Perceived Conversation Quality in Spontaneous Interactions

Raman, Chirag; Prabhu, Navin Raj ; Hung, Hayley

DOI
10.1109/TAFFC.2023.3233950

Publication date
2023

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Perceived Conversation Quality in Spontaneous Interactions

Chirag Raman, Navin Raj Prabhu, *Member, IEEE*, and Hayley Hung, *Member, IEEE*

Abstract—The quality of daily spontaneous conversations is of importance towards both our well-being as well as the development of interactive social agents. Prior research directly studying the quality of social conversations has operationalized it in narrow terms, associating greater quality to less small talk. Other works taking a broader perspective of interaction experience have indirectly studied quality through one of the several overlapping constructs such as rapport or engagement, in isolation. In this work we bridge this gap by proposing a holistic conceptualization of conversation quality, building upon the collaborative attributes of cooperative conversation floors. Taking a multilevel perspective of conversation, we develop and validate two instruments for perceived conversation quality (PCQ) at the individual and group levels. Specifically, we motivate capturing external raters' gestalt impressions of participant experiences from thin slices of behavior, and collect annotations of PCQ on the publicly available MatchNMingle dataset of in-the-wild mingling conversations. Finally, we present an analysis of behavioral features that are predictive of PCQ. We find that for the conversations in MatchNMingle, raters tend to associate smaller group sizes, equitable speaking turns with fewer interruptions, and time taken for synchronous bodily coordination with higher PCQ.

Index Terms—Perceived conversation quality, spontaneous interactions, social and behavioral sciences, group interactions

1 Introduction

Picture a spontaneous interaction such as a daily social conversation at work or home. The quality of such conversations is of importance towards both our well-being as well as the development of interactive technologies that influence our daily lives. At an individual level, conversation quality is directly associated with our happiness and life satisfaction [1], [2]. Furthermore, human judgement of conversation quality is a common measure for the evaluation of artificial conversation agents [3], [4]. Despite its importance, little prior research has directly studied conversation quality or jointly considered the factors affecting its perception.

One challenge is that conversation quality is not directly measured, and needs to be inferred from observable verbal and non-verbal behavioral cues. This has led to some research viewing conversation quality in narrow terms, considering only isolated attributes of the conversation. For instance, Milek et al. [1] and Mehl et al. [2] consider greater conversation quality to correspond to less small talk and information exchange at more than a trivial level of depth. On the other hand, taking a broader view of conversation quality runs into another challenge: its potential intersection with several overlapping social concepts. These include rapport [5], bonding [6], interest-levels [7], and involvement [8] amongst others. When studied towards the development of interactive dialogue agents, the focus has been on the verbal content of non-spontaneous dyadic conversations with a chatbot [3], [4]. In the second ConvAI2 Challenge, the human judgment of quality was evaluated simply as a measure of enjoyment through the question *"How much did you enjoy talking to this user?"* [4]. See et al. [3] conducted a large-scale study to identify the fine-grained factors governing human judgments of full conversations. Even here, the human judgment of overall quality is expressed in terms of the *humanness* and *engagingness* of artificially generated verbal dialogues. Moreover, the recording of spontaneous conversations in a way that enables the transcription of verbal content constitutes a privacy concern with ethical implications [9], [10]. Consequently, while individual factors have been studied in isolation, joint consideration of the multiple aspects of conversation quality in natural, spontaneous conversations remains a knowledge gap.

In this work, we take the perspective that such a *holistic* characterization of the quality of multiparty spontaneous interactions is an important objective in the development of socially intelligent systems. For instance, consider a social robot approaching a conversing group of people, as illustrated in Fig. 1. Here, a perception of the group's experience of the conversation as a whole could aid the social agent in developing more nuanced policies of approach. Furthermore, an estimate of each individual's experience could then aid the agent in developing personalized adaptive strategies to conduct the subsequent interaction smoothly.

- *Chirag Raman and Hayley Hung are with the Intelligent Systems, Delft University of Technology EEMCS, 2628 XE Delft, Zuid-Holland, The Netherlands. E-mail: {c.a.raman, h.hung}@tudelft.nl.*
- *Navin Raj Prabhu is with the Signal Processing Lab and Organisation Psychology Lab, University of Hamburg, 20146 Hamburg, Germany. E-mail: lr.navin@yahoo.com.*

Manuscript received 24 March 2022; revised 18 December 2022; accepted 22 December 2022. Date of publication 3 January 2023; date of current version 29 November 2023.
This work was supported by the Netherlands Organization for Scientific Research (NWO) under the MINGLE project under Grant 639.022.606.
This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.
(Corresponding author: Chirag Raman.)
Recommended for acceptance by M. CHETOUANI.
This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2023.3233950, provided by the authors.
Digital Object Identifier no. 10.1109/TAFFC.2023.3233950

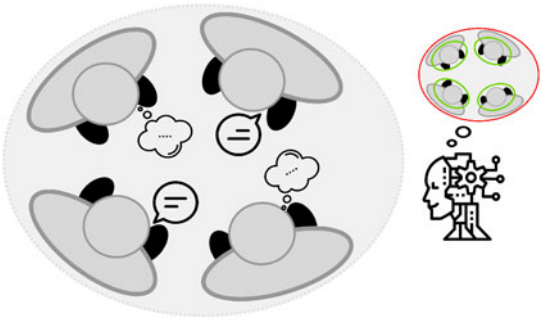


Fig. 1. Conceptual illustration of individual experiences existing in the perception of interacting partners, and how an external perceived measure of individual-level (green) and group-level (red) experience is relevant for the development of artificial interactive social agents.

In addition to a holistic characterization, we specifically argue for a *perceived* measure of conversation quality in this work, at both the individual and the group levels. This is in contrast to existing efforts for quantifying quality-related aspects of conversations, which have largely focused on self-reported measures after interactions [5], [6], [11], [12]. While such measures attempt to estimate an individual's true experience in situ, they also suffer several drawbacks including desirability bias [13], egoistic bias [13], [14], and recall bias and cognitive errors [15]. On the other hand, a perceived measure of experience quantifies how participants seem to be experiencing the interaction to an external third-party observer [7], [8], [16]. While such a measure may not capture the true experience, it closely models how we conduct interactions based on imperfect estimates of our conversation partners' experiences, and is therefore also useful towards the development of machines with social intelligence.

Concretely, we make three contributions in this work. First, we introduce the novel measure of *Perceived Conversation Quality* (PCQ) towards quantifying social experience in spontaneous interactions by jointly considering potentially overlapping related constructs. Second, we present an instrument for collecting annotations of PCQ at both the individual and the group level. We validate the instrument on the publicly available MatchNMingle dataset [9] of mingling interactions following a speed-dating event. Third, we present insights into the behavioral features that predict PCQ through confirmatory statistical analysis and empirical data-driven analysis.

Our preliminary work on this topic was presented in [17], which described the proposed instrument and analysis of annotations. The experiments we present in this manuscript (Section 5 onward) are completely new. Moreover, this manuscript is a complete rewrite; compared to our prior publication the manuscript now includes a clearer (i) overall presentation and motivation, (ii) organization of related literature, and (iii) description of the process of conceptualizing, validating, and analyzing PCQ.

2 Related Work

Spontaneous interactions are considered to be non task-directed, unconstrained, and typically occurring in natural situations [18], [19], [20]. In such a dynamic conversation setting, several constructs emerge. These include descriptors of interpersonal relationships amongst participants (e.g., rapport [5] and bonding [6]), or those which capture qualitative attributes of the interaction (e.g., involvement [8], [21], engagement [22], and interest-levels [7]).

2.1 Rapport and Bonding

Rapport and bonding have been widely studied as a pairwise phenomena using self-reported measures [5], [6], [11]. Müller et al. [5] define rapport as "the close and harmonious relationship in which interaction partners are 'in sync' with each other". The authors used a self-reported questionnaire adapted from Bernieri et al. [23] to measure rapport for every pair of individuals within small interaction groups. Another related social concept is bonding, which measures positive personal attachment including "mutual trust, acceptance, and confidence" amongst interacting pairs [24]. Based on this definition, Jaques et al. [6] studied bonding in human-agent interactions, using the bonding subscale of the *Working Alliance Inventory* (B-WAI) [24].

2.2 Involvement, Engagement, and Interest-Levels

Antil [21] defines involvement as "the level of perceived personal importance and/or interest evoked by a stimulus (or stimuli) within a specific situation". Following Antil's view of involvement as a non-binary variable, Oertel et al. [8] developed a 10-level annotation scheme for joint involvement of a group based on intuitive, listener-independent impressions of prosody and body and face movement. Oertel and Salvi [25] proposed a gaze-based method to relate group involvement to individual engagement in multiparty dialogue. Several researchers have conceptualized group cohesion to study its influence on task performance [26], in settings such as meetings [27], [28] and long-term crew missions [29], [30]. Gatica-Perez et al. [7] define group interest-levels as, "the perceived degree of interest or involvement of the majority of the group". The authors provided perceived annotations for interest-levels using audio-visual recordings of interactions, on a discrete 5-point scale. To this end, the external annotators were instructed to attend to interest-indicating activities such as note-taking, focused gaze, and avid participation in discussion. Note that these constructs have all been defined and studied in task-directed settings.

2.3 General Measures of Interaction Experience

In contrast to efforts focusing on specific social concepts, some recent approaches have proposed more general measures of experience in conversations. Cuperman and Ickes [12] introduced the *Perception of Interaction* (POI) questionnaire as part of a study to examine the effects of gender and personality traits on participant behaviors in dyadic interactions. The questionnaire collected self-reported measures of a participant's perception of their interaction experience. These aspects included the perceived quality of the interaction, the degree of rapport they felt they had with the other person, and the degree to which they liked the other person. This measure of interactions has been adapted by other works to study bonding [6] and interaction experience [31]. Lindley and Monk [16] follow the rationale that experience itself is difficult to quantify, but since it is entwined with social interaction, we might characterize experience by measuring aspects of conversation that are related to it. They studied several behavioral process measures and developed

the *Thin-Slice Enjoyment Scale* (TES): a measure of empathized enjoyment in social conversations from ratings of thin slices of behavior by naïve judges. In their factor analysis, the authors found that the judges viewed enjoyment and conversation fluency as being related. However, the POI was developed for self-reported measures, and neither work considered spontaneous interaction settings: Cuperman and Ickes [12] considered scripted dyadic interactions with confederates, while Lindley and Monk [16] developed the TES within the particular task-directed context of photo sharing.

3 Perceived Conversation Quality

3.1 Initial Conceptualization

The primary influences for our conceptualization of PCQ are the works of Edelsky [32], Lindley and Monk [16], and Cuperman and Ickes [12]. Specifically, from these works we motivate the rationale behind our choices of (i) focusing on the cooperative aspects of conversation towards conceptualizing PCQ, and (ii) rating thin slices of behavior to capture the gestalt impressions raters have of the continually unfolding conversation.

In an analysis of social interactions in a series of meetings, Edelsky [32] observed two contrasting styles of conversation, termed *cooperative floors* and *exclusive floors*. Cooperative floors are characterized by collaborative stretches of "free-for-all" conversation accompanied by a feeling of participants being "on the same wavelength" [32, p. 391]. (In contrast, the exclusive floor is owned by a single person with turns rarely overlapping.) This notion of the cooperative floor captures the sense of engagement associated with positive experiences, and has been since linked with informal social interactions [33], [34], [35] and enjoyment [36]. As such, we observe that Edelsky's notion of "on the same wavelength" strongly resonates with the POI questionnaire's focus on how interaction partners relate to each other [12]. Subsequent researchers have also derived qualitative measures of conversation based on the "free-for-all" aspects of Edelsky's description. These include conversational equality and freedom [16] (or interactivity [37]), and fluency through the occurrence of frequent turns [16], [38].

Ambady and Rosenthal [39] propose that thin slice judgments of behavior can be usefully made so long as the variables in question are observable and there is an affective or interpersonal component. They suggest that this is because such inferences are made through subconscious decoding of expressive behavior, with judgemental accuracy being strongly linked to "gestalt, molar impressions based on nonverbal behavior" [40, p. 439]. This result supports previous research showing that molar impressions, although vaguer and fuzzier, generally yield more useful information than the coding of specific behaviors without accounting for overall context. Researchers often encourage the formation of this gestalt impression by intentionally reducing information presented to raters, e.g., removing speech content while retaining tone of voice or extinguishing facial expressions [41]. In contrast, obtaining judgments of gestalt impressions is a natural fit for spontaneous interaction settings where recording speech or ego-centric perspectives is often not possible to preserve privacy [9], [10], [42].

3.2 Pilot Qualitative Interviews With Naïve Judges

We conducted pilot qualitative interviews with three naïve judges to verify if our initial conceptualization matched the lay interpretation of PCQ. All judges were students enrolled in technical Masters programs at the authors' university. The judges were shown unaltered recordings from the publicly available MatchNMingle (MnM) dataset [43], and asked what they thought of the conversations in the scene. Fig. 2 illustrates a snapshot of a scene from MnM. To obtain unbiased impressions, we didn't specify our focus on conversation quality, nor our conceptualization of it. All judges (i) described a continually evolving perception of participant experiences over the conversation lifetime, aligning with our choice of rating thin slices of behavior rather than a single rating for the entire conversation; (ii) described perception of individual experiences as well as the group as a whole, aligning with our choice of measuring PCQ at the individual- and group- levels separately; and (iii) identified the attributes of equal opportunity for speaking, smoothness of interaction, and interpersonal relationships that strongly resonates with the prior work that serves as our primary influences [12], [16], [32].

Fig. 2. A snapshot from the MatchNMingle dataset [43].

3.3 Definition and Constituents

Following our initial conceptualization and pilot interviews, we formalize PCQ of a spontaneous interaction as

> the degree to which participants in the spontaneous interaction appear to be on the same wavelength and maintain an equal opportunity floor, as perceived by an external observer.

Further, in the following subsections we present three constituents of PCQ that categorize the multiple social concepts associated with this definition.

3.3.1 *Interpersonal Relationships*

This constituent describes the degree of association between participants or the notion of being in-sync with one's interaction partners, using constructs such as rapport [5] and bonding [6]. More specifically, the constituent measures the degree to which an individual was accepted and respected by other individuals in the group or the degree to which the other individuals were paying attention to the individual. Increased bonding and rapport amongst interacting partners is widely acknowledged to result in improved collaboration, and improved interpersonal outcomes, thereby having a key influence on the PCQ.

3.3.2 *Nature of Interaction*

This constituent describes the degree to which the interaction was smooth and relaxed or forced and awkward. It captures the notion of whether the participants are having a

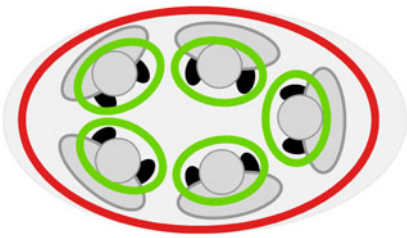


Fig. 3. Illustrating the scope of observation to measure the group-level (red) and individual-level (green) PCQ.

positive and pleasant experience, drawing upon the quality of interaction aspects of the POI [12].

3.3.3 Equal Opportunity

This constituent captures the *free-for-all* collaborative aspects of Edelsky's description of cooperative floors [32]. It describes the notion of equality of opportunity for participation shared amongst interacting partners, capturing the sense of cohesiveness and engagement in informal conversations. This includes factors such as conversation freedom [44], equality, and fluency [16] and an individual's opportunity to take the lead in the conversation [6], [12].

3.4 PCQ Questionnaires: A Multilevel Perspective

We devise two independent questionnaires to measure PCQ at the individual and group levels. This follows our broader multilevel perspective [45] of social interactions where constructs can be conceptualized at different levels, such as the individual, dyadic, and group levels. While prior works have often considered constructs at a single level (e.g., Müller et al. [5] consider rapport as a dyadic pairwise construct), a multilevel perspective aligns better with our pilot judges' descriptions of attributes pertaining to individuals and groups as a whole. Moreover, some prior works on conversation group dynamics have indeed also taken a multilevel perspective: Oertel and Salvi [25] distinguish overall group involvement from individual engagement, obtaining separate annotations at both levels. In the case of PCQ, our view is that an observer's perceptions of individual affect and behavior dynamically interact to contribute to an overall group-level perception. Fig. 3 illustrates the scope of observations towards measuring PCQ at each level.

The individual level captures what the quality of the conversation appears to be to a particular individual. The focus is on how the individual seems to be relating to their partners and participating in the conversation. Consequently, every individual receives a rating. Note that this perspective doesn't consider the individual's behavior in *isolation* by excluding the context of partner behaviors. Rather, the scope of consideration is restricted to what the individual seems to be experiencing. In contrast, the group level expands this scope of consideration to all interlocutors *as a whole*, focusing on their collective experience, resulting in a single group-level rating.

Concretely, we devise the PCQ questionnaires by drawing upon elements of the POI scale [12] and the TES [16]. However, since the POI was developed for self-reports rather than external perception, and neither was developed for spontaneous interaction settings, we adapt the specific items. First, all items were updated to address external observers and apply to group sizes beyond dyads. Second, privacy-preserving datasets of in-the-wild conversations often omit recording audio. So items referring to the verbal or paralinguistic content of speech were skipped, thereby relying solely on nonverbal cues for perception. Finally, we excluded original items that would require external raters to make significant speculations about participants' desires and opinions beyond what can be inferred from their observable behavior. These include questions related to interpersonal liking (e.g., *"I would like to interact more with the partner in the future"*), or degree of rapport (e.g., *"I felt that the partner was paying attention to my mood"*). From the varied descriptions of pilot judges on the matter, as well as internal author discussions, we deemed that answering such questions require external observers to make too many unverifiable assumptions for a useful perceived measure of conversation quality. We provide the two PCQ questionnaires in Supplementary Material Section 1, which can be found on the Computer Society Digital Library at online available http://doi.ieeecomputersociety.org/10.1109/TAFFC.2023.3233950.

4 Annotations, Validity, and Reliability

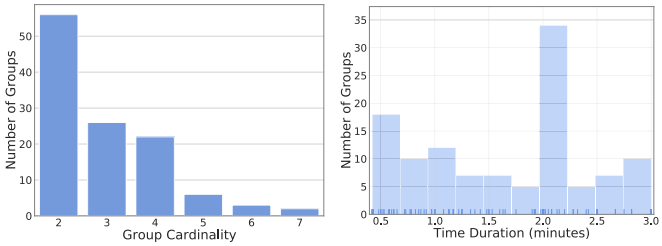
4.1 Dataset

We use the publicly available MnM dataset [43]. MnM is a multimodal dataset of in-the-wild free-standing mingling interactions. The recordings constitute a total of 30 minutes of interaction across three days, annotated for conversation groups using the spatial positions of the participants in video from overhead cameras. Fig. 2 illustrates a snapshot from the dataset. Conversation groups were operationalized using the framework of F-formations [46], where a unique group was considered to be an F-formation with a fixed number of interlocutors. The leaving or joining of one or more members was considered to give rise to new unique conversing groups. The authors of the dataset chose specific windows of 10 minutes per day for annotation with an aim to eliminate possible effects of participant acclimatization to being in a recorded mingling setting, and to maximize the density of participants in the scene. Over the 30 minutes 174 conversation groups were annotated. The duration of group conversation follows a mean of 1.91 min, std. of 2.13 min, median of 1.10 min, and a mode of 0.52 min. The provided data contains video from three of the five overhead cameras, and accelerometer readings from a sensor pack worn by each participant.

4.2 Annotation Procedure

The PCQ annotations were performed by only relying on overhead cameras *videos*. The MnM dataset contains only general audio from the overhead cameras, which is insufficient to reliably infer verbal cues of an individual, and close-talk microphone recordings are not available. However, the MnM dataset contains video recordings that capture rich non-verbal behaviors of participants from which a useful perception of conversation quality can be formed [7], [16].

We began by splitting the group conversations into multiple thin-slices [6], [47]. The distribution of group interaction duration in the data follows a median of 1.10 min and a mean of 1.91 min. For a fair comparison to conversations lasting around 1 minute, we split conversations of duration



(a) Group Cardinality (b) Duration of interactions

Fig. 4. Distribution of conversation group attributes from the MatchNMingle dataset.

greater than 2 minutes into independent slices of 1 minute each. Conversations of duration less than 2 minutes were untouched. We also omitted groups with a duration of less than 30 seconds. Note that studies on the predictive validity of thin slices of nonverbal behavior for other tasks have revealed (i) no clear pattern for optimal slice locations for 1 min slices within a longer slice [48]; and (ii) only some loss in predictive capacity for 1 min slices, while slices of duration 2 or 3 min were in general equal to 5 min slices in predictive capability [48], [49]. Considering these results along with the distribution of conversation duration in our data, we believe our choice of splitting conversations larger than 2 minutes into 1 minute slices to be reasonable. After the omission of groups lasting under 30 seconds, the total number of resulting conversation groups was 115. The distribution of group cardinality (number of participants) and interaction duration can be seen in Figs. 4a and 4b respectively.

We began by first conducting a qualitative annotation pilot with the same naïve judges who participated in the qualitative interviews. Note that these judges were not used for the final annotations. The goal of this pilot was to fine-tune the final annotation process using any initial feedback about the annotation procedure. The pilot annotators were presented with the videos of the individual thin-slices and asked to fill the two PCQ questionnaires. However, post-hoc interviews revealed two considerations. First, the annotators found the presence of free-standing conversation groups (FCGs) other than the one under consideration distracting. Second, the annotators suffered from fatigue while annotating longer conversations, especially while annotating both individual and group level PCQ. In light of this feedback, we manually cropped each FCG from the overhead video. To further reduce fatigue, annotators were given a period of two months to annotate all the slices, and were instructed to not annotate more than three groups per day.

The final annotations[1] were performed on a 5-point scale by three annotators. The annotators were chosen to be naïve judges in order to capture a general perception of conversation quality. The annotators were aged between 22 and 30 years, 2 females and 1 male. The age range matches overlaps with the reported age range of the participants in the data $(18-30)$ [9]. One of the annotators spent time internationally as a Masters student, matching the demographics of the participants. All annotators had completed education at least the Bachelors level. The annotators were provided with the independent conversation slices of cropped video clips and asked to fill out both PCQ questionnaires. The slices were provided to the annotators in randomized order for each annotator, to prevent any annotator bias which might occur from a chronological ordering of the clips.

4.3 Validity

When measuring intangible constructs such as PCQ, it is important to assess the validity [50], [51] of the proposed instrument. Broadly, validity deals with whether the instrument indeed measures what it claims to be measuring.

4.3.1 *Face Validity*

First we tested the face validity of our questionnaire items. Face validity is a consensus measure, and is checked to ensure that the raters accept the instrument [50]. This is done by asking the raters if the items seem valid. Both questionnaires passed the face validity test with full consensus.

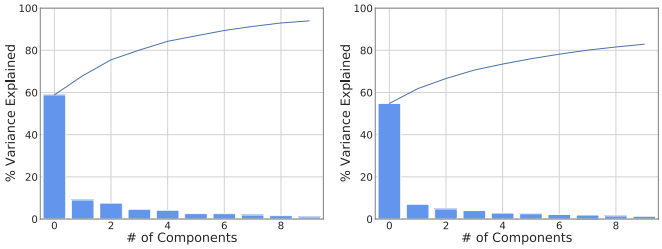
4.3.2 *Criterion and Construct Validity*

When prior trusted standards exist for a construct, a criterion-oriented study is common. Here validity can be established by showing that results of administering the instrument correlates with a contemporary criterion (e.g., a psychiatric diagnosis) or by proposing one instrument as a substitute for another (e.g., a multiple-choice form of spelling test is substituted for taking dictation) [51]. However, since PCQ is a novel conceptualization, prior trusted standards do not exist for it. In such cases where the attribute being measured is not "operationally defined", construct validity must be investigated [50], [51]. Construct validation is the gathering of evidence to support the interpretation of what a measure reflects, and addresses the question "What constructs account for variance in test performance?"

A typical approach for construct validation involves performing a factor analysis and investigating if items corresponding to one construct correlate with each other along a factor (convergent validity) and divert from items of other constructs (divergent validity) [50]. This works well for instruments with independent constructs (e.g., *gender* and *complexity of use* in Brinkman's mobile phone design questionnaire [50, Table 9]). However, such an analysis is unsuitable for situations like ours with overlapping constructs. Indeed, Cuperman and Ickes [12] decided to not reduce items from the POI to a smaller set of factors, following a precedent set by [52]. In contrast, we do perform a factor analysis, but rather than seeking the independence of factors, we investigate whether the loadings correspond to interpretable attributes of the constructs.

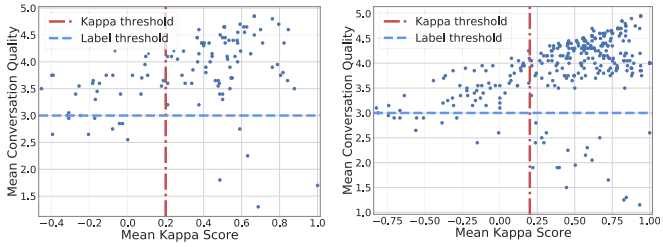
A principal component analysis (PCA) of the annotations showed that 71% and 65.2% of the variance at the group-level and individual-level respectively could be explained by the first principal component (see Fig. 5). Here, 1020 (3*340) and 345 (3*115) *thin-slice* samples were used for individual and group level PCQ (i.e., annotations from three annotators for each sample), respectively, with 10 features (the number of questionnaire items), which is greater than the variables-to-features ratio suggested to perform PCA [53]. From the plot of the data samples using the first two principal components in Fig. 6, we see that questions corresponding to

1. Annotations will be available on the MatchNMingle website at http://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/



(a) Group-level (b) Individual-level

Fig. 5. Eigenvalue distribution (bar chart) and the cumulative percentage of the explained variability (line plot).



(a) Group-level (b) Individual-level

Fig. 7. Scatter plot of the Mean Kappa score (κ) versus the Mean Conversation Quality score.

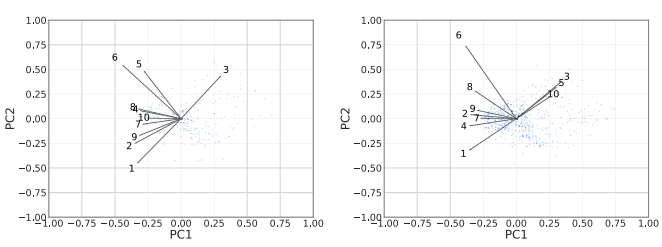
positive and negative orientations of PCQ cluster in opposite directions along the two components. Specifically the individual-level items pertaining to awkwardness (3), discomfort (5), and self-consciousness (10) load in the exactly opposite direction to the item about the individual looking relaxed (1). Of these, at the group-level only items 1 and 3 apply, and we see a similar pattern. Further, we also observe that the items pertaining to *equal opportunity* cluster separately: these correspond to items 5 and 6 about free-for-all participation at the group-level, and item 6 about taking lead at individual-level. Specifically, the highest loading of individual-level item 6 suggests that the taking lead in conversations accounts for the highest variance between individuals, which is intuitive given prior work on dominance in groups [54].

4.4 Reliability

To estimate inter-annotator agreement, we use the quadratic weighted kappa measure (κ) [55], a variant of the Cohen's kappa. The measure is especially useful when the annotation data is ordinal in nature. Fig. 7 plots the mean kappa score against the mean conversation quality score in a scatter plot similar to the analysis of inter-annotator agreement for cohesion performed by Hung and Gatica-Perez [27].

From the plots we see that there exists a linear relationship between mean kappa scores and mean conversation quality scores, suggesting that annotators agree better on conversations of higher perceived quality than conversations of lower perceived quality. Moreover, in the individual-level annotations, there exists a small cluster of samples where annotators tended to agree higher for lower conversation quality samples as well. In contrast, annotators never agree well for low conversation quality samples at the group-level.

To handle low inter-annotator agreement, following suggestions by Ringeval et al. [56], we performed zero-mean local normalization to remove annotator bias. Hung and Gatica-Perez [27] omit samples below $\kappa = 0.3$, and Ringeval et al. [56] obtain an average κ of ≈ 0.2 for all their emotion dimensions. Following these approaches, data samples at both the group- and individual- levels with $\kappa < 0.2$ were omitted from further analysis, where a $\kappa > 0.2$ indicates a reliability of *fair and above* [57].



(a) Group-level (b) Individual-level

Fig. 6. Plot of the factor loadings (black lines) and the samples (blue dots) in the first two principal components.

5 Modeling Conversation Quality

In this section we describe the experimental setup for our study of behavioral features that can be predictive of PCQ.

5.1 Preprocessing

We first preprocess the raw tri-axial acceleration signal from the wearable sensors to extract low-level features. First, each axis recording from the tri-axial accelerometer is standardized by calculating the z-score for each individual and axis, thereby removing the individual differences in movement intensity. Following prior work using wearable sensor data to study conversation dynamics [58], [59], [60], we compute the following features using the z-scores: the raw and absolute values for 3 axes each, and the euclidean norm of the raw values across axes, resulting in a total of 7 feature channels. Further, similar to [60], using a sliding-window filter, we denoise the feature channels by extracting statistical (mean, median and variance) and spectral features (log-bin values of power spectral density) from the respective sliding-windows. Drawing inspiration from [61], we also include features that are not preprocessed to circumvent any data loss from preprocessing. An analysis is also presented to understand their respective benefits (see Section 6.2.1).

5.2 Feature Extraction

5.2.1 Individual and Pairwise Features

We consider *pair-wise* bodily coordination features and *individual-level* turn-taking features to study PCQ. For bodily coordination, we extract three sets of features: *synchrony*, *convergence*, and *causality*. An overview of the individual and pairwise features extracted can be seen in Table 1.

Synchrony. Synchrony estimates the dynamic and reciprocal adaptation of the temporal structure of behaviors between interlocutors [62]. Following existing literature [11], [28], [60], we extract four unique measures of interpersonal synchrony: *Correlation*, *Time Lagged Correlation*, *Mutual Information*, and *Mimicry*. See Supplementary Section 2.1, available online, for feature extraction details.

Causality. Correlation does not adequately capture the causal effect [63]. We therefore extract two causality features:

TABLE 1
An Overview of the Four Sets of Individual- and Pair- Level Behavioral Features Extracted

	Attribute Category	Attribute Variant
Synchrony		
1	Correlation	correlation coefficient (ρ_{xy})
2	Time-lagged Correlation	min, max, argmin, argmax
3	Mutual Information	min, max, mean, variance
4	Mimicry	lag_min, lag_max, lag_mean, lag_variance, lead_min, lead_max, lead_mean, lead_variance
Causality		
5	Coherence	min, max
6	Granger's Causality	f_value
Convergence		
7	Symmetric Convergence	ρ
8	Asymmetric Convergence	lag, lead
9	Global Convergence	$d_1 - d_2$
Turn-Taking		
10	Conversation Equality	degree of equality
11	Conversation Fluency	percentage of silence, # back-channels
12	Conversation Syncrhonization	percentage of overlap, # successful interrupts, # unsuccessful interrupts

Coherence [64] and *Granger's Causality*. See Supplementary Section 2.2, available online, for feature extraction details.

Convergence. These features capture the increasing similarity between interacting partners over time [65], and have been shown to be predictive of mutual liking, attraction [60], [66], and social cohesion [28]. In this research, we use three unique estimates of convergence: *Symmetric Convergence*, *Asymmetric Convergence*, and *Global Convergence*. See Supplementary Section 2.3, available online, for feature extraction details.

Turn-Taking. MnM provides binary speaking status of participants annotated from video data. We extract turn-taking features using these annotations by assuming a speaking turn to be a continuous speaking activity segment separated by at least 500 ms of silence [16], [67]. Following existing literature [16], [27], [67], we extracted turn-taking features under three categories: *Conversation Equality*, *Conversation Fluency*, and, *Conversation synchronization*. Assuming a conversation of duration T and a group of N people, and denoting the i-th individual's binary speaking status as $\boldsymbol{s}^i = [s_1^i, \ldots, s_T^i]$, we have the percentage of speaking duration for i, $d^i_{\text{speak}} = (\sum_{t\in[T]} \boldsymbol{s}^i_t)/T$. The degree of equality for i is $eq^i = (d^i_{\text{speak}} - \bar{d})/\bar{d}$, where $\bar{d} = (\sum_{i\in[N]} d^i_{\text{speak}})/N$. As measures of fluency, we compute the percentage of individual silence $d^i_{\text{silence}} = 1 - d^i_{\text{speak}}$ and the number of back-channels (very short utterances of duration up to 2 s). As a measure of Syncrhonization, we consider the percentage of speech overlap, which is $d^i_{\text{o}} = (\sum_{t\in[T]} \mathbb{1}\{s^i_t = s^{j,j\neq i}_t\})/T$ for individual i, and the number of successful and unsuccessful interruptions, which are overlap durations when *turn-change* occurs and does not occur, respectively.

TABLE 2
Overview of the Statistical Analysis Performed

Dependent Variables	Independent Variable Sets	Statistical Models
IndivPCQ	Group cardinality	QLS Regression
GroupPCQ	Turn-taking, Bodily Coordination	LASSO Regression

5.2.2 Group-Level Features

Following [28], [30], we translate individual and pairwise features to group-level features using the feature aggregates *minimum*, *maximum*, *mean*, *mode*, *median* and *variance*. Specifically, for individual-level modeling, similar to Müller et al. [5] we aggregate over pairwise features involving that particular individual, and for group-level modeling aggregation is done over all the pairs in the group.

5.3 Experimental Setup

5.3.1 Statistical Analysis

We perform hypothesis-driven tests to study the effect of (i) group cardinality, (ii) turn-taking attributes and (iii) body coordination attributes on PCQ. We use the Quantile Least Squares (QLS) and Joint LASSO models for our hypothesis-driven analysis. The QLS analysis considers each set of behavioral features independently, while the Joint LASSO analysis accounts for the combined effect of all features, allowing for complementary insight. Due to the superior performance of models when no preprocessing was used (empirically explained in Section 6.2.1), for the statistical analysis tests, we only used the features without preprocessing.

Quantile Least Squares. QLS fits the regression to the conditional *median* of the dependent variable, in contrast to the conditional *mean* estimated by Ordinary Least Squares (OLS). Intuitively, the conditional median is more robust against outliers. More crucially, the QLS does not require the data to abide the assumptions of exogeneity and homoscedasticity like the OLS does. We find that the variance of the independent variables varies largely across quantiles (see Supplementary Fig. S5 for scatter-plots, available online), thereby violating the exogeneity and homoscedasticity assumptions. We therefore use the QLS model for our analysis.

Joint LASSO. While QLS is convenient in situations where classical parametric assumptions do not hold, it still suffers from effects of multicollinearity. We therefore use the QLS model to only study behavioral feature sets in isolation. However, to also account for the combined effect of feature sets, we perform a *joint* regression over all features using a LASSO model, which uses the *coordinate descent* [68] to fit the coefficients, thereby inducing sparsity to address multicollinearity. Subsequently, we perform a post-hoc Spearman's rank correlation on the LASSO filtered features.

An overview of the statistical tests performed can be seen in Table 2. We denote individual- and group- level PCQ as IndivPCQ and GroupPCQ respectively. In total, with two dependent variables, three sets of independent variables and three statistical models, 18 tests were performed. Bonferroni correction is applied to the p-values to correct for

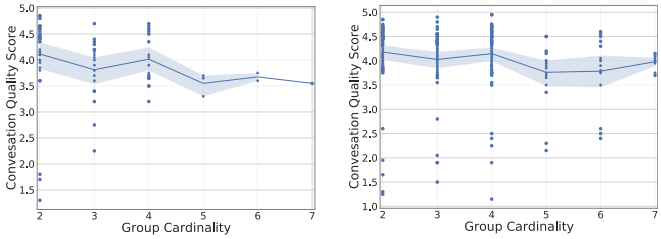


Fig. 8. GroupPCQ and IndivPCQ across group cardinalities.

multiple testing for each dependent variable. After Bonferroni correction a significance threshold of 0.005 was used for testing significance in all the analyses presented.

5.3.2 Analysis of Feature Extraction and Fusion

We perform data-driven analyses to study the effects of (i) window sizes for data preprocessing; (ii) fusion of attribute categories; and (iii) feature aggregators to compute group-level features from individual and pairwise features.

For these analyses, we treat predicting PCQ as a binary classification of low and high PCQ scores. A threshold of 3.0 (on the 5-point scale) is used to binarize the scores into low and high. As such, our annotations suffer from class imbalance, see Fig. 7 for the label threshold, and Supplementary Section 3, available online, for the class distribution. To address this, we employ the Synthetic Minority Oversampling technique (SMOTE) [69], which generates synthetic samples from the minority class. We use a logistic regression model trained with the elastic loss that combines the L_1 and L_2 penalties of the lasso and ridge regularization methods. Specifically, for each experiment we evaluate how the feature extraction or aggregation affects the predictive capability of the model. For dimensionality reduction, we perform PCA on the z-score standardized features, by selecting features that preserve the top 90% of variance in respective predictive tasks. As the performance metric, we use the area under the ROC Curve (AUC) score. The metric is calculated as the average across 5-folds in the cross-validation (CV) setting. A stratified k-fold CV was used to preserve the percentage of samples of each target class as the complete set, in the train and test partitions. Except when studying the effects of preprocessing, in all other experiments only features that are not pre-processed were used. Code for all experiments and analyses are available at https://github.com/LRNavin/conversation_quality.

6 Results

6.1 Statistical Analysis

6.1.1 Analysis of Group Cardinality

Existing research [70], [71], [72] has shown that behavior in group interactions varies with size of the group (group cardinality). Is this true for PCQ as well? We test the hypothesis:

> For an FCG, the PCQ changes with group cardinality.

From the plots in Fig. 8, we see that for both GroupPCQ and IndivPCQ the means for cardinalities of 2, 3 and 4 are higher than that of 5, 6, 7. The statistical tests reveal that IndivPCQ and GroupPCQ are significantly different across groups of different cardinality. We note that for all regression models, the β coefficient for the group cardinality variable is negative, suggesting that PCQ is inversely proportional to group cardinality. For example, the QLS model associates the cardinality attribute with $\beta = -0.2167$ and $\beta = -0.0833$ for IndivPCQ and GroupPCQ respectively (p-value=10^{-5}), indicating that people appear to have better quality conversations with fewer partners.

Post-hoc analysis testing for the differences in PCQ between cardinality pairs reveals that the IndivPCQ scores are significantly different in dyadic group interactions when compared to that of interactions in larger groups (cardinality ≥ 3). One possible alternate explanation of this result is that raters score PCQ more conservatively when there are more partners to pay attention to. Nevertheless, even if this were the case, it would be a valid characteristic of how people perceive behaviors in larger groups. Significant results were not observed for the post-hoc GroupPCQ comparisons, suggesting that no conclusions can be drawn with respect to GroupPCQ regarding pairwise differences with cardinalities. Note that this result should also be interpreted accounting for the small sample size for cardinalities ≥ 5.

6.1.2 Analysis of Turn-Taking Attributes

Turn-taking features have shown to be indicative of constructs such as enjoyment and cohesion [16], [27], [73]. We test the hypotheses:

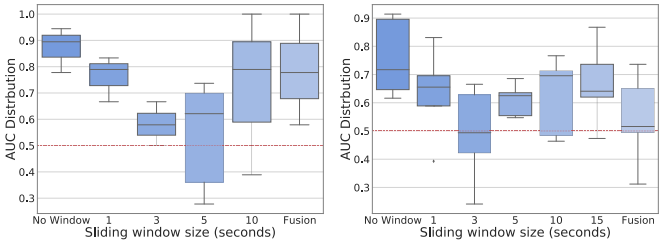
> In an FCG, turn-taking attributes (conversation equality, conversation fluency and conversation synchronization) are positively correlated with PCQ.

For IndivPCQ, the QLS model reveals that conversation equality and percentage of silence are the most significant attributes, with positive ($\beta = 0.2136, p = 10^{-4}$) and negative ($\beta = -0.5094, p = 10^{-4}$) correlations respectively. For GroupPCQ, QLS reveals that the number of successful and unsuccessful interruptions are the most significant attributes, with negative ($\beta = -0.0859, p = 0.001$) and positive ($\beta = 0.0956, p = 0.002$) correlations respectively. On the other hand, the LASSO and rank correlation models reveal a different set of significant attributes. For IndivPCQ, along with conversation equality and percentage of silence, the two interruption based attributes were also revealed to be significant. Similarly, for GroupPCQ, unlike the QLS, the two interruption attributes are found to be insignificant, while conversation equality, percentage of silence and number of backchannel attributes are found to be significant.

Intuitively, the result implies that observers consider group conversations with more equitable speaking turns and fewer interruptions to be of higher quality. An important thing to note here is that the complementary models associate all attributes with similar trends even though they differ on which attributes they consider to be of statistical significance. Even though the statistical significance of successful and unsuccessful interruptions differ when considered in isolation or jointly with other features, they are associated with negative and positive β's respectively, by all models tested.

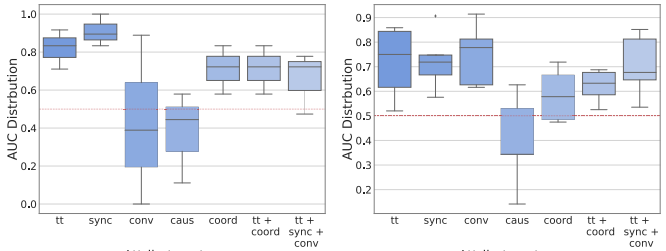
6.1.3 Analysis of Bodily Coordination Attributes

Coordination features across modalities such as bodily movements [60] and paralinguistic speech features [28] have been shown to be indicative of liking[60], attraction [60], and cohesion[28]. Here we test the hypothesis:



(a) GroupPCQ results (b) IndivPCQ results

Fig. 9. Results of the experiments on the predictive capabilities of different window-sizes.



(a) GroupPCQ results (b) IndivPCQ results

Fig. 10. Predictive performance of different feature fusion approaches. Attribute category and indices as in Table 1—*tt*: Turn-taking (10-12), *sync*: Synchrony (1-4), *caus*: Causality (5-6), *conv*: Convergence (7-9), *coord*: Bodily Coordination (1-9).

> In an FCG, bodily coordination features (synchrony, convergence, mimicry, and causality) are positively correlated with PCQ.

For the synchrony attributes, for both IndivPCQ and GroupPCQ we find that the *argmax* and *argmin* variants of lagged correlations are statistically significant attributes ($p = 0.003$). This suggests that the time taken to achieve maximum or minimum synchronous coordination has a significant effect on the conversation quality. We also note that for GroupPCQ, only correlation based features from the synchrony category were statistically significant, while other attribute sets (convergence and causality) were found to be statistically insignificant. For IndivPCQ, the *minimum* and *variance* of the convergence attributes were all statistically significant. This suggests that attributes capturing the least converging interacting pairs in a group are relevant to external observers. Moreover, we note that the minimum of the attributes are positively correlated, while the variance are negatively correlated. Further, the *maximum* and *minimum* of the lagged mimicry attributes were also statistically significant attributes. This suggests that pairs with high and low mimicry are relevant for estimating individual experience.

The Joint LASSO results indicate that several other feature sets also have a significant effect on IndivPCQ. Along with the *min*, *max*, *argmin*, and *argmax* attributes of the lagged correlation features, the non-lagged correlation were also significant. Moreover, the post-hoc rank correlation analysis associates different coefficient signs for some of the significant features. For example, lagged mimicry attributes are given negative β's by the rank correlation model but positive β's by LASSO. This suggests that there exists a non-linear monotonic relationships between these variables and IndivPCQ, causing the LASSO model to fail to explain this relationship, associating them with $\beta \approx 0$. One commonality between the two models is that both consider the *lagged* variant of mimicry features to be of more significance that the *lead* variant. For GroupPCQ, the LASSO and rank correlation analysis reveals that when jointly considered with other bodily coordination features, the lagged mimicry and convergence attributes are statistically significant.

6.2 Analysis of Feature Extraction and Fusion

6.2.1 Influence of Window Sizes

During data preprocessing we extract statistical and spectral features from the accelerometer data using the commonly used sliding window approach [58], [59], [60]. The choice of window-size influences a trade-off between noise-reduction and information loss. To understand the effect of this choice, we extract features using different window-sizes and evaluate the resulting change in the logistic regression model's predictive capability. The results are presented in Fig. 9 for respective sliding window-sizes, along with the fusion of features from all the window-sizes, denoted as "*Fusion*".

From Fig. 9, we see that the best performing features are the ones where no sliding-window technique was used for both GroupPCQ and IndivPCQ. This suggests that the smoothing of accelerometer readings results in a loss of information which hurts model performance. The results might also indicate that bodily coordination between interacting pairs occur at finer temporal granularity, which can be captured directly without the sliding-window approach. The model with no sliding-window based features is capable of predicting GroupPCQ with a mean AUC of 0.85 ± 0.07 and IndivPCQ with a mean AUC of 0.76 ± 0.13. Also, noting here that using no sliding-window achieves the least standard deviation in AUC scores.

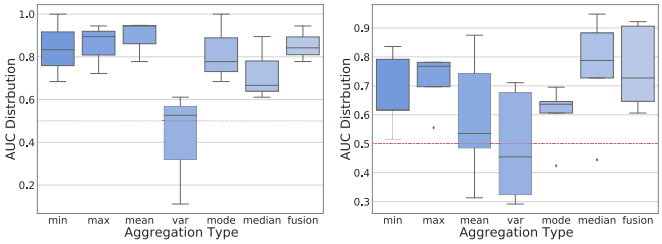
6.2.2 Influence of Fusing Attribute Categories

Here we study the influence of fusing different attribute categories on the performance of the logistic regression.

From the GroupPCQ results in Fig. 10a, we see that the synchrony attributes (mean AUC of 0.89 ± 0.04) and turn-taking attributes (mean AUC of 0.81 ± 0.06), are the best performing attributes. In contrast to the IndivPCQ results in Fig. 10b, the convergence attributes do not predict GroupPCQ well. Moreover, unlike for IndivPCQ, fusing turn-taking attributes with synchrony and convergence attributes does not improve GroupPCQ prediction, both interms of mean and variance AUC. From the IndivPCQ analysis, we see that convergence (mean AUC of 0.75 ± 0.12) and synchrony (mean AUC of 0.72 ± 0.12) based attributes perform well both by themselves and after feature-level fusion (mean AUC of 0.60 ± 0.10). We also observe that although turn-taking attributes are one of the best performing feature sets by themselves (mean AUC of 0.72 ± 0.15), fusing them with bodily coordination attributes reduces the standard deviation of AUC, 0.70 ± 0.09. The results also suggest that synchrony and convergence attributes are best predictors of IndivPCQ, both individually and fused.

6.2.3 Influence of Feature Aggregators

The last step of our feature extraction procedure is to use aggregators to combine pairwise features into group-level



(a) GroupPCQ results (b) IndivPCQ results

Fig. 11. Predictive performance of feature aggregators.

features, or aggregate over pairs containing an individual for individual-level modeling, following previous works [5], [28], [30]. Here we study how different aggregators affect the predictive performance of the logistic regression model.

From Fig. 11, we see that the the *mean* aggregation of the features performs the best with a mean AUC of 0.89 ± 0.08. The mean is a skewed average. In contrast, for IndivPCQ the unskewed average, the *median*, is the most informative, with an AUC of 0.78 ± 0.17. This is in line with inferences drawn by Nanninga et al. [28] while studying cohesion in meetings. For both IndivPCQ and GroupPCQ, the *variance* aggregator performs worst.

7 Discussion and Conclusion

In this work, we have conceptualized, validated, and analyzed a perceived measure of conversation quality by unifying overlapping constructs that have so far been largely studied in isolation in literature. While our core motivation has been to gain insight into how people perceive the individual and group experiences of others, we do not claim that our proposed method measures, or is meant to be a third-party proxy for, the *one true experience* of the individual or group in the scene. On the contrary, we suggest that these perceptions are indicative of empathized gestalt impressions people draw of others' experience as it unfolds. We argue that such a perceived measure should complement other self-reported measures of experience to gain richer insight into how these differ and identify the contextual factors that influence the perceptions.

Third-party ratings are always prone to be influenced by biases that are heavily embedded in our cultures. We recommend users of this research to be mindful that third-party perceptions are not the same as self-reported measures. This fundamentally influences the system design process. The motivation for taking a third-party perspective is to enable a study of whether such perceptions have agreement, and whether samples with high agreement have common behavioral manifestations. To develop systems for inferring an individual's actual social experience, we advocate for a participant-in-the-loop strategy that allows for the measuring of the actual experience while being mindful of the participants' consent.

Inter-rater agreement and annotation drift are important aspects to consider while collecting annotations for behavioral data. Annotation drift is an issue when the annotator's mental model of the measured phenomenon changes over time while the phenomenon remains constant. Accounting for drift is crucial when the annotation is used as an attribute of the underlying phenomenon rather than as an attribute of the third-party observer. This is the case for annotating phenomena such as facial action units, where the goal of the annotation is to represent the configuration of a person's facial muscles rather than the annotator's perception of it, so a systematic drift over time or disagreement amongst annotators is undesirable.

For a perceived measure like the one we are proposing, the central phenomenon being studied is an onlooker's perception. So, every perception is inherently valid. This argument is based on our understanding that the measure requires some projection of one's own experience onto the observed subjects when trying to empathize with their situation or take their perspective. Following the assumption that we construct narratives of other's behaviors, and that our appraisal of a situation is constructed based on our experiences, any drift occurring because of variations in one's experience can only provide (another) valid perspective on how the observed subject might be feeling. The same is true for variations in annotator agreement resulting from differences in perception of the annotators, either resulting from transient factors such as mood, or relatively stable factors such as personality and cultural background. For a perceived measure, we view all such perceptions as valid.

Designing the instrument to remove such variations would amount to artificially tampering with the phenomenon being measured. In our experiments we remove data with low inter-annotator agreement from the evaluation. However, this is because by design, the goal of the experiments is to gain insight into behavioral features that correlate with a high agreement on PCQ across raters. More broadly, we view the presence of low agreement on certain samples as a motivation for future work to explore more appropriate ways to embed subjectivity into the learning process when the goal is to train machine learning systems. Note that omitting the samples with low agreement from our experiments does not detract the validity of our measure. When the goal is to measure conversation quality as experienced by the individual or group in the scene, or even to use the third-party annotations as a proxy for the true experienced quality, we suggest treating the considerations of annotation drift and inter-rater agreement with care.

7.1 Limitations and Future Avenues

The data analyzed here was from spontaneous interactions in a single setting, that of mingling interactions following a speed-dating event. So, our findings pertaining to the individual features being indicative of PCQ ought to be interpreted within the scope of such a social context rather than being reflective of social behavior in all spontaneous interactions. As dedicated techniques for the non-invasive recording in-the-wild spontaneous interactions [74] continue to advance, it would be interesting to compare the effects of different social settings on the perception of PCQ using our proposed instrument.

Our operationalization of a conversing group follows the widely used framework F-formation [46]. However, recent evidence suggests that there might be multiple simultaneous conversations within a single F-formation containing more than four participants [72]. It would therefore also be interesting for future work to study PCQ within a single conversation floor rather than for the whole F-formation.

Finally, we have used three raters in this work to obtain our annotations. It would be useful for future works to use the proposed instrument to investigate systematic differences in perceptions of conversation quality across different cultures and demographics at scale.

ACKNOWLEDGMENTS

We thank Swathi Yogesh, Divya Suresh Babu, and Nakul Ramachandran for their time and patience in annotating the dataset, and Tiffany Matej Hrkalovic and Amelia Villegas Morcillo for the insightful discussions. Chirag Raman and Navin Raj Prabhu contributed equally to this work.

REFERENCES

[1] A. Milek et al., ""Eavesdropping on happiness" revisited: A pooled, multisample replication of the association between life satisfaction and observed daily conversation quantity and quality," *Psychol. Sci.*, vol. 29, no. 9, pp. 1451–1462, 2018.

[2] M. R. Mehl, S. Vazire, S. E. Holleran, and C. S. Clark, "Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations," *Psychol. Sci.*, vol. 21, no. 4, pp. 539–541, 2010.

[3] A. See, S. Roller, D. Kiela, and J. Weston, "What makes a good conversation? How controllable attributes affect human judgments," in *Proc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 1702–1723.

[4] E. Dinan et al., "The second conversational intelligence challenge (ConvAI2)," in *Proc. Int. Conf. Neural Inf. Process. Syst. Competition*, 2020, pp. 187–208.

[5] P. Müller, M. X. Huang, and A. Bulling, "Detecting low rapport during natural interactions in small groups from non-verbal behaviour," in *Proc. Int. Conf. Intell. User Interfaces*, 2018, pp. 153–164.

[6] N. Jaques, D. McDuff, Y. L. Kim, and R. Picard, "Understanding and predicting bonding in conversations using thin slices of facial expressions and body language," in *Proc. Int. Conf. Intell. Virtual Agents*, 2016, pp. 64–74.

[7] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2005, pp. I/489–I/492.

[8] C. Oertel, C. De Looze, S. Scherer, A. Windmann, P. Wagner, and N. Campbell, "Towards the automatic detection of involvement in conversation," in *Analysis of Verbal and Nonverbal Communication and Enactment*. Berlin, Germany: Springer, 2011, pp. 163–170.

[9] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 113–130, First Quarter 2021.

[10] C. Raman, J. Vargas-Quiros, S. Tan, E. Gedik, A. Islam, and H. Hung, "ConfLab: A rich multimodal multisensor dataset of free-standing social interactions in-the-wild," 2022, *arXiv:2205.05177*.

[11] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *Proc. IEEE Int. Conf. Privacy Secur. Risk Trust Social Comput.*, 2011, pp. 613–616.

[12] R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "Disagreeables"," *J. Pers. Social Psychol.*, vol. 97, no. 4, pp. 667–684, 2009.

[13] D. A. Northrup, *The Problem of the Self-Report in Survey Research*. New York, NY, USA: Inst. Social Res., York Univ., 1997.

[14] J. Garcia and A. R. Gustavson, "The science of self-report," *APS Observer*, vol. 10, no. 1, 1997.

[15] L. J. R. Norman, M. Bradburn, and S. K. Shevell, "Answering autobiographical questions: The impact of memory and inference on surveys," *Science*, vol. 236, pp. 157–167, 1987.

[16] S. E. Lindley and A. F. Monk, "Measuring social behaviour as an indicator of experience," *Behav. Inf. Technol.*, vol. 32, pp. 968–985, Oct. 2013.

[17] N. R. Prabhu, C. Raman, and H. Hung, "Defining and quantifying conversation quality in spontaneous interactions," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 196–205.

[18] D. Reitter, J. D. Moore, and F. Keller, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proc. 28th Annu. Conf. Cogn. Sci. Soc.*, 2006, pp. 685–690.

[19] C. Oertel, S. Scherer, and N. Campbell, "On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1541–1544.

[20] D. Wyatt, T. Choudhury, and H. Kautz, "Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, pp. IV-213–IV-216.

[21] J. H. Antil, "Conceptualization and operationalization of involvement," *Adv. Consum. Res.*, vol. 11, no. 1, 1984.

[22] J. C.-Y. Hsiao, W.-R. Jih, and J. Y.-J. Hsu, "Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns," in *Proc. Workshop AAAI Conf. Artif. Intell.*, 2012, pp. 40–43.

[23] F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe, "Dyad rapport and the accuracy of its judgment across situations: A lens model analysis," *J. Pers. Social Psychol.*, vol. 71, no. 1, 1996, Art. no. 110.

[24] A. O. Horvath and L. S. Greenberg, "Development and validation of the working alliance inventory," *J. Counseling Psychol.*, vol. 36, no. 2, 1989, Art. no. 223.

[25] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," in *Proc. Int. Conf. Multimodal Interact.*, 2013, pp. 99–106.

[26] M. Casey-Campbell and M. Martens, "Sticking it all together: A critical assessment of the group cohesion–performance literature," *Int. J. Manage. Rev.*, vol. 11, pp. 223–246, 2009.

[27] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behaviour," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 563–575, Oct. 2010.

[28] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 206–215.

[29] Y. Zhang, J. Olenick, C.-H. Chang, S. W. J. Kozlowski, and H. Hung, "The I in team: Mining personal social interaction routine with topic models from long-term team data," in *Proc. 23rd Int. Conf. Intell. User Interfaces*, 2018, pp. 421–426.

[30] Y. Zhang et al., "TeamSense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," *Proc. Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, 2018, Art. no. 150.

[31] A. Cerekovic, O. Aran, and D. Gatica-Perez, "How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits," in *Proc. Int. Workshop Hum. Behav. Understanding*, 2014, pp. 1–15.

[32] C. Edelsky, "Who's got the floor?," *Lang. Soc.*, vol. 10, no. 3, pp. 383–421, 1981.

[33] J. Coates, "Gossip revisited: Language in all-female groups," in *Women in Their Speech Communities*. White Plains, NY, USA: Longman, 1989.

[34] M. Dunne and S. H. Ng, "Simultaneous speech in small group conversation: All-together-now and one-at-a-time?," *J. Lang. Social Psychol.*, vol. 13, no. 1, pp. 45–71, 1994.

[35] D. Tannen, *Conversational Style: Analyzing Talk Among Friends*. London, U.K.: Oxford Univ. Press, 2005.

[36] A. F. Monk and D. J. Reed, "Telephone conferences for fun: Experimentation in people's homes," *Int. Conf. Home-Oriented Inform. Telematics*, 2007, pp. 201–214.

[37] J. Carletta, S. Garrod, and H. Fraser-Krauss, "Placement of authority and communication patterns in workplace groups: The consequences for innovation," *Small Group Res.*, vol. 29, no. 5, pp. 531–559, 1998.

[38] O. Daly-Jones, A. Monk, and L. Watts, "Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus," *Int. J. Hum.-Comput. Stud.*, vol. 49, no. 1, pp. 21–58, 1998.

[39] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 111, no. 2, 1992, Art. no. 256.

[40] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness," *J. Pers. Social Psychol.*, vol. 64, no. 3, 1993, Art. no. 431.

[41] F. J. Bernieri, J. M. Davis, R. Rosenthal, and C. R. Knee, "Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect," *Pers. Social Psychol. Bull.*, vol. 20, no. 3, pp. 303–311, 1994.

[42] C. Raman, H. Hung, and M. Loog, "Social processes: Self-supervised meta-learning over conversational groups for forecasting nonverbal social cues," 2021, *arXiv:2107.13576*.
[43] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 113–130, First Quarter 2021.
[44] C. Lai and G. Murray, "Predicting group satisfaction in meeting discussions," in *Proc. Workshop Model. Cogn. Processes Multimodal Data*, 2018, Art. no. 1.
[45] S. W. J. Kozlowski and K. J. Klein, "A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes," in *Multilevel Theory, Research and Methods in Organizations: Foundations, Extensions, and New Directions*. Hoboken, NJ, USA: Wiley, 2000, pp. 3–90.
[46] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, vol. 7, Cambridge, U.K.: Cambridge University Press , 1990.
[47] N. A. Murphy et al., "Reliability and validity of nonverbal thin slices in social interactions," *Pers. Social Psychol. Bull.*, vol. 41, no. 2, pp. 199–213, 2015.
[48] M. Z. Wang, K. Chen, and J. A. Hall, "Predictive validity of thin slices of verbal and nonverbal behaviors: Comparison of slice lengths and rating methodologies," *J. Nonverbal Behav.*, vol. 45, pp. 53–66, 2020.
[49] N. A. Murphy et al., "Predictive validity of thin-slice nonverbal behavior from social interactions," *Pers. Social Psychol. Bull.*, vol. 45, pp. 983–993, 2018.
[50] W. Brinkman, *Design of a Questionnaire Instrument*. Commack, NY, USA: Nova Publishers, 2009, pp. 31–57.
[51] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests," *Psychol. Bull.*, vol. 52, no. 4, 1955, Art. no. 281.
[52] D. C. Funder and C. D. Sneed, "Behavioral manifestations of personality: An ecological approach to judgmental accuracy," *J. Pers. Social Psychol.*, vol. 64, no. 3, 1993, Art. no. 479.
[53] D. J. Mundfrom, D. G. Shaw, and T. L. Ke, "Minimum sample size recommendations for conducting factor analyses," *Int. J. Testing*, vol. 5, no. 2, pp. 159–168, 2005.
[54] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 501–513, Mar. 2009.
[55] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychol. Bull.*, vol. 70, no. 4, 1968, Art. no. 213.
[56] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE 10th Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.
[57] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.
[58] H. Hung, G. Englebienne, and J. Kools, "Classifying social actions with a single accelerometer," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 207–210.
[59] E. Gedik and H. Hung, "Speaking status detection from body movements using transductive parameter transfer," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 69–72.
[60] Ö. Kapcak, J. Vargas-Quiros, and H. Hung, "Estimating romantic, social, and sexual attraction by quantifying bodily coordination using wearable sensors," in *Proc. IEEE 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, 2019, pp. 154–160.
[61] E. Gedik and H. Hung, "Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, 2018, Art. no. 163.
[62] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affective Comput.*, vol. 3, no. 3, pp. 349–365, Third Quarter 2012.
[63] J. Aldrich et al., "Correlations genuine and spurious in Pearson and yule," *Stat. Sci.*, vol. 10, pp. 364–376, 1995.
[64] D. C. Richardson and R. Dale, "Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension," *Cogn. Sci.*, vol. 29, no. 6, pp. 1045–1060, 2005.
[65] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 2779–2782.
[66] J. Michalsky and H. Schoormann, "Pitch convergence as an effect of perceived attractiveness and likability," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2253–2256.
[67] C. Lai, J. Carletta, and S. Renals, "Modelling participant affect in meetings with turn-taking features," in *Proc. Workshop Affect. Social Speech Signals*, 2013.
[68] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, pp. 1–22, 2010.
[69] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[70] E. Gedik and H. Hung, "Detecting conversing groups using social dynamics from wearable acceleration: Group size awareness," *Proc. Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 4, 2018, Art. no. 163.
[71] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the Organization of Conversational Interaction*. Amsterdam, The Netherlands: Elsevier, 1978, pp. 7–55.
[72] C. Raman and H. Hung, "Towards automatic estimation of conversation floors within F-formations," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos*, 2019, pp. 175–181.
[73] S. E. Lindley and A. F. Monk, "Social enjoyment with electronic photograph displays: Awareness and control," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 8, pp. 587–604, 2008.
[74] C. Raman, S. Tan, and H. Hung, "A modular approach for synchronized wireless multimodal multisensor data acquisition in highly dynamic social settings," in *Proc. Int. Conf. Multimedia*, 2020, pp. 3586–3594.

Chirag Raman received the BEng degree in information technology from the University of Mumbai, India, in 2010, and the master's degree in entertainment technology from Carnegie Mellon University, USA, in 2013. He is currently working toward the PhD degree with the Socially Perceptive Computing and Pattern Recognition Labs, TU Delft, The Netherlands, since 2018. Between 2013 and 2018, he worked as a research associate with Disney Research, a Lead iOS and UX developer with ProductionPro, and a senior research engineer with the Language Technologies Institute, Carnegie Mellon University. His research interests include multimodal machine learning, generative modeling, affective computing, computer vision, and computer graphics.

Navin Raj Prabhu (Member, IEEE) received the BTech degree in computer science from SRM University, India, in 2015, and the MS degree in computer science from the Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands, in 2020. Currently, he is working toward the PhD degree with the Signal Processing Lab and Organisation Psychology Lab, University of Hamburg, Hamburg, Germany. His research interests include affective computing, social signal processing, deep learning, uncertainty modelling, speech signal processing, and group affect.

Hayley Hung (Member, IEEE) received the PhD degree in computer vision from the Queen Mary University of London, in 2007. She is an associate professor with the Socially Perceptive Computing Lab, TU Delft, The Netherlands, where she works since 2013. Between 2010-2013, she held a Marie Curie fellowship with the Intelligent Systems Lab, University of Amsterdam. Between 2007-2010, she was a post-doctoral researcher with IDIAP Research Institute in Switzerland. Her research interests include social computing, social signal processing, computer vision, and machine learning.