the similarity between vector representations of sequential sentences generated as sums of word vectors, weighted by inverse document frequency, which had the highest correlation with the sum of PANSS ratings of "Incoherent Speech" and "Conceptual Disorganization" (measure: "sentidfseq"; Spearman rho(32) =-0.419, p=0.008). I computed the minimum coherence for each transcript across all pairs of sequential sentences. I then averaged these across all transcripts to obtain a single coherence score for each participant for each testing day.

5.2.3.4 Speech graph connectedness

As described in detail in Chapter 4, I also converted each transcript into a graph using the SpeechGraphs software package (https://www.neuro.ufrn.br/softwares/speechgraphs see Mota et al., 2012); however, in this experiment, we used LSC across the full transcript, without implementing step size or bin size procedures. Amongst the measures computed by this package, I selected the largest strongly connected component (LSC), the measure with the highest correlation with clinical ratings of incoherent speech and conceptual disorganization. I calculated an average LSC value for each participant for each testing day.

5.2.3.5 Pronoun frequency

As described in detail in Chapter 4, I used Linguistic Inquiry and Word Count (Pennebaker et al., 2015) to measure pronoun use, a lexical indicator shown to be correlated with social cognition in previous work (Buck et al., 2015b). I computed a mean pronoun frequency score for each participant for each testing day.

5.2.4 Statistical Analyses

I conducted analyses in R (version 4.0.3; R Core Team 2022) and Python (version 3.6, Van Rossum, G., & Drake, F. L., 2009). I removed any participants who did not complete both the emotionally evocative video response task and the animated shapes task. Thus, I joined the data set of 84 participants used in chapter 4 with data from the remaining participants (for the IA task) to acquire a data set of 82 participants. One patient was removed from the oxytocin analysis due to incomplete data, leaving 30 patients for that analysis. I used unpaired t-tests to evaluate group differences in intentional alignment and the existing NLP measures. I also compared each NLP measure's performance in distinguishing patients and controls using the area under the receiver

operating characteristic (AUROC) curves estimated using leave-one-out cross-validation. I used Pearson correlations to assess relationships between intentional alignment, clinical characteristics, and existing NLP measures. To assess oxytocin's effects on performance, I used paired t-tests and explored potential moderators using correlation analyses.

5.3. Results

5.3.1 Sample characteristics

On average, patients were older than controls and had fewer years of education (**Table 18**). Note that I did not match groups on education, given that decreased educational attainment is often a consequence of schizophrenia, and matching may obscure group differences and generate misleading results (Resnick, 1992).

Table 18. Sample clinical characteristics. PANSS-positive = Positive and Negative Syndrome Scale items reflect positive symptoms. CAINS-EXP and CAINS-MAP = Clinical Assessment Interview for Negative Symptoms Expressivity Subscale and Motivation and Pleasure Subscale, respectively. CPZ = Chlorpromazine. AmNART = American National Adult Reading Test. RFS = Role Functioning Scale.

	Patients (n=31)	**Controls (n=51)**	**Patients vs. Controls**
	Mean (SD)	Mean (SD)	
Age	34.87 (12.51)	29.08 (8.69)	p=0.028, d=0.563
Education years	14.34 (2.75)	15.33 (1.53)	p=0.013, d=0.610
PANSS-positive	10.06 (3.84)	-	-
CAINS	16.03 (8.67)	5.88 (3.70)	p<0.001, d=1.674
CAINS-EXP	2.45 (2.75)	0.37 (0.80)	P<0.001, d=1.154
CAINS-MAP	12.29 (6.67)	5.41 (3.47)	p<0.001, d=1.397
CPZ equivalents	196.25 (192.42)	-	-
AmNART	30.29 (7.29)	34.61 (5.38)	p=0.006, d=0.670
Accuracy	2.34 (0.806)	3.13 (0.443)	p<0.001, d=1.305
Intentionality	1.5 (0.453)	1.65 (0.312)	p=0.011, d=0.396
Hinting Task	11.32 (4.67)	14.88 (2.80)	p<0.001, d=0.977
RFS	19.87 (5.30)	25.51 (2.13)	p<0.001, d=1.542

5.3.2 Group differences in intentional alignment

Consistent with hypothesis H1, I found that patients showed impaired intentional alignment with neurotypical controls (intANC) (M=0.992, SD=0.003) compared to controls (M=0.995, SD=0.002), t(45.693)=-4.957, p<0.001; Cohen's d=1.248). However, this finding did not hold for intAAL. Though they had lower values on average, patients did not meet the threshold for statistical significance for impaired intentional alignment with assigned labels (M=0.294, SD=0.028) compared to controls (M=0.304, SD=0.026), t(59.714)=-1.508, p=0.137, d=0.350. As expected based on prior literature, the speech graphs of patients were also significantly less connected (M=39.279, SD=8.46) than those of controls (M=48.591, SD=9.62), t(69.838)=-4.587, p<0.001, *d*=1.012). Contrary to the analysis of responses to evocative videos in Chapter 4, I did not find significantly less coherent speech among patients using the sentidfseq measure (M=0.502, SD=0.109) relative to controls (M=0.517, SD=0.089), t(54.091)=-0.665, p=0.5086; d=0.159. I also did not find a significant difference in pronoun usage among patients (M=7.236, SD=3.47) relative to controls (M=7.822, SD=2.85), t(54.205)=-0.792;d=0.189. See **Figure 9**.

5.3.3. Relationships between intentional alignment and clinical characteristics

I also examined correlations between intentional alignment and patients' baseline clinical characteristics (see **Table 19**). Consistent with hypothesis H2, I found a significant association between accuracy and intentionality as derived from human ratings for the Frith-Happe Animations for both intANC and intAAL. Coherence, connectedness, and pronoun use were not significantly correlated with accuracy or intentionality scores. In terms of the Hinting Task, there were no correlations between intANC (r=0.337, p=0.064) or intAAL (r=0.143, p=0.443).

Since intANC is not correlated with neurocognition (unlike emoANC), no partial correlations were taken for intANC. Since intAAL is correlated with LNS and AMNART, I also examined partial correlations while controlling for LNS and AMNART scores for intAAL. The association between intAAL and accuracy and intentionality performance persisted and maintained statistical significance for accuracy and intentionality with both covariates (LNS and AMNART).

Neither intAAL nor intANC was significantly correlated with total negative symptoms, either subscale of negative symptoms, or functioning in the patient population.

Table 19. Correlations with intentional alignment with neurotypical controls in the patient group.

Boldface indicates statistical significance.

	intANC		**intAAL**	
	Pearson's *r*	***p*-value**	**Pearson's *r***	***p*-value**
Age	-0.025	0.895	-0.103	0.580
Education years	0.050	0.789	0.026	0.891
Accuracy	**0.425**	**0.017**	**0.584**	**0.0006**
Intentionality	**0.495**	**0.005**	**0.456**	**0.001**
Hinting Task	0.143	0.443	0.337	0.064
PANSS-E	-0.212	0.251	-0.153	0.411
CAINS	-0.048	0.796	-0.116	0.535
CAINS-EXP	-0.203	0.275	-0.141	0.456
CAINS-MAP	0.048	0.796	-0.060	0.750
CPZ equivalents	-0.218	0.247	-0.163	0.381
AmNART	-0.022	0.905	**0.359**	**0.047**
LNS	-0.043	0.820	**0.379**	**0.035**
Category fluency	0.149	0.425	0.128	0.493
RFS	0.006	0.974	0.216	0.244

5.3.4. Relationships between intentional alignment and existing NLP measures

I explored whether each measure was associated with existing NLP measures to characterize the intentional alignment measures further. I fit a linear regression model to each pair of measures, with group status as a covariate. In the whole sample, intentional alignment was not significantly associated with coherence (beta=-0.003, p=0.283), speech graph connectedness (beta=5.287e-05, p=0.0625), or pronoun use (beta=-2.626e-04, p=0.975). Examining the same relationships with correlations that do not account for group status within the patient group only, these relationships shifted as follows: coherence (r=-0.170, p=0.360), speech graph connectedness (r=0.230, p=0.101), and pronoun use (r=0.071, p=0.706).

Across the entire sample in the linear regression model with case status as a covariate, intAAL was significantly associated with coherence (beta=-0.108, p<0.001), speech graph connectedness (beta=0.001, p<0.001), and pronoun use (beta=0.213, p=0.0251). intAAL was also significantly associated with intANC (beta=3.761, p=0.002). Within the patient group, intAAL was not significantly associated with existing NLP methods: coherence (r=-0.326, p=0.073), speech graph connectedness (r=0.332, p=0.068), and pronoun use (r=0.126, p=0.498).

I also explored how intentional alignment predicted participants' group status (patient vs. control) relative to the other NLP measures given previous work using coherence (Elvevåg et al., 2007), connectedness (Mota et al., 2017), and lexical characteristics (Buck et al., 2015b) to identify schizophrenia. IntANC as an individual predictor yielded an AUROC of 0.734 (n=82). intANC's AUROC approached that of speech graph connectedness, which yielded an AUROC of 0.751. intAAL as an individual predictor yielded an AUROC of 0.556 (n=82). intAAL performs similarly to the remaining NLP measures (coherence=0.525; pronoun use=0.522). Combining any two measures did not meaningfully improve prediction. Combining intANC with

the best-performing measure (speech graph connectedness) in a logistic regression model resulted in an AUROC of 0.751, lower than speech graph connectedness alone. See **Figure 9**.

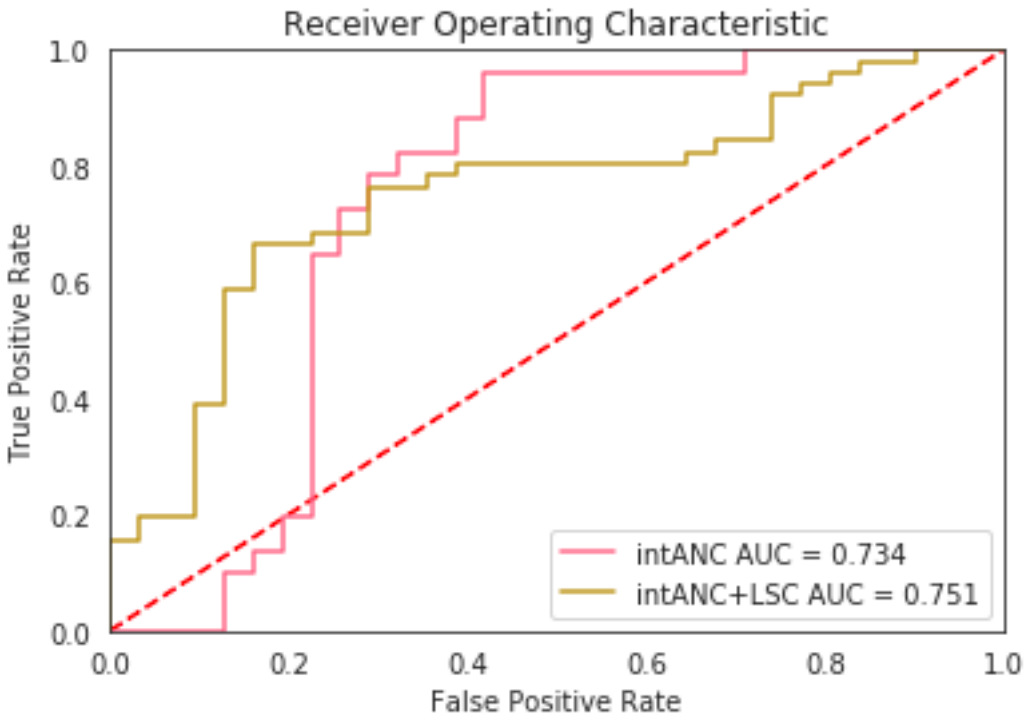


Figure 9. ROC curve and AUCs for group prediction using intANC as an individual predictor and combined with the top-performing NLP measure (LSC) in a logistic regression model.

5.3.5 Oxytocin effects on intentional alignment among patients

Inconsistent with hypothesis H3, oxytocin administration was not associated with better intANC among patients (M=0.991, SD=0.0031) relative to placebo (M=0.992, SD=0.0029), t(29)=-1.101, p=0.320, d=0.139. Also inconsistent with this hypothesis, oxytocin administration was not associated with better intAAL among patients (M=0.293, SD=0.0266) relative to placebo (M=0.290, SD=0.0255), t(29)=-0.643, p=0.523, d=0.110. Oxytocin administration was also not

associated with better accuracy (p=0.918) or intentionality (p=0.452) with assigned labels among patients relative to placebo. Oxytocin was associated with increased coherence (delta coherence M=0.031, SD=0.079); t(29)=2.156, p=0.039, d=0.251. I found no significant effect of oxytocin on pronoun usage or speech graph connectedness.

Figure 10. Associations between oxytocin administration and automatically computed language measures. Performance of the control group is shown for reference. A. intANC. B. intAAL. C. Coherence.

Figure 10a. Association between oxytocin administration and intANC

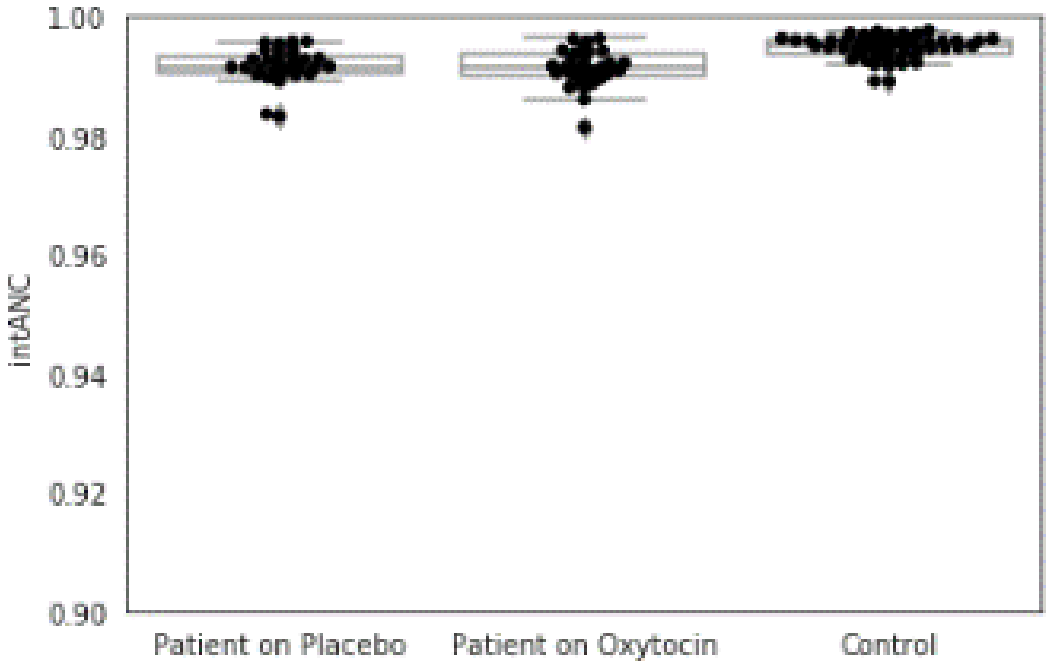


Figure 10b. Association between oxytocin administration and intAAL

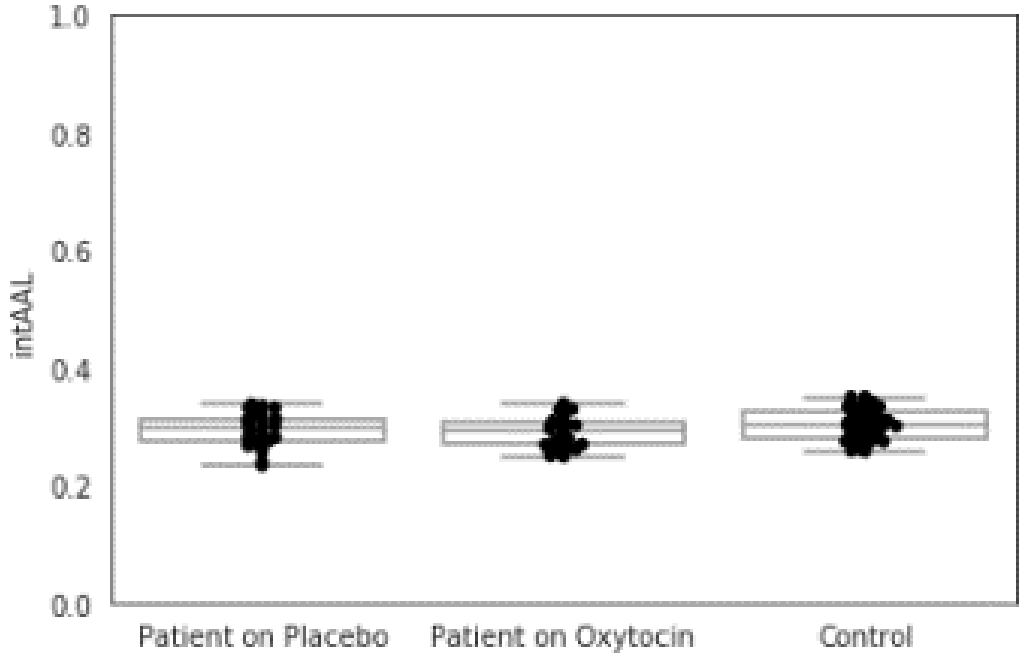
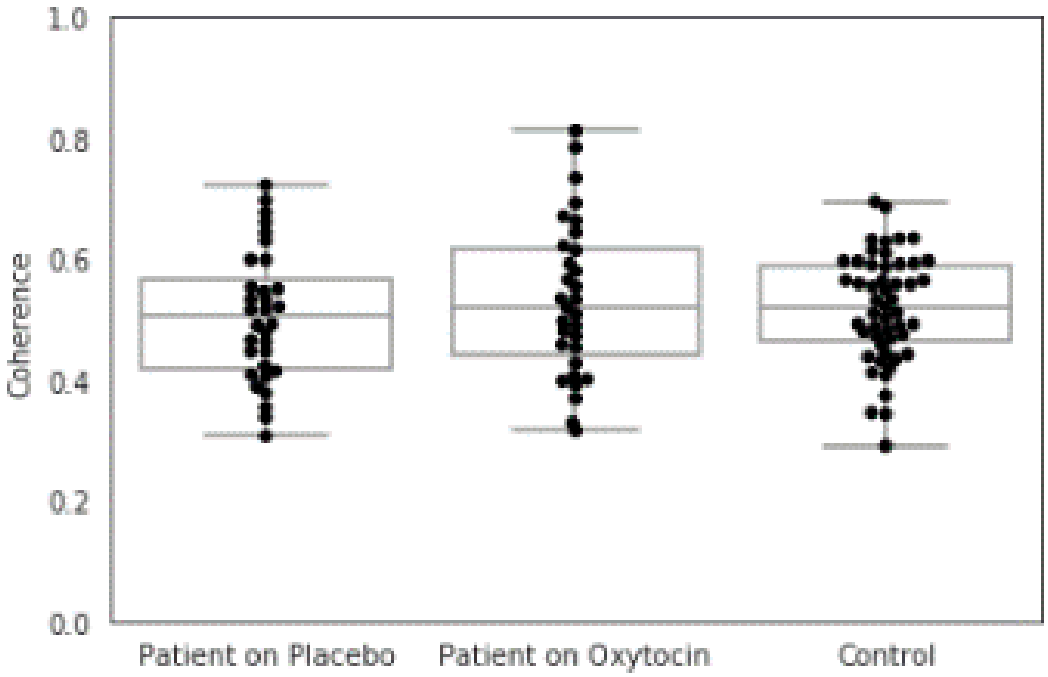


Figure 10c. Association between oxytocin administration and coherence



5.4. Discussion

Table 20: IntANC and IntAAL in relation to desiderata for a social cognition measure

Desideratum (hypothesis)	**IntANC**	**IntAAL**
1. Automated	√	√
2. Scalable	√	√
3. Case/control differentiation (H1)	√	x
4. Agreement with validated measures of social cognition (H2)	√	√
5. Correlation with symptom/endophenotype scales	x (0/2)	x (0/2)
5a. Negative Symptoms	x	x
5b. Functioning	x	x
6. Distinct from neurocognition	√ (3/3)	x (1/3)
6a. AMNART	√	x
6b. LNS	√	x
6c. Category Fluency	√	√
7. Detects oxytocin effects (H3)	x	x

As with emotional alignment, both measures satisfy the first two desiderata (**Table 20**) by design. They are objectively derived using automated methods in a reproducible manner. However, it should be noted that participants may approach the task differently each time, so test-retest reliability remains in question. They are both also scalable due to their computational nature. Also consistent with emotional alignment, potential limits to scalability include the computational resources required to conduct the analysis (e.g. space to host the set of neural word embeddings) and the need for transcription (until accurate automated transcription is readily available).

While many of the checkmarks in **Table 20** are assigned based on statistical significance, given the small sample size it is worth noting that in many cases the intANC and intAAL measures either show or trend in the expected direction for evaluations related to the remaining desiderata I have defined for a measure of mentalizing. Thus, it may be the case that upon evaluation with a larger sample, more of these desiderata would be met. However, this cannot be established based on the current results. With respect to the third desideratum, intANC shows group differences at a significant level. IntANC is also the best predictor of all of the alignment measures (including emotional alignment) for identifying group differences. In contrast, intAAL lacks the expected predictive power in discriminating cases from controls. Thus, only intANC exemplifies the third desideratum (detecting differences between groups). Of note, group differences are also not established by the Frith-Happe Animations manual rating scales for accuracy or intentionality. These scores were also poor predictors of patient versus control status (AUROC=0.55). This underscores the need for new methodologies to assess mentalization deficits in this population.

Both intANC and intAAL readily satisfy the fourth desideratum, showing significant correlations in the patient population with human judgment for accuracy and intentionality on the Frith-Happe Animations. The magnitudes of these correlations are surprising, with the *lowest* being 0.42 (accuracy to intANC), considering the small size of the patient population. These results can be viewed as strong support for the validity of these automated metrics of intentionality for this task. The two measures should intuitively correlate as they are measures of performance of the same task. In contrast, the Hinting Task and the Frith-Happe animation task are both intended to measure mentalization. However, these tasks and the way in which they are scored are different.

It is not surprising that correlations with scores derived from the same task (Frith-Happe) are stronger than those derived from a different one (Hinting).

The combination of the trend toward correlation with the Hinting Task and the fact that intAAL shows a higher correlation with the human-judged accuracy than intANC indicate that intAAL may be a better indicator of mentalizing. However, intANC appears to represent an aspect of the disorder that is different in neurotypical controls and patients and is also more strongly correlated ($r=0.495$ vs. 0.456 for intAAL) with human judgment of intentionality (rather than descriptive accuracy) for this video task. While this picture does not suggest a clear advantage for either method, these findings provide evidence that both intANC and intAAL measurements reflect aspects of mentalization and suggest that they may provide complementary approaches to measuring it. However, it appears the ANC methods again have a slight advantage with respect to the number of the proposed desiderata that it exemplifies (see **Table 20**).

While intANC and intAAL provide objective, scalable measures for mentalization-based on the Frith-Happe Animations, the remaining desiderata related to symptoms, functioning, and existing measures of social cognition were not satisfied. Neither intANC nor intAAL is associated with measures of negative symptoms (CAINS) or functioning (RFS) in the patient population at a significant level. However, this is also true for human ratings of task performance. Neither accuracy nor intentionality is associated with measures of functioning in the patient population at a significant level. In contrast, Hinting Task scores are significantly associated with negative symptoms and functioning, consistent with literature indicating that functional outcomes are linked to social cognitive deficits (Couture et al., 2006; Fett et al., 2011). Thus, it appears that the novel alignment measures developed in this work correspond with manual evaluations of responses to the Frith-Happe animations. However, assessments of responses to the Frith-Happe

Animation appear to capture different aspects of mentalizing than the Hinting Task, aspects that may be less critical to the ability to function in society.

With respect to the sixth desideratum, I found intANC to discriminate from all aspects of neurocognition in the patient population. IntAAL was significantly associated with the letter-number sequence task and verbal IQ in the patient population. Thus, intANC appears to possess better discriminant validity than intAAL. Only intANC satisfies the sixth desideratum by discriminating from neurocognition.

Furthermore, neither measure changed detectably with oxytocin administration. The lack of change in intentional alignment with oxytocin administration is consistent with the reference scores on this task. Neither accuracy nor intentionality scores were significantly different with the administration of oxytocin. Unfortunately, the Hinting Task was only taken at baseline, so no alternative method of establishing oxytocin effects is available in this data set. By the measures available, the data suggest a negative result for the effect of oxytocin on mentalizing.

Finally, intANC and intAAL added value to emoANC as a significant predictor of human ratings of accuracy, intentionality, and the Hinting Task. Furthermore, intANC significantly added to the predictive value of emoANC in a logistic regression model. This significance as an additive measure was unexpected because both measures (mentalizing and emotion processing) are on the same principal component for social cognition (Browne et al., 2016). However, the measures are not significantly correlated with emoANC and appear to capture different social cognition aspects.

Results indicate that this work has resulted in two novel automated measures that correlate well with human ratings of the accuracy and intentionality of responses to the Frith-Happe

Animations. However, the accuracy and intentionality scores for the Frith-Happe Animations in this sample do not correlate with the Hinting Task, negative symptoms, and functioning, nor do they show differences with oxytocin.

Only three studies identified during the scoping review (Chapter 2) directly involved both mentalization and a psycholinguistic property. Scores on the Hinting Task have been associated with word memory and sign language acquisition, with increased scores indicating increased word memory or an increase in speed of acquisition (Horton, 2010; Horton & Silverstein, 2008). However, this work is focused on the measurement of the correlation between neurocognition and social cognition rather than establishing linguistic indicators for mentalization. Buck and colleagues developed four linguistic indicators of social cognition, which were compared to a composite measure that included Hinting Task scores (Buck et al., 2015b). However, the measure selected from this work, pronoun use, did not exhibit the desired psychometric properties for a linguistic indicator of mentalization. This suggests that our automated intentional alignment measures of mentalization represent an advance over methods developed in previously published work; however, there are many limitations to this analysis and it should be viewed as a proof-of-concept.

Prior work on automated measures of Formal Thought Disorder (FTD) primarily focused on group differences. While measures of semantic coherence and speech graph connectivity have been shown to predict case status in previous studies (Bedi et al., 2015; Corcoran et al., 2018), in the context of the current data, only speech graphs were significantly different between cases and controls. Significant differences in semantic coherence, which were apparent in the context of responses to emotionally evocative videos (see Chapter 4), were not detected in the context of this task. This indicates that differences with automated measures of semantic coherence are

task-dependent, as suggested by the seminal work of Elvevåg et al. (2017). However, the estimated coherence measure from clinical interviews associated with scores on a scale that measures the intensity of positive symptoms for schizophrenia ($r=-0.422$, $p=0.018$), indicating that coherence, as measured in the context of these participants, remains associated with positive symptoms.

As noted in Chapter 4, the Hinting Task is a psychometrically validated measurement, which in previous work has been shown to have psychometric properties that make it acceptable for use in clinical trials (Pinkham, 2018). While this is the only measure of mentalizing recommended by the authors of the final SCOPE study (Pinkham et al., 2018), it is not repeatable, and requires assessment by a trained administrator in a laboratory setting.

Thus, it follows that the intentional alignment measures should be considered as the first proof of concept for an NLP measure of mentalization. Relative to the Hinting Task and the Frith-Happe Animations ratings, the strengths of the intentional alignment measures are that they appear to lack ceiling effects and are automated and hence readily scalable. Furthermore, their objectivity addresses the problem of inter-rater reliability, which has been a concern with human ratings of these tasks (Pinkham et al., 2018).

This study has a number of limitations. As with emotional alignment, the findings may not generalize to other tasks, and their extension to other tasks would require both the definition of a set of assigned labels and a way to relate responses to them. The Alignment Paradigm requires an adequately sized set of effectively evocative stimuli, and it is unlikely that intentional alignment would perform well without stimuli designed to evoke mentalization. Similarly, it seems that measures of coherence are task-sensitive. It is clear from the results in Chapter 4 that the negative finding that the measure of semantic coherence used in this study (sentidfseq) did

not discriminate well between cases and controls is inconsistent with its performance on this task in the context of emotionally evocative video stimuli. When comparing analyses of responses to these two types of evocative stimulus, there are vast differences between the performance of the baseline methods (coherence, connectivity, and pronoun use) across stimuli. No baseline method performs well in the context of responses to both of the emotionally and intentionally evocative data sets.

There are also limitations related to challenges of finding a well-established and valid mentalizing measure to use as a comparator. Both the Frith-Happe Animations ratings and the Hinting Task have ceiling effects, lack repeatability and require assessment by a trained administrator (Pinkham et al., 2018). The two extant measures of mentalizing used as comparators here are different from one another (e.g., the Hinting Task does not involve video stimuli) as well. Intentional alignment may capture a subcomponent of mentalizing better captured by one of these (or different) tasks. Additionally, there are several limitations to the computational approach. The neural word embeddings used to relate transcripts to intentions were trained on the Wikipedia and Common Crawl corpus. The language in these corpora may differ from that of typical spoken language. For example, Wikipedia contains edited reference material. Furthermore, there may be measurement difficulties related to language presented by the use of labels. Specifically, video names (e.g. “mocking”) may misleadingly represent their content. Using one word as an assigned label is problematic because there may be multiple ways to mentalize intentions exhibited by the same interaction. For example, an interaction might be considered from either the perspective of the red triangle (e.g. “chasing”) or that of the blue triangle (“e.g. fleeing") in a "geometric shapes” video.

The convergent validity of the alignment measures is indicated by their association with the Frith-Happe Animation ratings but not reinforced by the Hinting Task. This may suggest that intentional alignment measures provide an automated way to code the Frith-Happe animation task, but not one with sufficient convergent validity with other measures of mentalizing.

Furthermore, the lack of correlation with functional outcomes and detectable differences with oxytocin also indicate psychometric limitations, in particular related to criterion validity and responsiveness to treatment. Importantly, this concern also applies to the Frith-Happe Animation human ratings, which themselves do not exemplify a number of the desiderata for a desirable metric.

Future work will concern further examining and optimizing the measures. As with emotional alignment, larger samples are needed to establish further the significance of the correlations between these measures and measurements of relevant aspects of SSD. Comparing the measure to other nonsocial and social cognitive measures (such as the Hopkins Verbal Learning Test-Revised and BLERT) is also an important next step. Modifications including different videos, and the use of NLP methods trained on different corpora may also provide opportunities for improvement. Studies involving a more demographically diverse sample are needed to assess the generalizability of the findings of the study. If fully-powered trials indicate an optimized version of the intentional alignment measures is valid, inclusion as outcomes in clinical trials for pharmaceutical and psychological treatments for deficits in mentalizing may be a feasible next step. Further work could include studies that assess the clinical utility of these measures to see if they are reliable monitoring tools.

As with emotional alignment, intentional alignment could provide the basis for a quick and reliable tool for clinicians to monitor symptoms over time; however, this would require the

development of a novel interface. This could strengthen efforts at ongoing monitoring of social cognition in clinical care. Like emoANC, these methods have the potential to help patients track and reduce their symptoms over time through social cognitive remediation. In summary, with further development the methods have the potential for deployment on mobile devices to allow for ambulatory data collection, and for self-administration for feedback at the patient level.

In conclusion, I created two intentional alignment methods that measure mentalization deficits in schizophrenia, or how well a patient is aligned with a normative response. Together with the measures of emotional alignment described in Chapter 4, these measures are – to my knowledge – the first measures of social cognitive ability using NLP methodology. While adaptations and further psychometric examination is necessary, the correlation between these measures and human-assigned scores for the task concerned indicates that mentalizing deficits in schizophrenia can be assessed using NLP. Mapping this new domain could have broad-reaching implications for treating, monitoring, and rehabilitating patients with schizophrenia and other conditions characterized by mentalizing deficits.

Chapter 6: Concluding Remarks

Maybe each human being lives in a unique world, a private world different from those inhabited and experienced by all other humans. . . If reality differs from person to person, can we speak of reality singular, or shouldn't we really be talking about plural realities? And if there are plural realities, are some more true (more real) than others? What about the world of a schizophrenic? Maybe it's as real as our world. Maybe we cannot say that we are in touch with reality and he is not, but should instead say, His reality is so different from ours that he can't explain his to us, and we can't explain ours to him. The problem, then, is that if subjective worlds are experienced too differently, there occurs a breakdown in communication ... and there is the real illness.

— Philip K. Dick

6.1 Innovation and Contributions

6.1.1 The Alignment Paradigm

The Alignment Paradigm is a new automated approach to quantifying the similarity of responses to a defined set of stimuli between two groups. Specifically, I define a set of parameters of interest (the labels or normative responses) and use automated methods to measure the relatedness between them and an individual's responses. While the measurements used in this case were developed using Natural Language Processing (NLP), the paradigm can be applied to any aspect of a response in which interpretable features can be derived using automated methods. As such, the Alignment Paradigm represents a novel, generalizable approach toward the

development of scalable measures of cognitive, emotional and expressive changes related to mental illness.

The specific instantiations of this paradigm developed in this dissertation are also novel in conception, particularly in their application of neural language representations to elicit interpretable features from transcribed responses. I will focus my discussion on the contributions related to the two best-performing methods: emoANC and intANC, though the points presented here apply to emoAAL and intAAL also.

6.1.2 Emotional Alignment with Neurotypical Controls (emoANC)

emoANC is a novel automated method that measures emotional alignment with a normative response. There have been few studies of emotion using natural language processing in schizophrenia. To the best of my knowledge, this work is the first to use neural language models to analyze emotion in the language of people with schizophrenia. With emoANC, this provides an interpretable representation of the emotions expressed in participant transcripts and an objective quantitative measurement of deviation from normative responses.

6.1.3 Intentional Alignment with Neurotypical Controls (intANC)

While there have been efforts to evaluate linguistic associations with mentalizing in schizophrenia using LIWC (Buck et al., 2015b), to the best of my knowledge, this work is the first to apply NLP methods for automated assessment of whether or not an individual accurately recognizes intentions (i.e. is able to mentalize appropriately).

6.2 Implications

6.2.1 Diagnosis

Emotional and intentional alignment may add a valuable component to diagnosis of schizophrenia-spectrum disorders as well as other disorders characterized by social-emotional impairments by objectively measuring abnormalities that are not evaluated quantitatively or with a high degree of granularity for clinical purposes. Results indicate that either the emoANC and intANC methods could be used independently as predictors of case status, as shown in Chapters 4 and 5. Furthermore, emoANC improved the performance of a predictive model of case/control status based on established NLP methods for the assessment of language in schizophrenia. While automated diagnosis of schizophrenia is not the primary motivating use case for this research, these results suggest that emoANC provides additional information of diagnostic utility beyond that captured by previously-validated measures of coherence, connectivity, and pronoun use.

6.2.2 Monitoring

These measures also can potentially improve patient monitoring by psychiatrists and psychologists through repeated assessment, given the availability of an adequate number of video stimuli to prevent administration of the same stimulus to a single participant. Tasks to facilitate monitoring could be performed in the clinic or – leveraging the scalable, automated nature of these metrics – within a digitally-delivered remote or ambulatory care application. Automated detection of a social cognitive deficit could be used to identify patients for cognitive or pharmacological therapies designed to address this specific deficiency and monitor responsiveness to these treatments over time.

6.23 Treatment

The work presented in this dissertation has implications for treating patients with schizophrenia, as it includes an analysis of oxytocin as a potential therapy for social cognitive deficits. Results presented in Chapter 4 suggest that oxytocin improves the emotion processing ability of patients and adds to the evidence that it can be used to treat social cognitive deficits in schizophrenia.

This suggests that alignment-based methods could be used in clinical trials, either to identify individuals likely to benefit from therapies developed to improve social cognition or as an outcome variable of interest. For example, emotional alignment could be used as an outcome measure for emotion processing deficits. Similarly, intentional alignment could be used as a measure of mentalization. Emotional alignment and intentional alignment are scalable, enabling researchers to analyze data captured from patients at home longitudinally to acquire objective estimates of treatment effects over time. Furthermore, the potential utility of these methods is not limited to schizophrenia-spectrum disorders. People with autism spectrum and bipolar disorders can also have social cognitive deficits, suggesting the potential to assess the effects of treatments that target this transdiagnostic construct across different conditions. Future work is needed to test and validate these tools in different clinical populations. Of note, rehabilitation of social cognitive impairments in schizophrenia may already be possible via social skills training (Turner et al., 2018). This suggests a role for automated measures in the context of cognitive remediation interventions, providing training tools with the capacity for immediate feedback. As such, a possible extension of this work would be the development of a scalable, cost-efficient, and readily-deployable application for patients to train to improve their social cognitive deficits related to mentalizing and emotional experience. This application could provide real-time scores of social cognitive alignment to people with schizophrenia. It could also include explainable

artificial intelligence methods (such as Shapley values) to explain the responses to participants. Through iterative feedback on the task, participants could learn to improve their responses to be 'normative' in an effort to reintegrate into society after a psychotic episode.

6.3 Limitations

There are several limitations to this study, some of which I plan to address in future work. First, the methods I have developed to date can only identify *expressed* emotional or intentional experience. However, there may be a mismatch between how an individual feels and how these feelings are expressed in speech (Gard & Kring, 2009). Similarly, for mentalizing, it is possible that an individual has come to a normative assessment of the intended intention without expressing this in language. Extensions of the Alignment Paradigm toward non-verbal features, such as emotions derived from facial expressions (see Zeng et al., 2018), may be able to address this limitation in the context of particular tasks partially. However, further research is required to establish the utility of non-verbal stimuli for alignment-based methods. Second, these measures require emotionally or intentionally evocative videos, and further research is required to establish the extent to which the Alignment Paradigm can be applied to other tasks. Also, further research is needed to assess the utility of other approaches to the selection of assigned labels to serve as interpretable features. The assigned labels for the emotional dataset were based on valence ratings that were not unanimous, and it may be the case that modeling disagreement between annotators would lead to a more robust "emoAAL" metric (in contrast, emoANC uses the set of emotions defined by Cowan and Keltner in place of these valence labels – this additional granularity may explain its improved performance). The assigned labels for the intentional dataset were based on the titles of the videos, which were single words that describe the

intentions each video was intended to elicit, with the implicit assumption that there is only one "correct" way to respond to each video.

Furthermore, neither the neural network used to extract emotions nor the neural word embeddings used to identify intentions were trained on speech samples gathered in a clinical context. Rather, they were trained on annotated text from social media (i.e. Reddit) or text scraped from the internet, respectively. The authors of these language samples may not be representative of a demographically diverse schizophrenic population. For example, the majority of Reddit users (54%) are from the United States (*The Demographics of Reddit*, visited on July 26, 2022). One way to address this in future work is to consider using corpora developed to capture dialectical variants for additional pre-training or, with annotation, for fine-tuning, such as the Corpus Of Regional African American Language (Kendall and Farrington, 2021).

Other limitations concern the established measures of social cognition that were available as points of comparison. The social cognitive validation for both measures had limitations due to the Hinting Task not directly measuring emotion processing and the disagreement between the manual assessments from the Hinting Task and the Frith-Happe ratings that both ostensibly quantify mentalizing ability. Furthermore, all clinical characteristics were assessed with scales that had ceiling effects, so a large portion of the variance may not have been captured. With respect to the population studied, the sample size of participants is likely too small to fully represent the heterogeneity of people with schizophrenia spectrum disorders (including but extending beyond the concern that participants are exclusively male). Also, measures were taken at one or two time points which cannot capture fluctuations on account of the fluid nature of the illnesses concerned. Further evaluations with larger, more diverse samples are required to establish the generalizability of my findings beyond these data.

6.4 Conclusion

In this work, I created four novel measures for measuring social cognition and evaluated them empirically for their ability to distinguish patients from control participants and their convergence with established (but manually rated) measures of social cognition. The results of these evaluations confirm that neural language representations can be used to assess social cognition. In addition, results suggest that some of these methods provide information that is complementary to established NLP-based methods used to quantify formal thought disorder in people with schizophrenia. Future work may include developing alignment methods to measure other aspects of cognition (and potentially other aspects of mental illness in general), integration into the care setting for assessment, training, or remote patient monitoring. Though this is a first effort, and future advances require adaptation and further study, these methods have potential downstream to address a fundamental gap in the assessment of schizophrenia.

References

Abell, F., Happé, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*(1), 1–16.

Aleman, A., Lincoln, T. M., Bruggeman, R., Melle, I., Arends, J., Arango, C., & Knegtering, H. (2017). Treatment of negative symptoms: Where do we stand, and where do we go? *Schizophrenia Research*, *186*, 55–62.

Andreasen. (1986a). *Scale for the assessment of thought, language, and communication (TLC). | Semantic Scholar*.

Andreasen, N. C. (1986b). Scale for the Assessment of Thought, Language, and Communication (TLC). *Schizophrenia Bulletin*, *12*(3), 473–482.

Bang. (2019). *Reduced DNA Methylation of the Oxytocin Receptor Gene Is Associated With Anhedonia-Asociality in Women With Recent-Onset Schizophrenia and Ultra-high Risk for Psychosis | Schizophrenia Bulletin | Oxford Academic*.

Bedi et al. (2015). *Automated analysis of free speech predicts psychosis onset in high-risk youths | npj Schizophrenia*.

Bonfils, K. A., Luther, L., Firmin, R. L., Lysaker, P. H., Minor, K. S., & Salyers, M. P. (2016). Language and hope in schizophrenia-spectrum disorders. *Psychiatry Research*, *245*, 8–14.

Bradley, E. R., Brustkern, J., Coster, L. D., Bos, W. van den, McClure, S. M., Seitz, A., & Woolley, J. D. (2020). Victory is its own reward: Oxytocin increases costly competitive behavior in schizophrenia. *Psychological Medicine*, *50*(4), 674–682.

Browne, J., Penn, D. L., Raykov, T., Pinkham, A. E., Kelsven, S., Buck, B., & Harvey, P. D. (2016). Social cognition in schizophrenia: Factor structure of emotion processing and theory of mind. *Psychiatry Research*, *242*, 150–156.

Buck, B., Minor, K. S., & Lysaker, P. H. (2015a). 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.

Buck, B., Minor, K. S., & Lysaker, P. H. (2015b). Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Comprehensive Psychiatry*, *58*, 138–145.

Buck, B., & Penn, D. L. (2015). Lexical characteristics of emotional narratives in schizophrenia: Relationships with symptoms, functioning, and social cognition. *The Journal of Nervous and Mental Disease*, *203*(9), 702–708.

Cohen, A. S., Alpert, M., Nienow, T. M., Dinzeo, T. J., & Docherty, N. M. (2008). Computerized measurement of negative symptoms in schizophrenia. *Journal of Psychiatric Research*, *42*(10), 827–836.

Cohen, A. S., & Minor, K. S. (2010). Emotional Experience in Patients With Schizophrenia Revisited: Meta-analysis of Laboratory Studies. *Schizophrenia Bulletin*, *36*(1), 143–150.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67–75.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67–75.

Corcoran, C. M., Mittal, V. A., Bearden, C. E., E. Gur, R., Hitczenko, K., Bilgrami, Z., Savic, A., Cecchi, G. A., & Wolff, P. (2020). Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*, *226*, 158–166.

Cotter, J., Bartholomeusz, C., Papas, A., Allott, K., Nelson, B., Yung, A. R., & Thompson, A. (2017). Examining the association between social cognition and functioning in individuals at ultra-high risk for psychosis. *Australian & New Zealand Journal of Psychiatry*, *51*(1), 83–92.

Couture, S. M., Granholm, E. L., & Fish, S. C. (2011). A path model investigation of neurocognition, theory of mind, social competence, negative symptoms and real-world functioning in schizophrenia. *Schizophrenia Research*, *125*(2), 152–160.

Couture, S. M., Penn, D. L., & Roberts, D. L. (2006). The Functional Significance of Social Cognition in Schizophrenia: A Review. *Schizophrenia Bulletin*, *32*(suppl_1), S44–S63.

Covington, M. A. (2012). CPIDR® 5.1 user manual. *Athens, GA: Artificial Intelligence Center, The University of Georgia*.

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382.

Cowen and Keltner (2017). *Self-report captures 27 distinct categories of emotion bridged by continuous gradients | PNAS*.

De Coster, L., Lin, L., Mathalon, D. H., & Woolley, J. D. (2019). Neural and behavioral effects of oxytocin administration during theory of mind in schizophrenia and controls: A randomized control trial. *Neuropsychopharmacology*, *44*(11), 1925–1931.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, *93*(1), 304–316.

Elvevag, Brita. (2017). *Thoughts About Disordered Thinking: Measuring and Quantifying the Laws of Order and Disorder | Schizophrenia Bulletin | Oxford Academic*.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Reznick, J. S. (2007). MacArthur-Bates CDI (Communication Developmental Inventory) Words and Sentences.

Fett, A.-K. J., Viechtbauer, W., Dominguez, M.-G., Penn, D. L., van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 573–588.

Fiszdon, J. M., Fanning, J. R., Johannesen, J. K., & Bell, M. D. (2013). Social cognitive deficits in schizophrenia and their relationship to clinical and functional status. *Psychiatry Research*, *205*(1), 25–29.

Gao. (2016). *Oxytocin, the peptide that bonds the sexes also divides them | PNAS*.
Gard, D. E., & Kring, A. M. (2009). Emotion in the daily lives of schizophrenia patients: Context matters. *Schizophrenia Research*, *115*(2), 379–380.
Goodman, S. H., Sewell, D. R., Cooley, E. L., & Leavitt, N. (1993). Assessing levels of adaptive functioning: the Role Functioning Scale. *Community mental health journal*, *29*(2), 119-131.
Green, M. F., & Horan, W. P. (2010). Social Cognition in Schizophrenia. *Current Directions in Psychological Science*, *19*(4), 243–248.
Green, M. F., Horan, W. P., & Lee, J. (2019). Nonsocial and social cognition in schizophrenia: Current evidence and future directions. *World Psychiatry*, *18*(2), 146–161.
Green MF, Penn DL, Bentall R, et al. Social cognition in schizophrenia: An NIMH workshop on definitions, assessment, and research opportunities. Schizophrenia Bulletin. 2008;34:1211–1220.
Guastella, A. J., Hickie, I. B., McGuinness, M. M., Otis, M., Woods, E. A., Disinger, H. M., Chan, H.-K., Chen, T. F., & Banati, R. B. (2013). Recommendations for the standardisation of oxytocin nasal administration and guidelines for its reporting in human research. *Psychoneuroendocrinology*, *38*(5), 612–625.
Halverson, T. F., Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2022). Brief battery of the Social Cognition Psychometric Evaluation study (BB-SCOPE): Development and validation in schizophrenia spectrum disorders. *Journal of Psychiatric Research*, *150*, 307–316.
Horton, H. K. (2010). Linguistic ability and mental health outcomes among deaf people with schizophrenia. *The Journal of Nervous and Mental Disease*, *198*(9), 634–642.
Horton, H. K., & Silverstein, S. M. (2008). Social cognition as a mediator of cognition and outcome among deaf and hearing people with schizophrenia. *Schizophrenia Research*, *105*(1–3), 125–137.
Jin, H., & Mosweu, I. (2017). The Societal Cost of Schizophrenia: A Systematic Review. *PharmacoEconomics*, *35*(1), 25–42.
Jones, K. S. (1973). Index term weighting. *Information Storage and Retrieval*, *9*(11), 619–633.
Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American Journal of Psychology*, *120*(2), 263–286.
Kane, J. M., & Marder, S. R. (1993). Psychopharmacologic Treatment of Schizophrenia. *Schizophrenia Bulletin*, *19*(2), 287–302.
Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The Positive and Negative Syndrome Scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*, *13*(2), 261–276.
Kring, A. M. (2013). *The Clinical Assessment Interview for Negative Symptoms (CAINS): Final Development and Validation | American Journal of Psychiatry*.
Kucharska-Pietura, K., & Mortimer, A. (2013). Can Antipsychotics Improve Social Cognition in Patients with Schizophrenia? *CNS Drugs*, *27*(5), 335–343.
Leucht, S., Samara, M., Heres, S., Patel, M. X., Woods, S. W., & Davis, J. M. (2014). Dose equivalents for second-generation antipsychotics: the minimum effective dose

method. *Schizophrenia bulletin*, *40*(2), 314-326.

Lieberz. (2020). *Kinetics of oxytocin effects on amygdala and striatal reactivity vary between women and men | Neuropsychopharmacology*.

Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., Pan, X., Rutledge, R. B., & Ma, Y. (2019). Oxytocin modulates social value representations in the amygdala. *Nature Neuroscience*, *22*(4), 633–641.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

MacDonald. (2010). *The Peptide That Binds: A Systematic Review of Oxytocin and... : Harvard Review of Psychiatry*.

Mancuso, F., Horan, W. P., Kern, R. S., & Green, M. F. (2011). Social cognition in psychosis: Multidimensional structure, clinical correlates, and relationship with functional outcome. *Schizophrenia Research*, *125*(2), 143–151.

Michael, Green. (2016). *Impact of Cognitive and Social Cognitive Impairment on Functional Outcomes in Patients With Schizophrenia | Psychiatrist.com*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, *26*.

Minor, K. S., Bonfils, K. A., Luther, L., Firmin, R. L., Kukla, M., MacLain, V. R., Buck, B., Lysaker, P. H., & Salyers, M. P. (2015). Lexical analysis in schizophrenia: How emotion and social word use informs our understanding of clinical presentation. *Journal of Psychiatric Research*, *64*, 74–78.

Moe, A. M., Breitborde, N. J. K., Bourassa, K. J., Gallagher, C. J., Shakeel, M. K., & Docherty, N. M. (2018). Schizophrenia, narrative, and neurocognition: The utility of life-stories in understanding social problem-solving skills. *Psychiatric Rehabilitation Journal*, *41*(2), 83–91.

Mota et al. (2012). *Speech Graphs Provide a Quantitative Measure of Thought Disorder in Psychosis*.

Mota et al. (2017). *Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance | npj Schizophrenia*.

Mota, N. B., Furtado, R., Maia, P. P. C., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific Reports*, *4*(1), 3691.

Nelson, H. E., & Willison, J. (1991). *National adult reading test (NART)* (pp. 1-26). Windsor: Nfer-Nelson.

Oettl, L.-L., Ravi, N., Schneider, M., Scheller, M. F., Schneider, P., Mitre, M., da Silva Gouveia, M., Froemke, R. C., Chao, M. V., Young, W. S., Meyer-Lindenberg, A., Grinevich, V., Shusterman, R., & Kelsch, W. (2016). Oxytocin Enhances Social Recognition by Modulating Cortical Control of Early Olfactory Processing. *Neuron*, *90*(3), 609–621.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports*, *10*(3), 799–812.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. 26.

Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social Cognition Psychometric Evaluation: Results of the Final Validation Study. *Schizophrenia Bulletin*, *44*(4), 737–748.

Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: the word-in-context dataset for

evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Piskulic, D., & Addington, J. (2011). Social cognition and negative symptoms in psychosis. *Psychiatry Research*, *188*(2), 283–285.

Quintana. (2016). *Low dose intranasal oxytocin delivered with Breath Powered device dampens amygdala response to emotional stimuli: A peripheral effect-controlled within-subjects randomized dose-response fMRI trial—ScienceDirect*.

Reeder, C., Smedley, N., Butt, K., Bogner, D., & Wykes, T. (2006). Cognitive predictors of social functioning improvements following cognitive remediation for schizophrenia. *Schizophrenia Bulletin*, *32 Suppl 1*, S123-131.

Resnick, S. M. (1992). Matching for education in studies of schizophrenia. *Archives of General Psychiatry*, *49*(3), 246–246.

Rey, A. (1964). L'examen clinique en psychologie, Paris: Presses Universitaires de France, 1964. *Chemotherapy and objective cognitive functioning*, *95*.

Rezaii et al. (2019). *A machine learning approach to predicting psychosis using semantic density and latent content analysis | npj Schizophrenia*.

Romero-Martínez, Á., Sarrate-Costa, C., & Moya-Albiol, L. (2021). A Systematic Review of the Role of Oxytocin, Cortisol, and Testosterone in Facial Emotional Processing. *Biology*, *10*(12), 1334.

Rubin, L. H., Li, S., Yao, L., Keedy, S. K., Reilly, J. L., Hill, S. K., Bishop, J. R., Sue Carter, C., Pournajafi-Nazarloo, H., Drogos, L. L., Gershon, E., Pearlson, G. D., Tamminga, C. A., Clementz, B. A., Keshavan, M. S., Lui, S., & Sweeney, J. A. (2018). Peripheral oxytocin and vasopressin modulates regional brain activity differently in men and women with schizophrenia. *Schizophrenia Research*, *202*, 173–179.

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A Systematic Review of the Prevalence of Schizophrenia. *PLOS Medicine*, *2*(5), e141.

Salva. (2013). *Deficits in domains of social cognition in schizophrenia: A meta-analysis of the empirical evidence—PubMed*.

Social Cognition Psychometric Evaluation: Results of the Final Validation Study | Schizophrenia Bulletin | Oxford Academic. (n.d.).

Strauss, G. P., Chapman, H. C., Keller, W. R., Koenig, J. I., Gold, J. M., Carpenter, W. T., & Buchanan, R. W. (2019). Endogenous oxytocin levels are associated with impaired social cognition and neurocognition in schizophrenia. *Journal of Psychiatric Research*, *112*, 38–43.

The Demographics of Reddit: Who Uses the Site? (n.d.). Alphr.

Torres, A., Mendez, L. P., Merino, H., & Moran, E. A. (2002). Improving social functioning in schizophrenia by playing the train game. *Psychiatric Services (Washington, D.C.)*, *53*(7), 799–801.

Turner et al. (2018). *Meta-Analysis of Social Skills Training and Related Interventions for Psychosis | Schizophrenia Bulletin | Oxford Academic*.

Valentine, T., Block, C., Eversole, K., Boxley, L., & Dawson, E. (2020). Wechsler Adult Intelligence Scale-IV (WAIS-IV). *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*, 457-463.

Vaskinn, A., & Horan, W. P. (2020). Social Cognition and Schizophrenia: Unresolved Issues and New Challenges in a Maturing Field of Research. *Schizophrenia Bulletin*, *46*(3), 464–470.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. *Autism Research: Official Journal of the International Society for Autism Research*, *4*(2), 149–154.

Woolley, J. D., Chuang, B., Lam, O., Lai, W., O'Donovan, A., Rankin, K. P., Mathalon, D. H., & Vinogradov, S. (2014). Oxytocin administration enhances controlled social cognition in patients with schizophrenia. *Psychoneuroendocrinology*, *47*, 116–125.

Wyszomirska, J., Martyniak, E., & Bąk-Sosnowska, M. (2020). It is no joke. Metaphorical language and sense of humor in schizophrenia. *Psychiatria Polska*, *54*(4), 687–700.

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., & Cohen, T. (2020). *The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder*. PsyArXiv.

Xu, W., Wang, W., Portanova, J., Chander, A., Campbell, A., Pakhomov, S., Ben-Zeev, D., & Cohen, T. (2022). Fully automated detection of formal thought disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). *Journal of Biomedical Informatics*, *126*, 103998.

Yolland, C. O. B., Carruthers, S. P., Toh, W. L., Neill, E., Sumner, P. J., Thomas, E. H. X., Tan, E. J., Gurvich, C., Phillipou, A., Rheenen, T. E. V., & Rossell, S. L. (2021). The Relationship between Negative Symptoms and Both Emotion Management and Non-social Cognition in Schizophrenia Spectrum Disorders. *Journal of the International Neuropsychological Society*, *27*(9), 916–928.

Zeng et al. (2018). *Facial expression recognition via learning deep sparse autoencoders—ScienceDirect*.

Appendix

A. Validating the GoEmotions Neural Network with the Evocative Videos Dataset

GoEmotions is a data set based on Reddit and the model used was fine-tuned to this data set. Thus, it may not accurately represent the data. To evaluate the transferability of the model two authors manually rated the transcripts for emotions.

To evaluate GoEmotions we took a simple 3 step approach.

1.) Compare GoEmotions output with two manual raters on 159 transcripts (20% of the data)
    a. Each rater rated whether the emotion was present
    b. The guide used for these transcripts was the guide in the appendix of Demsky et al. (2020).
2.) Acquire a ground truth emotion score by taking all of the emotions rated present by both raters.
3.) Acquire an F-measure scores for comparison by using the human raters as ground truths and the GoEmotions model as predictors.

The results are shown in **Table A1** below. The table below shows the F1, Recall, Precision, and number of responses for each emotion. A response was identified as having the emotion rated present by both raters. The bottom row is the macro-average of the columns and it illustrates a macro-average F1 score of 0.37.

Table A1: GoEmotions Validation Ratings.

F1	Recall	Precision	Emotion	n
0	0	0	admiration	18
0.760563	0.72973	0.7941176	amusement	34
0.4	0.285714	0.6666667	annoyance	3
0	0	0	caring	2
0.25	0.4	0.1818182	confusion	11
0.428571	0.5	0.375	curiosity	8
0	0	0	desire	2
0.285714	0.166667	1	disapproval	1
0.818182	0.75	0.9	disgust	10
0.769231	0.833333	0.7142857	excitement	7
0.647059	0.846154	0.5238095	fear	21
0	0	0	gratitude	3
0.754098	1	0.6052632	joy	38
0.4	1	0.25	love	4
0.580645	0.5625	0.6	nervousness	15
0.206897	0.176471	0.25	neutral	12
0	0	0	optimism	4
0	0	0	realization	15
0.5	0.5	0.5	relief	2
0.611111	0.916667	0.4583333	sadness	24
0.370604	0.433362	0.3909647		11.7

The analysis resulted in a macro-average F-score of 0.37 when validating GoEmotions. This F-score is below the macro-average F-score of 0.46 from the original work by Demsky et al. One potential reason for the drop in F-score is the nature of the transcripts, which are different from Reddit Posts. Another potential reason is that about 40% of the language evaluated was language from people with schizophrenia, a population with known emotion processing deficits. However, this may also be a result of the limited representation of certain emotion labels in this set. The weighted macro-average F-score was approximately 0.5 which is comparable to the unweighted macro-average reported for this model with a more balanced evaluation set.