Goals, Tasks, and Bonds: Toward the Computational Assessment of Therapist Versus Client Perception of Working Alliance

Alexandria K. Vail[1], Jeffrey Girard[2], Lauren Bylsma[3], Jeffrey Cohn[4], Jay Fournier[5], Holly Swartz[3], Louis-Philippe Morency[6]

[1] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2] Department of Psychology, University of Kansas, Lawrence, Kansas, USA
[3] Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[4] Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[5] Department of Psychiatry, Ohio State University, Columbus, Ohio, USA
[6] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

***Abstract*— Early client dropout is one of the most significant challenges facing psychotherapy: recent studies suggest that at least one in five clients will leave treatment prematurely. Clients may terminate therapy for various reasons, but one of the most common causes is the lack of a strong *working alliance*. The concept of working alliance captures the collaborative relationship between a client and their therapist when working toward the progress and recovery of the client seeking treatment. Unfortunately, clients are often unwilling to directly express dissatisfaction in care until they have already decided to terminate therapy. On the other side, therapists may miss subtle signs of client discontent during treatment before it is too late. In this work, we demonstrate that nonverbal behavior analysis may aid in bridging this gap. The present study focuses primarily on the head gestures of both the client and therapist, contextualized within conversational turn-taking actions between the pair during psychotherapy sessions. We identify multiple behavior patterns suggestive of an individual's perspective on the working alliance; interestingly, these patterns also differ between the client and the therapist. These patterns inform the development of predictive models for self-reported ratings of working alliance, which demonstrate significant predictive power for both client and therapist ratings. Future applications of such models may stimulate preemptive intervention to strengthen a weak working alliance, whether explicitly attempting to repair the existing alliance or establishing a more suitable client-therapist pairing, to ensure that clients encounter fewer barriers to receiving the treatment they need.**

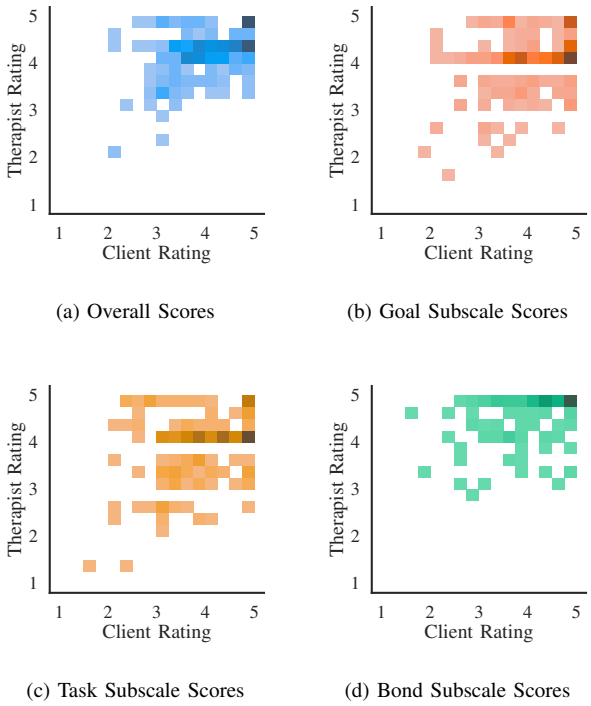
I. INTRODUCTION

Previous research has established that the strength of the relationship between a client and their therapist is a robust predictor of positive therapy outcomes [24], [29], [32]. Much of the current psychological literature on the client-therapist relationship pays particular attention to what is known as the *working alliance*. Although many variations on the definition of 'working alliance' can be found, there is a consensus on the central idea that the working alliance captures the *collaborative* aspect of the therapist-client relationship [3], [24]. Higher therapist-reported and especially client-reported ratings of the working alliance have been strongly associated with reduction of the client's symptoms and concerns [14], [23], [24], but also with other positive therapy outcomes such as reduced drug abuse and recidivism [30] and improved medication compliance [13]. Of particular note is the recognized relationship between the strength of the working alliance and client dropout [13], [27], [41]. Proactive detection is especially valuable in this case: by the time a client has decided to quit therapy, the time for potential intervention has already passed. Understanding the complexity of the therapist-client relationship is crucial for informed treatment decision-making.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

Unfortunately, measuring the strength of a working alliance faces several challenges. Most recorded ratings of the working alliance are obtained by self-reports from the client and their therapist, who are also participants in the interaction; previous research has documented significant divergences in these two participants' perception of the working alliance. Clients are often hesitant to express feedback or concerns [36], [37]: many clients do not express any concern at all until they have already decided to discontinue treatment [21]. On the other hand, therapists often miss subtle signs of client discontent during therapy sessions [36]. Alarmingly, some studies have even demonstrated that therapists may perform worse than random chance at identifying signs of client frustration or annoyance [20], [32]. Several attempts have been made to evaluate the reliability of third-party human observers, but to date, observer ratings of the working alliance have repeatedly emerged as the least valuable predictors of therapy outcomes [24], [32].

The primary aim of this paper is to explore the use of computational behavior analysis to overcome the obstacles facing the objective measurement of the working alliance. Our analysis focuses primarily on head gestures and turn-taking behaviors, as these features have been identified as essential signals in the detection of similar measures of relationship [6], [16]. We begin with a set of inferential analyses to explore general trends in behavior that may indicate a participant's perception of the working alliance. Given these identified patterns, we develop a series of predictive models to estimate the working alliance ratings provided by the therapist and the client. Following this, we perform a set of ablation studies to examine the value of including specific categories of behavioral features, such as therapist behavior versus client behavior or head gesture features versus turn-



(a) Overall Scores (b) Goal Subscale Scores

(c) Task Subscale Scores (d) Bond Subscale Scores

Fig. 1. Heatmap distributions of client and therapist rating of working alliance and its subscales

taking features. Finally, we conclude by discussing some of the most notable takeaways revealed by these results and the promising directions for future work.

II. RELATED WORK

To date, there has been little to no computational behavior analysis of working alliance in psychotherapy. However, there is a large volume of published studies in the computational literature that explores a similar construct: *rapport*, which can broadly be defined as mutual attentiveness, amiability, and receptivity between interaction participants [42]. Rapport differs from the working alliance in several fundamental ways, but one of the most notable differences is that rapport is generally considered to be 'other-focused', in which the primary goal is to develop a relationship between participants [42]. In contrast, the working alliance is 'task-focused', in which developing the relationship is secondary to the accomplishment of mutual goals [3]. The working alliance is more commonly described in asymmetric interactions, such as between therapist and client or teacher and student [24]. However, both concepts are related to relationship-building, and given the relative paucity of studies investigating working alliance computationally, we draw insight from the considerable amount of literature on the similar concept of rapport.

In previous studies of dyadic interaction, different behaviors have been shown to be related to rapport-building. One such behavior is head gestures: nodding is recognized as one of the most valuable indicators of rapport between human participants [42]. To a lesser extent, head shakes are also related to rapport in therapeutic contexts [43]. A growing body of literature has investigated the incorporation of rapport-building when designing virtual agents; gestures of both head and hands have been identified as some of the most influential behaviors for inclusion [16], [38].

Significant attention has also been paid to turn-taking behaviors in 'listening' agents [8]. Appropriate backchanneling (verbal and nonverbal) is critical to developing user trust [2]. Similarly, increased pauses have been recognized as positively impacting rapport-building, in terms of waiting to 'grab the floor' after a partner's dialogue turn but also within a turn, allowing the partner to 'grab the floor' themselves [7]. Taking longer dialogue turns — speaking for longer periods before transitioning to the partner — significantly impairs the development of rapport between participants [6]. Given that the therapeutic setting is an asymmetric interaction, 'listening' behaviors are especially pertinent in this context.

III. DATASET

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

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

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

A. Ratings of Working Alliance

Following the conclusion of each therapy session, both therapist and client participants completed the therapist and client versions of the revised short-form Working Alliance Inventory (WAI; [19]), a widely-used measure of alliance in therapy. The WAI consists of three subscales capturing three aspects of working alliance:

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

TABLE I

SAMPLE ITEMS FROM BOTH THERAPIST AND CLIENT VERSIONS OF THE WORKING ALLIANCE INVENTORY

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

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

Fig. 2. Inferential model specification in formula notation

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

Each subscale consists of a set of statements which the individual rates on a five-point Likert-type scale ranging from 'seldom true' to 'always true'. Representative items for each subscale are presented in Table I, and the distribution of scores observed in our dataset is illustrated in Figure 1.

B. Head Gesture Annotation

Head motion for each participant was automatically measured using the OpenFace facial behavior analysis toolkit [1]. Gestures of interest in the present study were limited to head nods (vertical motion along the pitch dimension) and head shakes (horizontal motion along the yaw dimension). A low-resource algorithm was selected to classify head gestures based on prior work using basic dimensions of motion [25], [26], [44]. Although these works derived head motion from the motion of a particular facial landmark between the eyes, our implementation instead incorporates head motion derived from the head tracking features provided by OpenFace [45]. Total distance traveled along each dimension was calculated over a rolling window of one second, and gestures were detected based on the top quartile of distance traveled within one second.

C. Speaking Turn Annotation

We define a 'speaking turn' as a contiguous speech segment from a single speaker until a non-speaking pause longer than one second. To determine speaking turns throughout the session, we performed speaker diarization (i.e., identifying when each speaker is actively speaking) using a voice activity detection algorithm available through openSMILE [12]. By applying this detection algorithm to each of the two participant microphones (client and therapist), the resulting annotations indicate whether the client or the therapist is presently speaking or, occasionally, if both are speaking.

TABLE II

SUMMARY STATISTICS FOR FEATURES DERIVED FROM HEAD GESTURES AND TURN-TAKING BEHAVIORS

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

IV. ANALYSIS

The present analysis consists of three stages. We begin with a set of inferential models to identify meaningful relationships between participant behaviors and working alliance ratings. We then incorporate these behaviors into a set of predictive models to estimate working alliance ratings. Finally, we perform a set of ablation studies to examine the value of including specific categories of behavior features: (1) client behavior vs. therapist behavior, and (2) head gestures vs. turn-taking behaviors.

Our feature set is primarily composed of the two sets of features derived from head gestures and speaking turns, as described in Sections III-B and III-C. Four additional features were derived from head gestures and turn-taking behaviors

TABLE III

CLIENT RATINGS — POPULATION-LEVEL EFFECTS FROM INFERENTIAL MODELS OF WORKING ALLIANCE RATINGS

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

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

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

A. Inferential Analysis

Due to the nested structure of our recorded client-therapist interactions, we utilize a multilevel modeling approach to account for multiple sessions per client and multiple clients per therapist. Recognizing the multilevel structure of such interactions is critical, as these observations are not wholly independent, and such dependencies could bias parameter estimation or model building during training time [10]. We follow an established method for decomposing longitudinal data into three separate components [17].

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

We approach our models from a Bayesian perspective. Bayesian methods provide a means of augmenting pre-existing domain knowledge (in the form of a prior distri-

TABLE IV

THERAPIST RATINGS — POPULATION-LEVEL EFFECTS FROM INFERENTIAL MODELS OF WORKING ALLIANCE RATINGS

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

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

bution) with data-driven updates (in the form of observed data) to construct more robust models than either technique can achieve individually [15]. These analyses were performed using the `bambi` Python package [5], a high-level interface for the probabilistic programming framework PyMC3 [40]. Models were estimated through Markov chain Monte Carlo [33] via the No-U-Turn Sampler algorithm [22]. The model specification is presented in Figure 2. This equation describes the form of the model, in which each term includes an implied coefficient: these coefficients are parameters estimated during training time.

Interpretation of these models requires examining the resulting posterior distribution (the estimated distribution after observed-data updates) for each model parameter. To quantify these posterior distributions, we measure the posterior median and the 89% highest density interval (HDI). These two measures help us study the central tendency and spread, respectively, for each of the model parameters (also known as *effects*). The posterior median minimizes absolute error; the 89% HDI is common in Bayesian analysis as it is more stable than the 95% HDI [28]. To understand the significance of the observed results, we also calculate the probability of direction (pd), a metric ranging between 50% and 100%, indicating the probability that a given parameter has the same sign as the posterior median [31]. We interpret pd values greater than 95% as 'significant' and pd values greater than 99% as 'highly significant'. Tables III and IV present the results obtained from the inferential analyses of client and therapist working alliance ratings, respectively. Note that each row of the table indicates a separate model and that client behavior models were examined independently from therapist behavior models.

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

therapist. In both cases, the working alliance ratings were more associated with the behavior of the person providing the ratings than with the behavior of their partner.

B. Predictive Models

To evaluate the predictive power of head gestures and turn-taking behaviors in estimating working alliance ratings, we developed a set of models targeting each WAI subscale. Using the therapist-level, client-level, and session-level aggregated features (see Section IV-A for details), we evaluated three predictive modeling procedures: support vector regression (SVR; [11]), Elastic Net [46], and random forests [4]. These algorithms were selected based on their ability to perform well on small datasets.

Model hyperparameters were automatically selected using a nested leave-one-therapist-out cross-validation approach in order to minimize train-test data contamination. For each therapist ($n = 11$), all sessions conducted by that therapist were designated as the test set, while all other sessions were designated as the training set. Within the training set, validation for each fold was performed similarly: the sessions from one therapist were used for validation, while the remaining sessions were used for training. Features were recomputed for each training run to ensure that they do not rely on values from the test set. Prediction performance during validation and testing was measured using the root mean squared error (RMSE) metric. A benefit of RMSE over other similar metrics (e.g., the coefficient of determination R^2) is its definition in the same units as the output variable — in this case, working alliance ratings — and its stability in smaller datasets. Table V compares the test-set performance for each prediction model. For comparison, we also include a baseline model predicting the mean from the training set. All three models performed above the baseline model: the SVR and Elastic Net models tended to achieve the lowest RMSE.

C. Ablation Studies

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

V. DISCUSSION

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

Participant ratings of the working alliance are largely uninformed by the behavior of the other participant. A consistent theme throughout these results is the suggestion that client behaviors do not offer much insight into therapist ratings and similarly that therapist behaviors do not offer much insight into client ratings. This result corroborates prior work suggesting a frequent disconnect between therapist and client perception of the alliance [36], [37]. Also of note is the trend that client behaviors appear to hold more predictive power toward client ratings than therapist behaviors hold toward therapist ratings. This result is a valuable finding, as previous work has established that client ratings of the working alliance are the most reliable indicators of positive therapy outcomes, compared to therapist and observer ratings [24].

Head gestures tend to be more reflective of the task-oriented components of the working alliance, while turn-taking behaviors tend to be more reflective of the relationship-oriented component. As in many similar multimodal analyses [35], [39], our results identify trends in the salience of particular behavioral signals during the prediction of different outcome measures (Table VI). We note that turn-taking behaviors (speaking turn length and wait time) were primarily associated with the relationship-oriented component of the working alliance ratings — the bond subscale. On the other hand, head gestures (head nods and head shakes) were associated mainly with the working alliance ratings' task-oriented components — the goal and task subscales. There are similarities between these connections and those identified in studies of rapport, which recognize head gestures as more 'contentful' interaction signals [16], [43] and turn-taking patterns as more indicative of trust and respect [42]. We also note that the derived features (listening nods and listening shakes) were more predictive of the goal and task subscales than the bond subscale. This result could be attributed to prioritization among behavior signals, indicating that head gestures are a 'stronger' signal than turn-taking behaviors.

Beyond simply being *uninformed* by the partner's behavior, in certain cases, working alliance ratings are *misinformed* by the partner's behavior. A comparison of the behavior patterns associated with client ratings (Table III) and therapist ratings (Table IV) reveals a few notable divergences. In one case, an increase in nodding on the part of the therapist was generally associated with the therapist providing *higher* ratings on the goal subscale. However, this same therapist behavior was associated with *lower* goal and task subscale ratings from the client. Similarly, when clients nodded more frequently when listening, clients tended to provide *higher* ratings on all subscales, but therapists tended

TABLE V

PERFORMANCE METRICS OF PREDICTIVE MODELS: ROOT MEAN SQUARE ERROR, MEDIAN AND STANDARD DEVIATION

	Client Ratings				Therapist Ratings			
	Overall	Goal	Task	Bond	Overall	Goal	Task	Bond
Baseline	0.82 (0.21)	0.86 (0.21)	0.94 (0.24)	0.85 (0.22)	0.39 (0.31)	0.61 (0.36)	0.61 (0.42)	0.36 (0.29)
SVR	**0.63 (0.22)**	0.69 (0.22)	0.74 (0.19)	**0.60 (0.25)**	**0.31 (0.27)**	0.42 (0.30)	**0.50 (0.36)**	**0.30 (0.23)**
Elastic Net	0.65 (0.23)	**0.66 (0.22)**	**0.68 (0.23)**	0.65 (0.23)	0.37 (0.25)	**0.42 (0.31)**	0.53 (0.35)	0.32 (0.23)
Random Forest	0.72 (0.18)	0.73 (0.20)	0.73 (0.18)	0.77 (0.19)	0.38 (0.21)	0.43 (0.26)	0.58 (0.28)	0.36 (0.29)

TABLE VI

PERFORMANCE METRICS OF ABLATION STUDIES: ROOT MEAN SQUARE ERROR, MEDIAN AND STANDARD DEVIATION

	Client Ratings				Therapist Ratings			
	Overall	Goal	Task	Bond	Overall	Goal	Task	Bond
Client Behavior	**0.64 (0.25)**	**0.69 (0.23)**	**0.71 (0.26)**	**0.70 (0.28)**	0.64 (0.31)	0.80 (0.44)	0.84 (0.46)	0.64 (0.33)
Therapist Behavior	0.95 (0.29)	1.02 (0.30)	1.04 (0.28)	1.03 (0.30)	**0.44 (0.34)**	**0.55 (0.40)**	**0.71 (0.45)**	**0.38 (0.30)**
Gesture Features	0.70 (0.23)	0.75 (0.25)	0.76 (0.26)	0.78 (0.30)	0.51 (0.33)	0.61 (0.40)	0.63 (0.47)	0.47 (0.36)
Turn-Taking Features	0.71 (0.25)	0.77 (0.23)	0.78 (0.33)	**0.73 (0.29)**	0.53 (0.36)	0.64 (0.37)	0.65 (0.47)	**0.44 (0.30)**
Gest. + Turn. Features	**0.67 (0.27)**	**0.74 (0.26)**	**0.73 (0.27)**	0.74 (0.25)	**0.49 (0.34)**	**0.59 (0.38)**	**0.60 (0.42)**	0.45 (0.31)

to provide *lower* ratings. These results seem to be consistent with other research, which found that therapy participants often 'misread' the behavioral cues of their partner [20], [36]. Despite this, our computational models were capable of predicting both participants' self-reported ratings of working alliance with moderate accuracy (Table V).

VI. CONCLUSION

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

Together these results provide important insights into the challenges facing assessment of the working alliance during therapy and how computational behavior analysis holds promise for addressing these obstacles. Further research might explore the role of personal characteristics (e.g., personality, sociodemographics) or the client's psychiatric concerns (e.g., anxiety, depression), as the influence of these factors on nonverbal behavior is well-established [9], [34]. Although the sample of participants in this work is diverse and representative of the population in one community, generalizations to broader populations dissimilar to this one will require additional data collection and repeat analysis. A natural progression of this work would also include other behavioral signals, such as facial expressions or acoustic patterns in speech. The understanding gained through this line of research can foster the development of systems providing early detection of a weak working alliance, allowing for preemptive intervention and reduction in the barriers facing clients seeking treatment.

VII. ACKNOWLEDGMENTS

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

REFERENCES

[1] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the Thirteenth IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[2] T. Bickmore and J. Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2001)*, pages 396–403, Seattle, Washington, United States, 2001. ACM Press.

[3] E. S. Bordin. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3):252–260, 1979.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] T. Capretto, C. Piho, R. Kumar, J. Westfall, T. Yarkoni, and O. A. Martin. Bambi: A simple interface for fitting Bayesian linear models in Python. *arXiv:2012.10754 [stat]*, 2021.

[6] J. Cassell. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1/2):89–132, 2003.

[7] J. Cassell. Towards a model of technology and literacy development: Story listening systems. *Journal of Applied Developmental Psychology*, 25(1):75–105, 2004.

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

[9] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Proceedings of the Third International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, pages 1–7, Amsterdam, Netherlands, 2009. IEEE.

[10] J. De Leeuw and E. Meijer, editors. *Handbook of Multilevel Analysis*. Springer-Verlag, New York, 2008.

[11] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In M. C. Mozer, M. Jordan, and T. Petsche, editors, *Proceedings of the Ninth International Conference on Neural Information Processing Systems (NIPS 1996)*, volume 9. MIT Press, 1997.

[12] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. In *Proceedings of the Thirty-Eighth IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pages 483–487, 2013.

[13] A. F. Frank. The role of the therapeutic alliance in the treatment of schizophrenia: Relationship to course and outcome. *Archives of General Psychiatry*, 47(3):228, 1990.

[14] L. Gaston, C. Marmar, D. Gallagher, and L. Thompson. Alliance prediction of outcome beyond in-treatment symptomatic change as psychotherapy processes. *Psychotherapy Research*, 1(2):104–112, 1991.

[15] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, third edition, 2015.

[16] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *Intelligent Virtual Agents*, volume 4722, pages 125–138. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[17] E. L. Hamaker and B. Muthén. The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25(3):365–379, 2020.

[18] M. Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1):56–62, 1960.

[19] R. L. Hatcher and J. A. Gillaspy. Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research*, 16(1):12–25, 2006.

[20] C. Hill, B. Thompson, and M. Corbett. The impact of therapist ability to perceive displayed and hidden client reactions on immediate outcome in first sessions of brief therapy. *Psychotherapy Research*, 2(2):143–155, 1992.

[21] C. E. Hill, E. Nutt-Williams, K. J. Heaton, B. J. Thompson, and R. H. Rhodes. Therapist retrospective recall impasses in long-term psychotherapy: A qualitative analysis. *Journal of Counseling Psychology*, 43(2):207–217, 1996.

[22] M. D. Hoffman and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[23] A. O. Horvath and L. S. Greenberg. The development of the Working Alliance Inventory. In *The Psychotherapeutic Process: A Research Handbook*, Guilford Clinical Psychology and Psychotherapy Series, pages 529–556. Guilford Press, New York, NY, US, 1986.

[24] A. O. Horvath and B. D. Symonds. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of Counseling Psychology*, 38(2):139–149, 1991.

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

[26] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *Proceedings of the Workshop on Perceptive User Interfaces (PUI 2001)*, pages 1–5, New York, NY, USA, 2001. Association for Computing Machinery.

[27] A. M. Kokotovic and T. J. Tracey. Working alliance in the early phase of counseling. *Journal of Counseling Psychology*, 37(1):16–21, 1990.

[28] J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, Boston, edition 2 edition, 2015.

[29] M. J. Lambert, editor. *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. John Wiley & Sons, Hoboken, N.J, 6th ed edition, 2013.

[30] L. Luborsky. Therapist success and its determinants. *Archives of General Psychiatry*, 42(6):602, 1985.

[31] D. Makowski, M. S. Ben-Shachar, S. H. A. Chen, and D. Lüdecke. Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, 10:2767, 2019.

[32] D. J. Martin, J. P. Garske, and M. K. Davis. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, 68(3):438–450, 2000.

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

[34] S. Nowicki and M. P. Duke. Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, 18(1):9–35, 1994.

[35] E. M. Provost, Y. Shangguan, and C. Busso. UMEME: University of Michigan Emotional McGurk Effect data set. *IEEE Transactions on Affective Computing*, 6(4):395–409, 2015.

[36] D. L. Rennie. Clients' deference in psychotherapy. *Journal of Counseling Psychology*, 41(4):427–437, 1994.

[37] R. H. Rhodes, C. E. Hill, B. J. Thompson, and R. Elliott. Client retrospective recall of resolved and unresolved misunderstanding events. *Journal of Counseling Psychology*, 41(4):473–483, 1994.

[38] L. D. Riek, P. C. Paul, and P. Robinson. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3(1-2):99–108, 2010.

[39] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the Seventh Annual Audio/Visual Emotion Challenge (AVEC 2017)*, pages 3–9, Mountain View California USA, 2017. ACM.

[40] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

[41] S. M. Saunders, K. I. Howard, and D. E. Orlinsky. The therapeutic bond scales: Psychometric characteristics and relationship to treatment effectiveness. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, 1(4):323–330, 1989.

[42] L. Tickle-Degnen and R. Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293, 1990.

[43] P. Tsui and G. L. Schultz. Failure of rapport: Why psychotherapeutic engagement fails in the treatment of Asian clients. *American Journal of Orthopsychiatry*, 55(4):561–569, 1985.

[44] H. Wei, P. Scanlon, Y. Li, D. S. Monaghan, and N. E. O'Connor. Real-time head nod and shake detection for continuous human affect recognition. In *Proceedings of the Fourteenth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2013)*, pages 1–4, 2013.

[45] A. Zadeh, Y. C. Lim, T. Baltrušaitis, and L.-P. Morency. Convolutional experts constrained local model for 3D facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW 2017)*, pages 2519–2528, Venice, 2017. IEEE.

[46] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.