IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis

LUCIEN MAMAN[1], ELEONORA CECCALDI[2], NALE LEHMANN-WILLENBROCK[3], LAURENCE LIKFORMAN-SULEM[1], (Senior Member, IEEE), MOHAMED CHETOUANI[4], GUALTIERO VOLPE[2], GIOVANNA VARNI[1]

[1]LTCI, Télécom Paris, Institut polytechnique de Paris, Palaiseau, 91120, France
[2]Casa Paganini - InfoMus, DIBRIS, University of Genoa, Italy
[3]Department of Industrial/Organizational Psychology, University of Hamburg, Germany
[4]Institute for Intelligent Systems and Robotics, Sorbonne University, CNRS UMR7222, Paris, France

Corresponding author: Lucien Maman (lucien.maman@telecom-paris.fr)

This paper has been partially supported by the French National Agency (ANR) in the frame of its Technological Research JCJC program (GRACE, project ANR-18-CE33-0003-01, funded under the Artificial Intelligence Plan).

ABSTRACT This paper presents *GAME-ON* (*Group Analysis of Multimodal Expression of cohesiON*), a multimodal dataset specifically designed for studying group cohesion and for explicitly controlling its variation over time. Cohesion is here addressed according to the Severt's theoretical multidimensional integrative framework. More specifically, GAME-ON focuses on the social and task dimensions of the instrumental function of cohesion. The dataset consists of over 11 hours of synchronized multimodal recordings (audio, video, and motion capture data) of 17 small groups (3 persons) playing a social game, i.e., an escape game. The game consists of several tasks designed to manipulate the variation of cohesion over time. GAME-ON includes annotations consisting of self-assessment of cohesion and other constructs such as emotions, leadership, and warmth and competence. A first statistical analysis of these annotations shows that we successfully manipulated all the relative variations of cohesion (between tasks) over time. This holds for all tasks except for one where we observed a significant variation of cohesion in the opposite direction than expected. The dataset will be publicly available for research purposes. The motivation of our work is to provide the scientific community with an asset for studying cohesion and other group phenomena.

INDEX TERMS Cohesion, Group interaction analysis, Multimodal dataset, Social Signal Processing

I. INTRODUCTION

SOCIAL signal processing (SSP) is a multidisciplinary research domain aimed at enabling machines to sense, recognize, and display human social signals, that is the multimodal expression of attitudes towards social contexts [1]. To date, one of the most challenging tasks addressed by SSP is automated group interaction analysis. Analyzing group behavior entails both technological and social difficulties due to the patchwork of simultaneous one-to-one and one-to-many interactions that establish and evolve over time. As group members explicitly and implicitly interact to coordinate their actions and achieve objectives, so-called emergent group states develop over time. These states are social processes that result from the micro-level affective, behavioral and cognitive interactions among group members, through the micro-processes of group interaction (e.g., [2], [3]). Emergent states include pivotal group phenomena such as group trust, conflict, leadership, transactive memory system, and cohesion and they are an important aspect in modelling the dynamic process of group problem solving [4]. Studies about emergent states cover a broad range of domains and group contexts such as sports, army or business. Emergent states have been consistently demonstrated to influence desirable group outcomes such as group effectiveness and performance [5]–[7]. However, the definition of group emergent states as phenomena that originate in dynamic group interactions and dynamically evolve and change over time makes them notoriously difficult to capture. Advances in SSP have the potential to address this problem in the broader literature, especially when they embrace interdisciplinary collaborations in order to advance our understanding of dynamic group processes [8]. Yet, to the best of our knowledge, no previous dataset in the SSP domain has explicitly focused on emergent group states and their underlying dynamics, despite the relevance of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

emergent states for team science and practice. Whereas most existing SSP datasets entail spontaneous or scripted group interactions, they do not target a specific emergent group state. A possible exception concerns the ELEA dataset addressing emergent leadership in groups [9]. ELEA, however, did not refer to emergent group states but rather focused on the emergence of individual leaders in group interactions.

Our contribution to group interaction analysis is GAME-ON (Group Analysis of Multimodal Expression of cohesiON), a multimodal dataset designed *ad hoc* to address group cohesion and to control its dynamics. With dynamics, we mean the variation of cohesion over the time of the data collection, i.e., its increase or decrease between one task the data collection consists of and the next one. We focus on cohesion because this phenomenon has received more scholarly attention than any other emergent group state [10]. The GAME-ON dataset consists of multimodal (audio, video, and motion capture data) synchronized recordings of small groups (3 persons) playing an escape game, that is a game where the players, in a limited amount of time, have to escape a room by collaborating and solving puzzles and other tasks. The design and implementation of GAME-ON was driven by our motivation to build an interdisciplinary scientific community working on emergent group states and to provide researchers with a *unique* multimodal dataset. Through tight collaboration between computer scientists and psychologists, we developed a rich setup using breakthrough technology in a synchronized way, and a scenario grounded on psychological models of cohesion. This will allow researchers from several communities to use the data and to collaboratively explore research questions and methodological workflows.

The paper is organized as follows. Section 2 reviews theoretical models of cohesion, major behavioral and automated tools to assess it, and existing datasets to investigate group interactions. Section 3 describes the design and technical setup used to record GAME-ON. Section 4 presents a statistical analysis of participants' perceptions of cohesion, leadership, warmth and competence of their group members, and individual emotional states. Conclusions in Section 5 end the paper.

II. BACKGROUND

A. COHESION

Cohesion is one of the most studied emergent states [11], involving both group emotions [12] and goals [13]. Meta-analytic evidence consistently showed positive linkages between cohesion and group performance, leading researchers to focus attention on understanding how to enhance it (see [14] for a recent review). Researchers on Psychology, however, suggested different definitions of cohesion, data collection techniques and methodologies to observe this emergent state over the last century, making it difficult to compare findings across studies and limiting the ability to advance science and practice [14], [15].

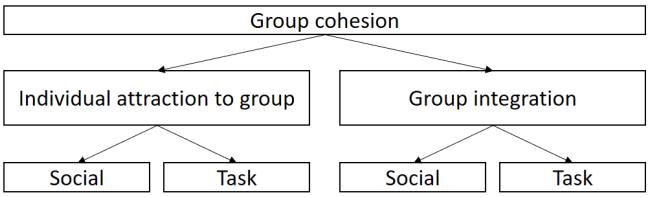


FIGURE 1: Carron's model of cohesion. It has 2 dimensions (*Individual attraction to the group* and *Group integration*) that in turn are expressed in the task and social dimensions.

1) Theoretical models

The first definition of cohesion, given by Lewin in the 1940s under the framework of the field theory [16], referred to it as "*a group characteristic that depends on its size, organization and intimacy*" [17]. Following Lewin's work, Festinger defined cohesion as the "*total field of forces causing members to remain in the group*" [18]. Theses forces pointed to different dimensions of cohesion. However, due to the difficulty to control and measure the impact of each force, researchers continued to consider cohesion as a uni-dimensional construct. Later, researchers started to focus either on the forces related to the social dimension [19] or on those related to the task dimension of cohesion [20], [21]. These studies had a relevant impact on the development of multidimensional models of cohesion that grounded and refined these forces as 2 distinct dimensions: social and task cohesion.

Since the 1980s, the idea that cohesion is a multidimensional construct is well accepted. Carron was among the first to propose a multidimensional model of cohesion [22] (see Figure 1) that was adopted by many scholars as the reference model to describe cohesion. This model comprises 2 major dimensions: *Individual attraction to the group* and *Group integration*. Individual attraction to the group represents all the reasons that would motivate a group member to remain in the group, while group integration represents the degree of unification of the group. Each one of these dimensions can manifest as a task or a social dimension. The task dimension relates to the degree of commitment to group tasks and goals. The social dimension relates to the relationships and friendships between group members. This model had a substantial impact on the research field of cohesion and led to the creation of the Group Environment Questionnaire (GEQ) [22] to measure cohesion.

More recently, Severt *et al.* [23] proposed an integrative framework taking into account Carron's model and other researchers' ideas and improvements (i.e., [24]–[27]). This framework posits that cohesion can be categorized by 2 main functions, an *affective* function and an *instrumental* function. Figure 2 summarizes these key concepts.

The affective function of cohesion refers to all the aspects that highlight the emotional impact on a group member and, by extension, the group as a whole (e.g., behaviors or elements of an interaction such as cooperation or exchange). Severt and his colleagues divided it into 2 dimensions that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

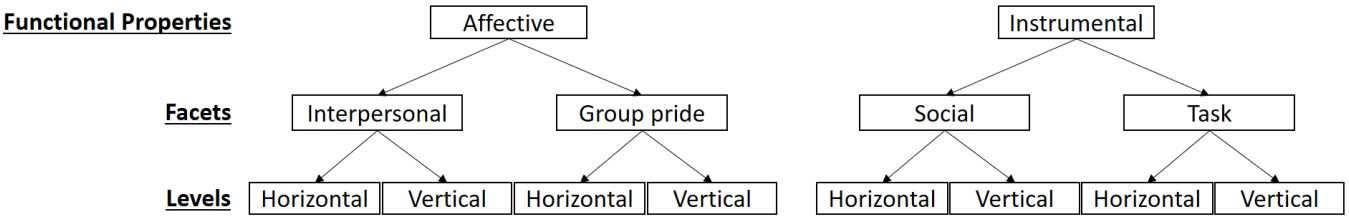


FIGURE 2: Severt's multidimensional integrative framework of cohesion. This is divided into 2 functional properties (*Affective* and *Instrumental*). Each property has 2 dimensions (or facets), which are also divided into 2 levels (*horizontal* and *vertical*).

they refer to as facets. First, the *interpersonal* dimension lies on how much one likes, dislikes, or hates the other group members. It can be viewed as a force acting between people that tends to draw them together and to resist their separation. The second one, the *group pride* dimension, results from a deep sense of belonging to a group as a whole. It creates a sense of community which strengthens the bonds of unity. A group member may be attracted to the group because being part of it is viewed as an honor [20]. This dimension emphasizes the importance that members place on identifying themselves to the group and being part of it [25]. Friendship bonds and the desire to identify to a group are often signals of the emergence of cohesion through its affective dimensions. A group of coworkers going out for an event outside of work hours is an example of the emergence of interpersonal cohesion whilst observing group members wearing group t-shirts is an example of group pride cohesion.

The instrumental function of cohesion refers to "*those aspects that highlight the goal- and task-based activities of the group*" [23]. Following Katz' statement about the instrumental function of cohesion [28], Severt *et al.* suggest that it is the instrumental function of cohesion that "*keeps the group intact so that it can achieve the set goals of the group, all the while maximizing the rewards gained from achieving those goals, and minimizing penalties or losses in the process*" [23]. Within the instrumental function of cohesion, Severt and colleagues distinguish between *social* and *task* cohesion. The social dimension refers to the social bonds between group members that are bound by the group's *working* relationship. It might be counterintuitive to categorize social cohesion as an instrumental function, but social bonds can indeed serve the group's goal. The higher social cohesion will be in a group, the more its members will value the relationships and friendships that the group provides [19], resulting in a positive climate where group members engage in high-quality social working relationships. An example of social cohesion is when group members play board games together during their lunch break.

Task cohesion relates to the degree of commitment to group tasks and goals. It is implied that group members need to share a sufficient level of confidence on the task(s) realization. An example of task cohesion is when a leader supports another group member by creating conditions that will ease the resolution of the task.

For each dimension of the 2 functional properties of cohesion, 2 levels can be distinguished according to hierarchy differences among members: *horizontal* and *vertical*. Horizontal cohesion concerns relations among group members of the same authority level, whereas vertical cohesion implies hierarchy and refers to the relations between a member of authority and a subordinate within the group context. It is important to differentiate these dimensions as cohesion can emerge from relationships among various type of groups and group members and across the entirety of the group's hierarchy. Cohesion also manifests differently according to the dimension and level of measurement.

Based on Severt's framework and Carron's model, we specifically designed our data collection in order to measure instrumental cohesion at a horizontal level. Our choice to focus on instrumental cohesion, with the two sub-dimensions of task and social cohesion, follows theoretical arguments that all groups form for a purpose, and even social groups have an instrumental basis (e.g., forming a social group in order to develop friendships; see [29]). The focus on social and task cohesion as part of the instrumental cohesion framework also aligns with the dominant approach in the current teams literature (e.g., [30]–[32]). Moreover, we decided to study cohesion at the horizontal level in order to have as many participants as possible (i.e., it is easier to find groups of friends than groups with a hierarchy). This decision also improves the applicability of our findings and external validity of our study setting as it corresponds contemporary trends of flattening organizational hierarchies and self-managed teams (e.g., [33]).

2) Methods to assess cohesion

Empirical efforts to assess cohesion began in the early 1950s and continue to this day. The vast majority of previous research on cohesion has relied on surveys and questionnaires, which provide static snapshots of the phenomenon and cannot account for the underlying dynamics (e.g., [12]). Researchers developed a range of questionnaires to measure cohesion in different types of groups. These include the Sport Cohesiveness Questionnaire (SCQ) [34], the Sport-modified Bass Orientation Inventory (SBOI) [35], and the Multidimensional Sport Cohesion Instrument (MSCI) [36]). These tools differ in the number of items (from 3 to 22 items), the types of assessment (self, external or both), and the answering format (either a forced choice or a 5-point

Likert scale). These questionnaires evaluated both task and social aspects of cohesion, but none of them convinced the community and criticisms arose due to the inconsistency in the definition of cohesion and incoherence in the variable measurements [37], [38], making it impossible to compare results across studies. Taking into account these criticisms, more recent studies focused on validating tools to assess and measure cohesion (e.g., [22], [26], [38]–[40]) with limited success (e.g., the Team Climate Questionnaire (TCQ) [41] or the Perceived Cohesion Scale (PCS) [26]).

Following these developments, Carron and colleagues designed the Group Environment Questionnaire (GEQ) [22] to assess cohesion in sport teams. This is an 18-items self-report survey with a 9-point Likert scale answering format. It has been extensively applied and its psychometric properties have been validated by several psychologists and sociologists (e.g., [42]–[46]). Due to its large popularity, the questionnaire was translated into several other languages (e.g., French [47], Arabic [48], and Italian [49]). Moreover, researchers adapted it in order to target other groups than sport teams (e.g., [50], [51]). Carron and Brawley encouraged researchers to modify the GEQ by using the original items, adapted to the target group (i.e., not changing the valence nor the grammatical construction of the items), and by removing inappropriate items [29].

In our data collection, we used the GEQ to assess the social and task dimensions of cohesion. We used the items related to Carron's social aspect of cohesion to measure the social dimension of Severt's framework. Indeed, Carron's definition of social cohesion covers a larger spectrum of behavior than Severt's definition (e.g., aspects of the affective dimensions belong to the social dimension regarding Carron's model).

3) Automated approaches to detect cohesion

Over the last decade, scholars started to focus on how to automatically detect and predict emergent group states, and in particular cohesion. Multiple approaches exist and are based on the work and observations made by sociologists and psychologists.

Some studies attempted to predict and measure cohesion via a unimodal approach. Using a linguistic style matching metric, Gonzales *et al.* proposed a way to predict cohesion and performance of small groups from verbal behavior during face-to-face and text-based computer mediated discussions [52]. Their metric, however, only relies on verbal communication and only takes the task dimension of cohesion into consideration. Giraldo and Passino also investigated task cohesion through the patterns of communication among group members and modeled a human group as "*a dynamical complex system whose dynamics are driven by task optimization and the interaction between subsystems that represent the members of the group interconnected according to a given communication network*" [53]. This model is very interesting because it includes the dynamics of the interactions, but it grounds on the simplistic definition of cohesion developed by Festinger in the 1950s [18].

Ghosh *et al.* proposed methods to automatically predict group cohesiveness in images from the GAF 3.0 dataset, focusing on facial expressions [54]. Their approach achieves near human-level performance in predicting a group's cohesion score, but this mainly concerns perceived cohesion and does not provide any insight on the underlying behavioral dynamics of cohesion.

Most of the recent studies focused on small groups' nonverbal cues, as nonverbal communication has been shown to be a more powerful predictor of group-level cohesion than verbal behavior [55]. Moreover, studies attempting to predict and measure cohesion using a multimodal approach, tend to yield better results than unimodal models.

Among multimodal models of cohesion, Hung and Gatica-Perez were the first to include both audio and video nonverbal descriptors to study cohesion through multiple dimensions in a meeting context [56]. They also collected annotations of cohesion provided by external observers to establish a reference for evaluating automated methods. Their results showed that the best performing features to estimate high and low levels of group cohesion during meetings were: the total pause time between each individual's turns during a meeting segment with audio cues, reaching a 90% classification accuracy, the total visual activity for each person in the meeting, getting to an 83% classification accuracy with visual cues, and the visual activity during periods of overlapped speech with audio-visual cues, hitting an 82% classification accuracy. In order to reach these performances, they used binary classifiers (e.g., SVMs). All the features described in [56] were either based on individuals or at the group level.

Nanninga and colleagues recently extended this work, integrating pairwise and group descriptors related to the alignment of para-linguistic speech behavior [57]. They found that such kind of descriptors outperform traditional turn-taking based descriptors, and they perform better on the estimation of the social dimension than on the task dimension of cohesion. They also showed how combining these 2 types of descriptors guarantees an optimal classification performance. The authors evaluated the performances of 2 supervised classification methods (a Gaussian Mixture Model and a Kernel Density Estimation) fed with nonverbal features (e.g., mimicry and, similarly to [58], turn-taking features). They performed well for classifying the social dimension of cohesion (low or high), for which they achieved a performance of 0.71 Area under the ROC Curve (AUC). Concerning the task dimension of cohesion, they managed to reach a performance of 0.64 AUC. In this study, however, they did not focus on how the task and social dimensions are related to each other over time. As cohesion is an emergent group state, integrating the temporal aspect in the cohesion process could lead to interesting results and discoveries.

Other studies investigated cohesion at a longitudinal level with the use of sociometric badges. It can be anything placed on a person or on its phone, that is able to track the person's movement and activity. The main advantage of such equipment is that it does provide an unobtrusive way of collecting

social and task-relevant interactions. A pioneering study was conducted by Olguin-Olguin and Pentland who built a commonplace wearable technology-based experimental platform for investigating face-to-face interactions of workers for a period of 20 working days [59]. They developed their own sociometric electronic badge to track group members and provide information about their nonverbal behavior and proximity extracted by the frequency of face-to-face interaction together with other sources, such as emails and performance data. Although this study addressed technological challenges on data collection from groups, it however did not directly focus on cohesion and its dimensions. Also, all the features collected through these sociometric badges were only based on individuals and no group level features were analyzed. Zhang *et al.* used the same kind of wearable sensors to study small group collaborations during long duration missions in confined spaces [60]. In order to recognize group members' affect states and group cohesion (i.e., over social and task dimensions), they collected and analyzed data from a group of 6 members involved in a 4-months simulation of a space exploration mission. They defined cohesion detection as a binary classification problem (negative or positive) and they used features in their models both from individual members and group as a whole. Their results show that group task cohesion can be correctly classified with a high performance of over 0.8 AUC. An interesting conclusion from this study is that quantifying behavior patterns including dyadic interactions and face-to-face communications is important in assessing the group process. Results are promising, but they concern a quite specific scenario (i.e., a NASA team that will go on a Mars expedition). Results would certainly be applicable to a military environment but would probably not apply to most of the groups.

Automatically measuring and evaluating cohesion (and emergent states in general) is still at its infancy. Previous studies suffer from a lack of publicly available data specifically designed for cohesion and, at present, all the models built to detect and measure cohesion are trained by using external assessment of cohesion only. Developing models integrating also self-assessment would help to gain insight into this complex emergent state. As shown in [61], indeed, external and self-assessment introduce different biases in the scores used to build labels for the models, respectively. Furthermore, most of the exploitable data only consist of audio and video content. Using technologies such as motion capture systems would also largely benefit the different communities studying cohesion, emergent states and social signal processing by giving more insights and opportunities to successfully model, predict and measure various constructs.

B. DATASETS

Most of the publicly available datasets that involve social interactions among at least 3 persons have been designed either to record social interactions in a specific context such as meetings (see [1] for a review) or in different environments to improve group and crowd recognition algorithms (see [62] for a review). Some of these datasets stand out from the state-of-the-art by introducing newest technologies and ways to record data (e.g., [63]–[65]).

Table 1 shows a selection of relevant datasets of group interactions of at least 3 persons. Moreover, it compares the GAME-ON dataset with respect to the characteristics of such datasets. Some datasets reported in Table 1 captured social interactions with unobtrusive technologies (e.g., video-cameras) in order to analyze natural interactions between participants (see for example [63], [65], [66]) and some used a specific context to elicit specific behaviors with the aim of automatically extracting multimodal signals (e.g., [9], [58], [64], [67]–[70] and see [71] for a review).

The rise of interest in the automatic detection and monitoring of emergent states led researchers to train their algorithms on existing datasets as collecting data in a multimodal fashion is a long and costly process. AMI and VACE datasets were among the first to try to capture groups interaction and many scholars used them in their studies. These datasets, however, did not focus on a specific emergent group state and were, *de facto*, not based on a particular theoretical model. The ELEA dataset [9] addressed emergent leadership in groups by using a well known meeting situation called the Winter Survival Task, a game where 2 participants have to identify objects (out of a predefined list) that would increase their chances of survival in a polar environment. ELEA, however, did not refer to emergent group states but rather focused on the emergence of individual leaders in group interactions. Nevertheless, these datasets include annotations that give the opportunity to use them for diverse studies. SALSA, MatchNMingle, MULTISIMO, AMIGOS and Canal9 also provide a substantial amount of self and external annotations used for identifying participants personality traits, roles, dominance, social cues, F-formations or emotions, easing automated extraction of features related to these measurements (e.g., leadership, agreements, social actions).

The authors of datasets recording free-standing conversational groups (e.g., [65] and [70]) argue that the recording process had none or very small impact on the interactions. These setups, however, are limited in terms of quantity and diversity of the sensors used. Oppositely, datasets using controlled experiments made an effort to record data with new technologies (360°cameras, Kinect, EEG, ECG or GSR) and contain a higher amount of interactions (e.g., [64]).

To the best of our knowledge, there is no existing dataset that explicitly addresses cohesion and controls its dynamics over time. We have the ambition to fill this gap by introducing a new multimodal dataset, GAME-ON, dedicated to the study of cohesion and more specifically to its instrumental dimensions. GAME-ON design is theoretically based on Severt's integrative framework of cohesion. The game context helped to engage participants and elicit natural reactions. Our dataset also provides a significant amount and diversity of data with the use of recent motion capture systems in addition to HD video and audio recordings. It also contains repeated self-annotations per participant about their perception of cohesion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

TABLE 1: A selection of social interactions datasets grouped by scenario. Some of the datasets focus on measuring a construct by using simple settings (i.e., [9], [58], [66], [68], [72]), while the other ones adopted sophisticated technologies (i.e., [63]–[65], [69], [70]). The Table also reports the kind of annotations (self or external), as well as the duration of the recordings available. GAME-ON stands out from the state-of-the-art datasets by providing a significant amount of multimodal data in a game scenario. Other distinctions are that it addresses group cohesion and explicitly controls the underlying interaction dynamics over time. GAME-ON also provides repeated self assessment of cohesion, leadership, emotion and warmth and competence.

						Multimodalities				
Dataset	**Scenario**	**Purpose**	**Group size**	**Duration**	**Annotations** ***Self (⋆), External (*)***	**Vision (HD)**	**Audio**	**Mocap data** ***Inertial***	**Mocap data** ***Optical***	**Other**
AMI (2005) [67]	Meeting	Individual actions, face behaviors, speech	4	167 meetings 100h	Agreements*, disagreements*, dominance*	✔	✔	×	×	×
VACE (2006) [68]	Meeting	Event interpretation, multimodal signal processing	5	N/A	Speaker segmentation*, speech transcription*, F-formations*[2]	✔	✔	×	✔	×
ELEA (2012) [9]	Meeting	Leadership, non verbal behaviors	3-4	40 meetings ~10h	Personality traits⋆, Big Five⋆, perceived leadership⋆⋆, dominance⋆⋆, competence⋆, likeness⋆, ranked dominance⋆⋆	✔	✔	×	×	×
SALSA (2017) [65]	Free Standing Conversational Group	Natural social interactions, F-formations	18	1h	Personality*, position*, head*, body orientation*, F-formation*	✔	✔	×	×	ID/RFID, bluetooth, Accelerometers
MatchNMingle (2018) [70]	Free Standing Conversational Group, speed dates	Automatic analysis of social signals and interactions	2-8	2h	HEXACO⋆, Self Control Scale⋆, Sociosexual Orientation Inventory⋆, social cues*, social actions* F-formations*	✔	✔	×	×	wearable devices recording triaxial acceleration and proximity
MULTISIMO (2018) [69]	Experiment Solving a quiz	Human-human interactions, groups' multimodal behavior	3	23 sessions ~4h	Personality⋆, experience⋆, speaker segmentation*, dominance*, transcripts*, turn-taking*, emotions*	✔	✔	×	×	360°camera, 2 Kinects
AMIGOS (2018) [64]	Experiment Watching videos	Affect, personality, mood	4	~9h	Big-Five⋆, PANAS⋆, valence⋆⋆, arousal⋆⋆, dominance⋆, liking⋆, familiarity⋆, emotions⋆	✔	✔	×	×	EEG, ECG, GSR
Canal9 (2009) [66]	TV Show	Body motions	3+	~43h	Role*, turn-taking*, agreements*, disagreements*, speaker segmentation*	✔	✔	×	×	×
The Idiap Wolf (2010) [58]	Game	Deceptive roles, group interaction	8-12	4 groups ~7h	Speaker segmentation*, roles identifications*	✔	✔	×	×	×
Panoptic (2019) [63]	Game	Capturing social interactions	3-8	65 sequences 5,5h	×	✔	✔	×	×	Massively Multiview System[3]
GAME-ON (2020)	Game	Cohesion, non verbal behaviors	3	17 groups ~11,5h (as a group) ~34.5h (as individual)	Cohesion⋆, leadership⋆, emotional state⋆, warmth and competences⋆	✔	✔	✔	✔	×

over time, giving insights on the dynamics of this emergent group state. We also collected data about participants' emotional states and their perception of leadership and warmth and competence of their group members. As reported in the literature, several emergent states can occur simultaneously and be closely related to each other (e.g., [73]). We are particularly interested in the relationships between cohesion and other emergent states such as leadership.

III. THE GAME-ON DATASET

A. DATA COLLECTION DESIGN

1) The game

Our data collection exploits a game scenario inspired by the rules of Cluedo[1] and is conceived as a simple *escape game*. Cluedo is a board game where 3 to 6 players try to figure out 3 main facts of a murder: the murderer, the location of the murder, and the murder weapon.

An *escape game* is a physical game in which a small team of players is fake locked in a room setup according to a specific theme. The players have to cooperatively discover clues, solve puzzles, and so on to accomplish a specific goal (e.g., escaping, finding an object, or solving a murder) in a limited amount of time. Social games, such as escape games, are a form of socially rich multi-party problem solving where people coordinate and like to spend time together to achieve common goals. They have been considered as a

[1]https://www.hasbro.com

[2]A F-formation is a set of possible configurations in space that people may assume while participating in a social interaction.

[3]http://domedb.perception.cs.cmu.edu/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

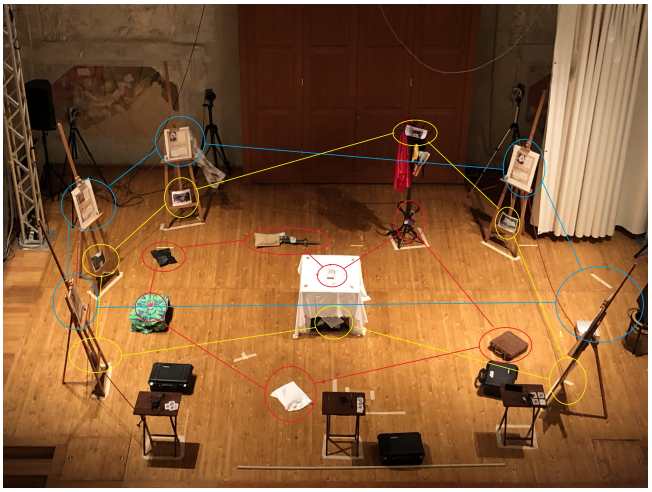


FIGURE 3: The game area and all the material required to solve the murder. Blue circles correspond to the posters of the suspects, yellow circles represent the places where the murder could occur and the potential weapons are circled in red. Near every table at the front of the scene, 3 distinct color marks (blue, green, red) are taped on the floor to indicate participants' personal area.

viable research methodology to address the subtle nuances of human-human communication by several research domains, from psychology [74] and neuroscience [75] to behavioral economics [76] and human computer interaction [77]. There exist, indeed, several datasets in which social games are exploited as an experimental tool for eliciting socio-affective behavior such as laughter [78] and deceptive behavior [58], or for evaluating interaction capture methods [63]. None of them, however, has been designed for studying a specific emergent state, following scholars' models and recommendations.

In the context of the GAME-ON, the game created an engaging experience for the participants and it allowed us to have a fine control on the measurement of the dimensions of cohesion by naturally breaking the whole interaction into distinct tasks. The game scenario was:

During the XIIth century, a brilliant mathematician, student of Fibonacci[4]*, was assassinated and his ghost is trapped into a theatre. Every year the ghost locks people there asking them to help him to discover* ***who*** *killed him,* ***with what*** *weapon and* ***where****.*

The participants had one hour to solve the murder and to escape from the theatre.

The scene (see Figure 3) contains 5 posters of the suspects, with a short description of their personality, 8 potential weapons, with a symbol attached to it and 7 different places where the murder could occur. The game is divided into 5 tasks, either timed or designed to not exceed a specific amount of time (see Table 2 for the detail of the timings). Participants were instructed that they should finish the game as quickly as possible. During each task, they could find different clues, helping them to solve the murder or unlock a new task of the game. Between each task, participants were asked to fill up questionnaires that were conceived as part of the game (e.g., once completed, they received a code for a locker containing the next instructions). Details about the questionnaires are provided in Section III-A4.

To create some competition between the groups and/or among the members of each group, we established a group and an individual leaderboard. This was based on the time participants took to solve the murder and on their performances on the different tasks. Leaderboards are an effective way to motivate participants through competition [79]–[81].

The design of the game has been tested and incrementally adjusted until the beginning of the data collection in order to ensure that the game flow was coherent and that the tasks were understandable by the participants (e.g., we displayed some hints on the wall to make sure that everyone could still progress in the game). The design also largely benefited from knowledge and from discussions with one of the authors, who is an organizational psychologist expert on emergent states.

2) Participants

The data collection took place at Casa Paganini in Genoa, Italy[5]. This is an ancient monumental building having a space, which was formerly used as a theatre. This space is now exploited as a location for experiments on movement analysis in naturalistic settings, and is endowed with a technological infrastructure for motion capture and multimodal recordings. We ran a campaign for recruiting participants through the website of the scientific project funding the data collection[6] and social media[7], mailing lists and the distribution of flyers. The protocol was approved by the Ethics Committee of the Department of Informatics, Bioengineering, Robotics and System Engineering of the University of Genoa, Italy. All subjects gave written informed consent.

In order to take part in the data collection, participants needed to be over 18 (legal age in Italy), to have a good understanding of written and spoken Italian (as all the rules, questionnaires and hints were in Italian) and to participate in a group of 3 friends without any hierarchical status among them. This last point is very important as we are only controlling the functional property of cohesion (see Severt's framework in Section II). Having participants considering themselves as friends allowed us to infer that the affective property of Severt's framework is constant over the time of the data collection. Indeed, we assumed that participants liked each other (interpersonal dimension) and that they were not ashamed to be part of the group (group pride dimension).

[4]Leonardo Fibonacci (c.1170 – c.1240–50) was an Italian mathematician from the Republic of Pisa. He is best known for his discovery of a particular number sequence, which has since become known as the Fibonacci Sequence

[5]http://www.infomus.org/index_eng.php

[6]https://grace.wp.imt.fr/

[7]Instagram and Twitter accounts: @Grace__Project

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

We also observed during the pre-tests, that having participants considering themselves friends, really impacted the spontaneity of the reactions and the dynamics of the group. Also, cohesion can take a long time to emerge in groups of strangers. For instance, previous studies show how cohesion is more volatile during the early phases of team functioning [82] and sustainable task cohesion emerges more quickly than does sustainable social cohesion [83]. A total of 17 groups (i.e., 51 persons) participated in the data collection. Participants' ages ranged from 21y to 33y (M = 25.3y, SD = 3.1y) with 69% identified as female and 31% identified as male. Participant's friendship duration ranged from 1 month to 22 years (M = 3.1y, SD = 2.5y). Concerning the escape game experience of the participants, 64.71% had never participated in an escape game before, 25.49% only tried once and 9.80% participated multiple times. Only 2 participants already went to an escape game together before.

Participants received a small gift having a value inferior to 10 euros as a nominal honorarium for their participation.

3) Procedure

First, we welcomed participants in a room next to the theatre and we asked them to read a description of the data collection, validated by the Ethical Committee. Then, they signed a consent form. Before starting the game, participants filled up a set of questionnaires too in order to assess their level of friendship, their experience in escape games, their perception of the group cohesion, participants' warmth and competence and, finally, their attitude towards group games. The questionnaires were filled up on an Android tablet. We lent one tablet per participant for the time of the game. More details and explanation of the above-mentioned questionnaires are in Section III-A4.

Then, participants entered the theatre. Researchers helped them to wear the motion capture suits and the radio-microphones. Then, a full check of the setup was done in order to make sure that the data was streamed properly.

Participants were allowed to interact freely on stage for few minutes to get acquainted with the sensors. Then, the game started with a pre-recorded audio-video presentation explaining the context and the rules. The presentation was displayed on a wall of the game area. This was done to avoid any bias in providing participants with instructions. Similarly, we used another presentation during the game, automatically displaying additional information, clues or reminders.

The game consisted of 5 tasks and was designed *ad hoc* to control the instrumental functional property of cohesion. Each task was conceived for a specific purpose in order to elicit a controlled variation of the social and task dimension of cohesion, i.e its increase or its decrease, of cohesion. In the following, we refer to those as Increase of Cohesion (I) and Decrease of Cohesion (D). The duration of each task was timed according to its difficulty and the feedback collected during the pre-tests. This is summarized in Table 2.

Figure 4 summarizes the flow of the game. Bubbles indicate the questionnaires administered before, during and after the game. In order to not break the dynamics of the game and to avoid weariness, we integrated the questionnaires in the game logic. In that way, we ensured that all the participants filled up all the questionnaires at the same moment of the game.

TABLE 2: Expected variation of cohesion per task and timed duration of each task. GAME-ON provides increase and decrease measurements of both social and task dimensions of cohesion. DS and IS refer to a decrease and an increase on social dimension, whereas DT and IT refer to a decrease and an increase on task dimension.

Task	Social dimension	Task dimension	Duration (min)
Task 1: Discovery	Decrease (DS)	Decrease (DT)	10
Task 2: Enigmas	DS	Increase (IT)	9
Task 3: The impossible	Increase (IS)	DT	7
Task 4: The weird object	IS	IT	7
Task 5: The presentation	IS	IT	8

Below, we report a detailed description of each of the 5 tasks.

– **Task 1: *Discovery* (DS & DT)**
Participants were asked to find 2 objects, a box and its key, hidden in the game area. The box contains the instructions and materials for the next task. Participants had up to 10 minutes to complete this task. By finding objects, they get bonus points, otherwise they lose points for their personal score on the leaderboard. This task was conceived to encourage participants to discover the game area while being in competition among them to find the objects in order to limit social interactions.

– **Task 2: *Enigmas* (DS & IT)**
17 enigmas were divided into the following different categories: 1) *Matchsticks*: these are rearrangement puzzles in which a number of matchsticks are arranged as squares, rectangles or triangles. The aim is to move one or a limited number of matchsticks to create a new shape. 2) *Logic*: these enigmas describe a specific situation or context and ask the participant to find a logical explanation to it. 3) *Numbers*: these problems require calculations and ask the participant to give a mathematical solution to the problem. 4) *Observation*: these enigmas propose visual scenes with squares or circles and participants need to link different objects together. We intentionally chose enigmas that require different skills to make sure that every participant could contribute. Participants had 4 minutes to split all the enigmas taking into account every participant's skills. This brainstorming was expected to elicit an increase of the task dimension (IT).
Once participants split the enigmas, or if the 4 minutes were over, they had to start working on them in dedicated areas of the stage. They were not allowed to talk, otherwise, they would lose points. We established this rule to limit social interactions. Every time a participant completed an enigma, she had to put it on a box located outside of the game area. This added some stress and we could observe interesting phenomena (e.g., we noticed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

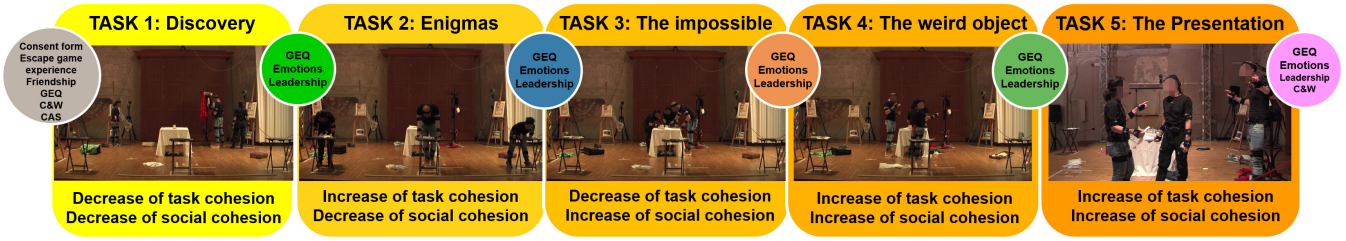


FIGURE 4: Timeline of the flow of the game. The questionnaires are displayed in chronological order before, between and after the tasks (see the bubbles in the figure). The expected variations on cohesion are indicated at the bottom of each image taken from the dataset.

that successful participants were often looked at by the other group members when they moved to the box).
Participants had 5 minutes to solve a maximum of enigmas. At the end of the game, we added or subtracted points to the group regarding the number of correct and wrong answers. All groups received a 4 minutes reward at the end of the last task.

– **Task 3: *The impossible* task (IS & DT)**
This task included 3 different sub-tasks. Participants still needed to collaborate as 2 out of 3 puzzles gave hints about the murderer and the weapon. The theme was Fibonacci. The group received 60 square pieces of paper of different sizes and colors with a number written on the front and a letter written on the back. One person had to reconstruct the Fibonacci sequence, another one had to reconstruct a palindrome spotted on a murderer poster, and the last one had to construct a Fibonacci clock indicating 3:45 pm.
On each weapon, a different Fibonacci clock was printed and participants had to find the clock indicating 3:45 pm to guess the weapon used for the murder.
We made this task impossible to achieve. Each problem required the same pieces of paper. Moreover, it had to be done within 7 minutes, adding some pressure to the participants. As each participant could not complete her part of the puzzle without negatively affect other members of their group, we expected a decrease of the task dimension of cohesion (DT), whereas the social dimension was expected to increase (IS) due to the high number of interactions provoked by a stressing situation.

– **Task 4: *The weird object* task (IS & IT)**
It consisted of guessing what an object was. Participants had to link it to the place of the murder. Then, the group had to write the answer on a paper and put it in a box. If they guessed it right, they earned extra points at the end of the game. This task was timed to 7 minutes.

– **Task 5: *The presentation* (IS & IT)**
The group had 4 minutes to provide a first solution to the murder in an original way (e.g., acting). At the end of the presentation, a red signal was always given by the researcher in charge of the session, indicating that they provided a wrong solution. This was designed to observe the group's reaction after failing. We gave them an extra 4 minutes to present a second solution. At the end of it, a green signal was always given, indicating that they found the solution. This was designed to observe the group's reaction after succeeding.

Task 4 and Task 5 required participants to be creative. We did this choice due to the fact that creativity enhances social interactions, eliciting situations with an increase of cohesion for the social dimension [84]. Also, the fact that the group had to reach a common decision was expected to amplify the task dimension of cohesion. In both Task 4 and Task 5, the 2 studied dimensions were expected to increase. At the end of the data collection session, participants were briefed about the detail, the aim and the context of the study. Moreover, researchers answered all of the participants' questions.
Before leaving the theatre, participants were asked to fill up a last questionnaire to obtain their feedback on the game.

4) Questionnaires

Participants were asked to complete questionnaires at the beginning and at the end of the data collection and after each task to further assess group cohesion as well as individual emotional states and opinions. We chose to adopt repeated measures at regular intervals to reach a good level of granularity and to be able to detect changes in the cohesion process. The questionnaires were presented in the same order after each task, but the order of the items of each questionnaire was randomized in order to keep participants' attention. Figure 4 also shows the order of the questionnaires.
As this data collection involved Italian speakers, we used validated Italian versions of each questionnaire, when they were available. Otherwise, we translated the items without changing the valence nor the grammatical construction of the question, according to the guidelines provided by Carron [29]. It is important to note that all the questionnaires were administrated separately, and we retained the original Likert scale format. In order to assess the consistency of the questionnaires, we ran an Explanatory Factor Analysis

(EFA) at every time they were administered (i.e., before the data collection, after each task, and at the end of the data collection, respectively). Results of each EFA showed that the items were loading into the expected number of factors, indicating the consistency of all the used questionnaires (see Appendix B). Moreover, we calculated Greatest Lower Bounds (GLB) to establish the reliability of the scales. GLB provides a viable option in cases of low number of items and small sample sizes [85]–[91]. All the GLBs were found to be over 0.7 indicating the reliability of all the used questionnaires used (see Appendix B). Here in the following a short description of each questionnaire is provided.

- **Cohesion**: We used the Group Environment Questionnaire (GEQ) [22], [49] to measure group members' self-assessment of cohesion. It consists of a 4-scale 18 items questionnaire, aimed at measuring cohesion in sport groups. Each item can have a score ranging from 1 ("Strongly disagree") to 9 ("Strongly agree") and was administrated before the data collection and after each task. Several studies have shown how the GEQ can be leveraged for addressing group situations in other contexts, for example in work meetings [92], [93] or in exercise classes [51] and even in different cultural contexts [47]. The first time we administrated the GEQ, before Task 1, we decided to discard the 2 following items as we considered that they were not related to the escape game context and hardly adaptable: "I'm not happy with the amount of playing time I get." and "Members of our team do not stick together outside of practice and games.". GEQ scores calculated from this first questionnaire were then used as a baseline for the analyses presented in Section IV.
  Concerning the questionnaires administered between the tasks, we used a shorter version of the GEQ as the answers to some items would not evolve during the time of the data collection. We discarded the 2 following items: "For me, this team is one of the most important social groups to which I belong." and "Some of my best friends are on this team.".
  We also slightly adapted the items without changing the valence nor the grammatical construct of the question. For example, "Our team members have conflicting aspirations for the team's performance." became "Our team members had conflicting aspirations for finding the key." after the *Discovery* task.
  We also decided to replace 2 items by ones from Michalisin *et al.*'s [93] study as we believe that they are close enough to the originals and more suited to our context. In that way, "I enjoy other parties rather than team parties." became "I wish I was on a different team." and "I do not like the style of play on this team." was replaced by "Our team does not work well together.".
  Our version of the GEQ used between the tasks contains 14 items: 8 related to the task dimension, and 6 to the social dimension (see Appendix A).

- **Warmth and competence (W&C)** [94]: This questionnaire is a set of 8 items to measure warmth and competence at inter-group, interpersonal and individual levels, answered on a 9-points Likert scale from 1 ("I completely disagree") to 9 ("I completely agree"). We used a round-robin rating, meaning that each participant had to rate all the other participants and themselves.
  Half of the items are related to the warmth dimension whilst the other half focus on the competence dimension. The warmth dimension captures traits that are related to perceived intent, including friendliness, helpfulness, sincerity, trustworthiness and morality whereas the competence dimension reflects traits that are related to perceived ability, including intelligence, skill, creativity and efficacy [95]. Participants were asked to fill up this questionnaire before and at the end of the data collection.

- **Competitivity**: The Italian version of the Competitivity Attitude Scale (CAS) questionnaire was used [96]. It consists of 10 items on participants attitude toward competition. This is a self-assessment questionnaire on a 5-point Likert scale from 1 ("Never true for me") to 5 ("Always true for me").
  This questionnaire was administered just before the *Discovery* task with a twofold aim: to foster participants competitiveness by having them reason about it, and to gain further information in participants attitude towards group games.

- **Emotions**: In order to get some insights on participants' emotions at each task, we asked them to answer to the question: "How do you feel?" by picking among 6 different emotional labels. Moreover, participants could select the "other" option and provide their own emotional label. The labels were selected by relying on the Emotion Theory by Roseman [97]. According to this theory, emotions depend on the subjective perception of the ongoing situation (i.e., one's own appraisal), in terms of causal attribution (the situation was caused by someone else, by the self or was due to external circumstances) or in terms of being consistent or not with one's goals and motivations.
  Each emotion can be identified by a specific combination of causal attribution and goal consistency (i.e., its appraisal configuration) [98]. For instance, a player winning a game may feel *pride* as a consequence of perceiving herself as responsible for the victory (causal attribution) and because winning satisfies her goal of being a good player (consistency with personal goals and motivations). According to their appraisal configuration, emotions can be categorized as positive or negative [99]. Following this, we selected 6 emotions that, given their specific appraisal configuration, might be elicited by the game. We selected 3 positive and 3

negative emotions: 2 of them resulting from an "other-caused" causal attribution (admiration and anger), 2 from a self-caused causal attribution (pride and shame) and 2 from a circumstances-caused causal attribution (happiness and frustration). We selected these specific emotions as they were the most relevant given the context of the game.

– **Leadership**: We used a set of 5 items on a 6-point Likert ranging from 1 ("Completely disagree") to 6 ("Completely agree"), following Gerpott *et al.*'s study recommendations [100] based on previous work [101], [102]. For the same reasons as in the W&C questionnaire, we decided to use a round-robin rating.

– **Motivation**: We used the Intrinsic Motivation Inventory (IMI) questionnaire [103]. This was initially conceived as a multidimensional measurement device intended to assess participants' subjective experience related to a target activity in laboratory experiments. It is on a 7-point Likert scale from 1 ("Completely disagree") to 7 ("Completely agree"). We decided to leverage this tool at the end of the data collection session as a guide for our debriefing phase. Having participants' opinion about the game and their enjoyment would be useful for further studies. With this in mind, we selected the Interest/Enjoyment and Perceived Competence subscales from the IMI.

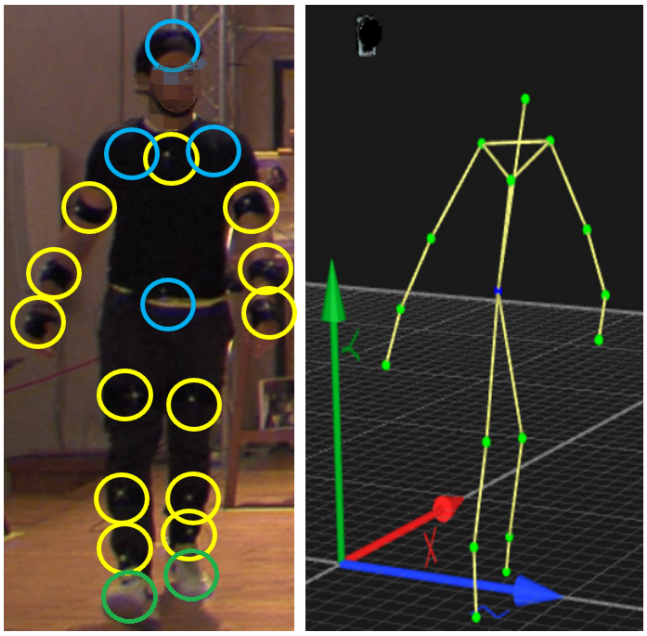


FIGURE 5: Position of the 17 IMU Shadow sensors and 17 Qualisys reflective markers (yellow and blue circles) on a participant and its associated reconstructed skeleton. Sensors circled in blue are positioned at the back of the participant (the 2 shoulders, the head and the hip). Sensors in yellow are at the front of the participant. Green circles correspond to the 2 Shadow sensors placed in the participant's shoes.

B. TECHNICAL SETUP

1) Equipment

To collect our dataset we built a rich setup that allowed us to manage data from different sources. Synchronization of the data was handled via hardware and software as explained in Section III-B2. We captured the behaviors of 3 persons interacting simultaneously. For this purpose, we adopted a hybrid motion capture approach combining together 3 Shadow inertial motion capture suites[8] with a Qualysis optical motion capture system[9]. This choice was made to take advantage of the strengths that each technology offers, correct the drifts that may occur in long recording sessions and have reliable measures. Shadow's suite is a wireless wearable system composed by 17 IMU sensors (3-axis accelerometers, gyroscopes, and magnetometers), placed on the body at some precise reference points (see Figure 5) plus 2 additional sensors, placed in the participants' shoes. In our setup, data was captured at 100Hz.

Qualysis configuration included 16 infra-red cameras optimally placed to cover the whole game area. Data was also captured at 100Hz. In order to have a perfect coupling between the 2 systems, 17 infra-red reflective Qualysis markers were attached on the Shadow's IMUs with Velcro straps. Additionally, audio and video were recorded. We used 3 wireless headsets microphones (AKG wireless set 800MHz with C555L headsets, Mono, 48kHz, 16 bits per sample), and 2 static professional JVC video-cameras (1280×736, 50fps) frontally (at about 9m from the center of the scene) and laterally (at about 4.5m from the center of the scene) placed with respect to the game area. Moreover, 2 additional Panasonic handy cameras (1920×1080, 50fps) completed the setup. These last 2 video-cameras were used as back-up cameras and were not synchronized.

For data acquisition and synchronization, we used 4 desktop PCs (I7 Intel processor, 8 GB DDR3 RAM, Windows 10x64), 1 devoted to audio capturing, 1 devoted to video capturing, 1 for the Qualysis system, and 1 for the Shadow system.

2) Software platform

Data recordings were handled by using EyesWeb[10], a software platform to support real-time capturing and processing of multimodal data streams. EyesWeb handles data synchronization by time-stamping each received frame or sample. Time-stamping is based on SMPTE time codes[11], with the additional possibility to use sub-sample accuracy. When the hardware supports it, the SMPTE signal is used as a reference clock. For example, the Qualisys system can receive an

[8] https://www.motionshadow.com/

[9] https://www.qualisys.com/

[10] http://www.infomus.org/eyesweb_eng.php

[11] See standard ST 12-1:2014, which is available at the SMPTE website: https://www.smpte.org/standards/document-index/ST

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

SMPTE signal as input, and lock to it. This mechanism is also used by the JVC video cameras. In such cases, the received samples are automatically timestamped by the capture device. Other devices are synchronized by EyesWeb, which timestamps each sample when it is received by the host computer. By means of these timestamps, EyesWeb can accurately play the data back with the same timings as they were captured. That is, this process preserves each raw signal native frame rate, when performing multimodal analysis.

In the case of the GAME-ON dataset, the frontal JVC camera was generating the SMPTE time codes, which were received by the lateral JVC camera, by the audio card of the PC for audio recordings, by the Qualysis system, and by the PC running the Shadow recorder. Thus, audio, video, and Qualysis recorders were all locked to the same SMPTE signal. The Shadow system generates its own timestamps. Shadow data, including the timestamp, were received by an *ad-hoc* C# console application connected to both the Shadow system and to EyesWeb. Shadow data was thus received by EyesWeb, and the correspondence between the SMPTE time code and the Shadow timestamp, for each Shadow sample, was recorded in a separate file, letting us manage synchronization between Shadow data and other data.

3) Data inspection

Post-processing included several steps. As data was recorded separately for each task, the first step was to trim the data to only keep the interesting content, discarding the moments where participants were filling questionnaires or were waiting for the others to start a new task.

We used ffmpeg[12] to trim our audio and video files and discarded the data that was not tasks related.

Then, the second step consisted of determining what data got lost for each sensor. Among all the groups (representing 11h36m16s of data) we had to discard 2 groups (1h16m48s), representing 11.03% of the data, due to connectivity problems between the C# application and the Shadow system, causing deep gaps in the data.

We used Qualisys technology to correct the small drift that Shadow, similarly to other inertial motion capture systems, may introduce, thus having more accurate coordinate values for each of the 17 points. We only needed to label 1 point (i.e. hip or head) with the Qualisys Track Manager (QTM) software to get the drift-corrected translation values for all the other points. We used the hip marker except for the frames where it was not visible. Concerning the video, we managed to save 100% of the files, whilst we lost 3.49% of the audio data, representing 24m16s of content. Missing audio is however available on the back-up cameras.

4) Data visualization

We developed an EyesWeb application to visually check that the motion capture data concerning the 17 points representing joints in the participants' skeletons was coherent. As the data was recorded and stored in a specific architecture and format, this application automatically selects and plays the audio, the video and the motion capture data files belonging to the same recording session in a synchronized way. Here below the organization of the recorded files:

```
Date of the session (e.g., 2019-10-28)
├─ audio
│  └─ Audio files (.aif)
├─ qtm
│  └─ Qualysis' Qtm files (.qtm)
├─ shadow
│  ├─ Shadow's CSV files
│  └─ Shadow's text files (timestamps)
└─ video
   ├─ Video files (.avi)
   └─ Video's text files (timestamps)
```

We recorded 1 audio file per participant and per task for a total of 15 audio files per group. We recorded 1 QTM file per task for a total of 5 QTM files per group. Concerning the Shadow data, we stored all the data in 1 CSV containing all the sensors values per participant per task and 1 text document per CSV file, storing the shared timestamps for a total of 30 files. We saved the frontal and lateral video recordings for each task, but also 1 text file per recording storing the shared timestamps, for a total of 20 files.

Figure 6 shows the EyesWeb application that we developed to visualize the data.

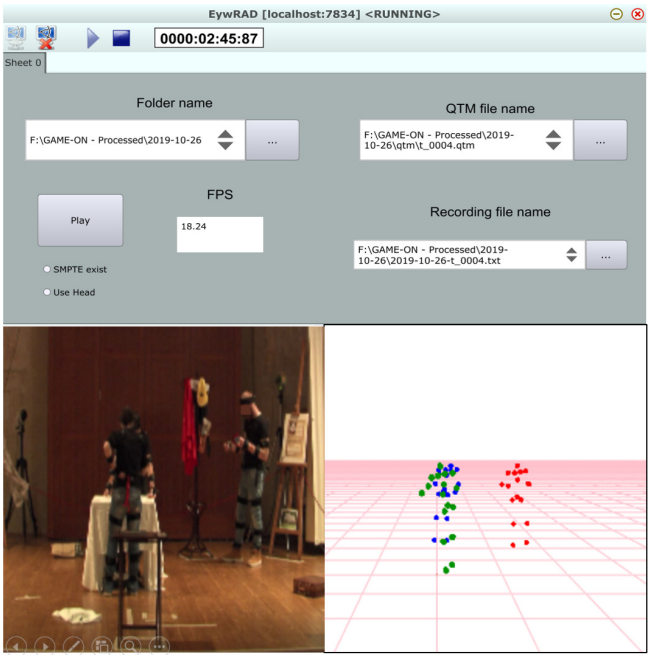


FIGURE 6: The EyesWeb application for visualizing synchronized data streams.

IV. DATA ANALYSIS

In this Section, we present a first analysis of the data gathered through the questionnaires. We used an alpha level of 0.05 for

[12]https://www.ffmpeg.org/

all statistical tests.

A. GEQ ANALYSIS

The following analysis aimed at understanding and evaluating the dynamics of cohesion over time, regarding its social and task dimensions. In Task 1, we looked at the variations of cohesion (i.e. increase or decrease) with respect to the baseline obtained from the first administration of the GEQ questionnaire before starting the data collection. In each of the other Tasks, we looked at the variations with respect to the previous one. Moreover, we also looked at the variation of cohesion measured at the beginning of the data collection and after each task.

In order to analyze such variations, we computed 2 self-assessment scores of cohesion from the GEQ questionnaire, for every participant and for each task. We named these scores as GEQ-Social and GEQ-Task, respectively. The former relates to the social dimension and it results from the sum of the items 1 to 6 reported in Appendix A. As there are 6 items, the minimum score possible was 6 and the maximum score possible was 54. The latter one corresponds to the task dimension and it results from the sum of the items 7 to 14 reported in Appendix A. As there are 8 items, the minimum score possible was 8 and the maximum score possible was 72. Figure 7 shows the boxplots of the GEQ-Social and GEQ-Task scores, respectively. In order to test the normality of the data, we used a Shapiro-Wilk test. In both cases, the test showed a significant departure from normality for both the social dimension ($W = 0.93$, $p < .001$) and the task dimension ($W = 0.96$, $p < .001$).

1) The social dimension

A non-parametric Friedman test of differences among repeated measures showed a significant difference between the GEQ-Social scores across tasks ($X^2(5) = 31.40$, $p < .001$). Post-hoc Conover's tests with a Bonferroni-adjusted alpha level confirmed that we managed to control the social dimension of cohesion accordingly to the sequence in Figure 4. In Task 1 and Task 2, we expected to break the social cohesion of the group, developed prior the data collection as participants were friends (from IS to DS). Then, we wanted to observe an increase of social cohesion in Task 3, Task 4 and Task 5 (from DS to IS).

Post-hoc tests showed a significant difference between the Baseline and all the tasks ($p < .001$)[13], proving that the game had an impact on the social dimension of cohesion. Moreover, as expected, we observed a significant decrease of social cohesion between Task 1 and Task 2 ($p < .001$, $Mdn = 45$ for Task 1, and $Mdn = 42$ for Task 2). We also observed a significant increase of social cohesion between Task 2 and Task 3 ($p < .001$, $Mdn = 42$ for Task 2 and $Mdn = 43$ for Task 3), and between Task 3 and Task 4 ($p < .001$, $Mdn = 43$ for Task 3, and $Mdn = 44$ for Task 4), indicating that the expected behavior was indeed obtained. Post-hoc tests also showed significant differences between Task 4 and Task 5 ($p = .015$). Again, the medians increased ($Mdn = 44$ for Task 4 to $Mdn = 45$ for Task 5), indicating that this last task can also be considered as IS.

[13] All the p-values presented are already Bonferroni-adjusted.

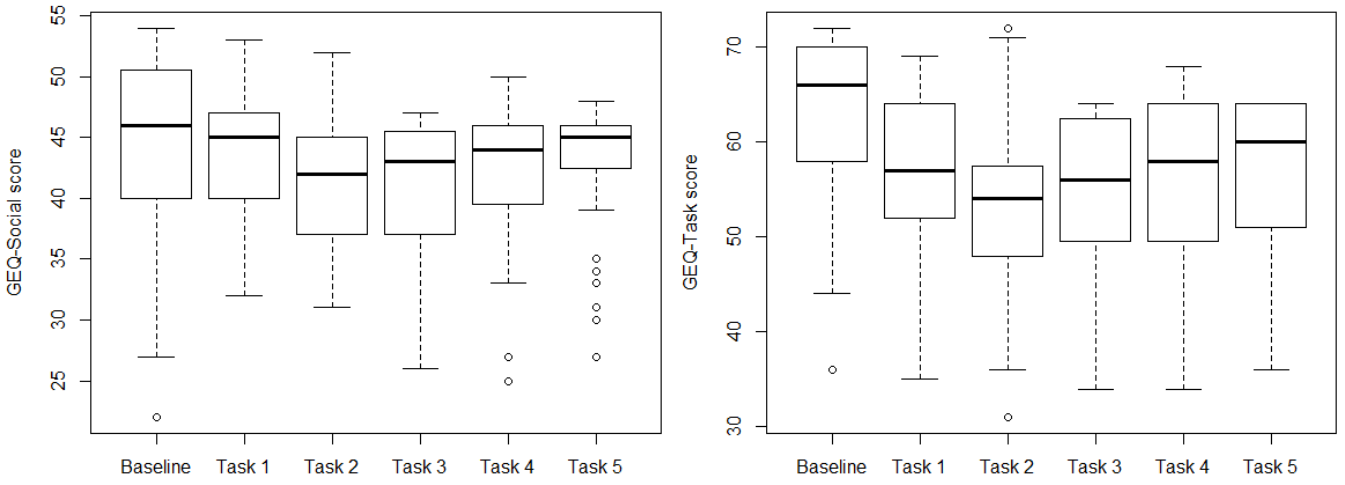
2) The task dimension

A non-parametric Friedman test of differences among repeated measures showed a significant difference between the GEQ-Task scores across tasks ($X^2(5) = 43.86$, $p < .001$). Post-hoc Conover's tests with a Bonferroni-adjusted alpha level showed, however, a different situation with respect to the one presented in Figure 4. We first expected task cohesion to decrease from Baseline to Task 1 (from IT to DT) and then, to observe an increase in Task 2, followed by another decrease in Task 3. Finally, we expected task cohesion to increase in Task 4 and Task 5.

Similarly to the results obtained for the social dimension of cohesion, post-hoc tests showed a significant difference between the Baseline and all the tasks ($p < .001$)[13], proving that the game had an impact on the task dimension. There also was a significant difference between Task 1 and Task 2 ($p < .001$), but medians decreased instead of increased as we expected ($Mdn = 57$ for Task 1, $Mdn = 54$ for Task 2).

Several explanations account for this result. A visual inspection of the video data showed that the participants did not fully understand the aim of Task 2. We noticed that the researcher in charge of the session had to remind the instructions more than once during the other tasks as participants were not following or understanding the guidance. Also, Task 2 was designed to allow time to participants (4 minutes) to organize the distribution of the enigmas among them. This was expected to result in an increase of task cohesion, but most of the groups rushed to the next phase of the task and randomly assigned enigmas. As participants were not allowed to interact during the second part of the task (5 minutes), it is very likely that their answers about the task dimension were biased by the decrease of social cohesion. Also, whereas we were aware that eliciting and measuring multiple changes of one single dimension over a very short period of time (i.e., the Task 1 – Task 2 – Task 3 sequence) was complicated, this indeed revealed more complicated than expected. In brief, we could only observe a significant decrease of task cohesion between Task 1 and Task 2 ($p < .001$, $Mdn = 57$ for Task 1, $Mdn = 54$ for Task 2) and between Task 1 and Task 3 ($p = .009$, $Mdn = 57$ for Task 1 and $Mdn = 56$ for Task 3), and a significant increase of task cohesion between Task 4 and Task 5 ($p = .028$, $Mdn = 58$ for Task 4 and $Mdn = 60$ for Task 5). Indeed, according to Conover's post-hoc results, there also was a marginally significant difference in task cohesion between Task 2 and Task 4 ($p = .063$), and a significant difference between Task 2 and Task 5 ($p < .001$). GEQ-Task scores in Task 3 and Task 5 were significantly different ($p < .001$) too. In summary, we can consider that GEQ-Task scores for Task 1 and Task 3 reflect a downward variation of task cohesion as the medians significantly decreased. Conversely, an upward

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access



(a) Boxplot of the GEQ-Social scores

(b) Boxplot of the GEQ-Task scores

FIGURE 7: Boxplots of the GEQ-Social and GEQ-Task scores per task. Medians of GEQ scores are represented by the bold black lines. White dots represent outliers. Figures 7a and 7b show that the medians first decreased from the Baseline (GEQ administered before the game) to Task 2 and then increased until Task 5, nearly going back to their original values. GEQ-Social and GEQ-Task could range from 6 to 54 and from 8 to 72, respectively. This is due to a higher number of task-related items.

variation is observed between Task 3, Task 4, and Task 5, so that we can conclude that task cohesion increased in Task 4 and Task 5.

In conclusion, despite that Task 2 was probably miss-evaluated, we still managed to control the direction of variation of the task dimension of cohesion over time.

B. LEADERSHIP ANALYSIS

A leadership score per participant per task was computed by summing up all the items scores of the leadership questionnaire. The participant with the highest score for a task was considered as the leader for this specific task. If 2 or more participants had the same highest score, we considered that no leader emerged for the task. If the same participant was a leader for at least 4 tasks over 5, we considered that this participant was clearly identified as the leader for the game.

Table 3 presents the percentage of the number of times a leader was identified per task. We remarked that the more participants interacted with each other, the more a leader was identified. We can explain this by the fact that most of the participants (96%) never played an escape game together before and leadership, for these specific tasks, emerged with time. Surprisingly, we could not identify any game leader. Except for 2 groups where only 1 leader was identified for only 1 task, we could note that, systematically, 2 participants over 3 were identified as leader during the game. This means that 1 member of the group was clearly identified as not taking any leadership for the game. This is a result concerning group roles that, in our opinion, is worth of further analysis.

TABLE 3: Percentage of the number of times a leader was identified per task.

	Task 1	Task 2	Task 3	Task 4	Task 5
Number of times a leader was identified (%)	70.59	70.59	76.47	82.35	94.12

Our leadership score, however, mixes both self and external assessments. According to Vinciarelli, each type of assessment has biases [61]. Further research will be carried out to investigate the impact of these biases on leadership analysis.

C. EMOTIONS AND MOTIVATION ANALYSIS

After each task, we asked participants to pick (and/or provide) the emotions that best described their feelings. Figure 8 shows the results, task by task. We observed that in the tasks eliciting an increase of cohesion in both dimensions (i.e., Tasks 4 and Task 5), happiness was the most dominant feeling, corresponding to 34.29% and 54.29% of the answers, respectively. In Task 1, the feeling of happiness was probably influenced by participants' excitation at the start of the game. We observed, however, 3 other emotions related to the discovery of the box or the key: *Proud*, *Frustrated* and *Admiration*. A participant was more likely to feel proud or frustrated depending on whether she found an object or not. Arguably, as participants were friends, one would more easily feel admiration toward one's group members.

In Task 2 and Task 3, participants felt frustrated (36.07% and 41.27% respectively). These 2 tasks were intentionally made difficult (or impossible) to complete. In Task 2, however, we observed a higher diversity in the answers. This is probably related to participants' appreciation of the quality of their own performance. We also noticed that happiness was either the first or the second most dominant emotion at every task of the game. This is in line with the results from the IMI questionnaire administered at the end of the game. We summed all the items scores and compared them in order to assess participants' level of enjoyment. A high score indicates a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

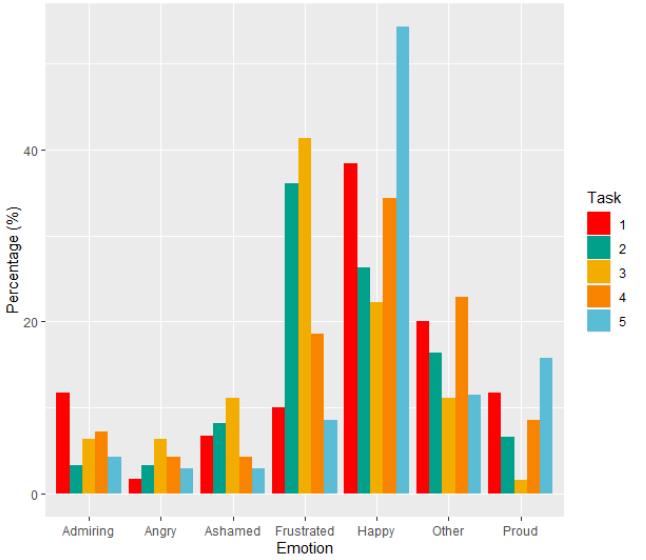


FIGURE 8: Percentages of the 6 emotional labels per task. The 2 most dominant emotional labels chose were "Happy" and "Frustrated". The "Other" category includes 19 different emotional labels provided by the participants.

high level of enjoyment from a participant. The minimum possible score was 14 and the maximum possible score was 98. Results vary from 25 to 77 (M = 58, SD = 10.22, Mdn = 61). Based on the scores' distribution, we assumed that a participant particularly enjoyed the game if her IMI score was strictly above the median, particularly did not enjoy the game if her IMI score was strictly below a threshold value, or felt neutral if her IMI score was comprised between the threshold and the median. Here, the threshold value was fixed to 52 by subtracting the standard deviation from the median, indicating a particularly negative experience given the range of the IMI scores. We reckon that 46.81% enjoyed the game, 25.53% did not, and 27.66% felt neutral about it.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced GAME-ON, a new multimodal dataset dedicated to the study of cohesion, and more specifically to its instrumental dimensions (social and task) at a horizontal level. This dataset presents the following advantages with respect to the other datasets available in the literature. First, GAME-ON's design was conceived to study cohesion and its variations over time, and it was based on a theoretical model of cohesion (i.e., Severt's conceptual framework of group cohesion [23]), unlike any other available datasets. Then, it includes multimodal data from several synchronized sensing systems and questionnaires responses (self-assessment) on cohesion and other related constructs (emotions, leadership, warmth and competence) from 51 participants, which is uncommon in Social Signal Processing studies. This will allow researchers to enrich the analysis of cohesion by probing its relations with other constructs. All the consolidated methods from computer vision, movement analysis, and speech processing can be applied on the dataset to extract features characterizing individuals as well as groups. Moreover, GAME-ON can also be used as a testbed for developing new algorithms for automated behavioral analysis.

GAME-ON, however, has some limitations. It only explores 2 facets of cohesion (i.e., social and task) over the 4 presented in Severt's framework [23]. In addition, the relatively short duration of each data collection session (i.e., 1 hour) is likely to have constrained the range of variation of cohesion we could observe. Moreover, despite the variety of data available in GAME-ON, it does not include physiological data that could be useful to enrich the understanding of cohesion and analytical methods.

As the statistical results show, except for task cohesion between Task 1 and Task 2 for which we observed a variation of cohesion in the opposite direction than expected, we successfully manipulated all the relative variations of cohesion (between tasks) over time.

Further analysis will concern the link between cohesion and participants' emotional states, as well as the link between cohesion and participants' perception of leadership and warmth and competence, respectively. Moreover, we will run an external annotation campaign on cohesion perception that will allow researchers to study differences between self-reported and observed cohesion.

GAME-ON will be publicly available for research purposes. We are confident that it will be a great asset for researchers studying cohesion and other emergent states in dynamic group interactions.

.

APPENDIX A THE GEQ QUESTIONNAIRE

Items related to the social dimension of cohesion

1) I did not enjoy socially interacting with the team.
2) I do not want to continue playing with this team.
3) I would rather solve the enigmas on my own than together.
4) We did not have fun during the task.
5) I would like to spend more moments like the previous one with this team.
6) I wish I was on a different team.

Items related to the task dimension of cohesion

7) I was unhappy with my team's level of desire to win.
8) This team did not give me enough opportunities to use my abilities when we shared the enigmas.
9) Our team was united in trying to solve as many enigmas as possible.
10) We all took responsibility for any loss or poor performance.
11) Our team members had conflicting aspirations for solving the enigmas.
12) If members of our group had problems while trying to resolve a problem, everyone wanted to help them.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3005719, IEEE Access

13) Our team members did not communicate freely about each members' responsibilities during our task.
14) Our team did not work well together.

APPENDIX B PSYCHOMETRIC PROPERTIES

We ran an Exploratory Factor Analysis (EFA) with oblique rotation (promax) to assess the consistency of the questionnaires. For the GEQ and the W&C scales, EFA was performed for both dimensions (i.e., social/task and warmth/competence, respectively), each time the questionnaire was administered (i.e., before the data collection, after each task, and at the end of the data collection, respectively). First, the Kaiser criterion was applied [104]; therefore all factors holding eigenvalues greater than 1 were retained. Then, we performed a Scree test to determine the number of factors to adopt. Results are explained for each scale below.

A. CONSISTENCY RESULTS (EFA)

EFA results suggested a one factor solution for each dimension measured by the GEQ (i.e., social and task) and the W&C scale (i.e., warmth and competence), thus supporting the idea of all the items related to a specific dimension loading into the same factor. This was true at each time we administered the questionnaires, indicating their consistency. Regarding the Leadership questionnaire, EFA results suggested a multiple factor solution. We observed that the items were loading into multiple factors (i.e., 2 factors for Task 2 and Task 4 or 3 factors for Tasks 1, Task 3 and Task 5). Our results can be explained by the fact that each task elicited and required different group dynamics and different aspects of leadership. This is in line with the functional leadership theory [105], according to which team leaders should adapt their behavior depending on the team needs during a specific situation. Hence, we opted for a more parsimonious solution relating all the different functions to one overall leadership factor.

Finally, regarding CAS and IMI scales, even if we did not modify the original questionnaires, we decided to verify their psychometric properties. EFA suggested a 2 factors solution which is in line with previous work on the CAS study [96] and coherent regarding the IMI scale as we only selected 2 subscales from the original questionnaire [103].

Table 4 reports all the percentages of variance explained by the retained factor(s).

TABLE 4: Percentage of variance explained by the retained factor(s)

		Percentage of variance explained by the retained factor(s)					
		Baseline	Task 1	Task 2	Task 3	Task 4	Task 5
GEQ	Social	44	31	38	47	56	61
	Task	44	37	24	39	41	43
W&C	Warmth	72	-				80
	Competence	77	-				78
Leadership		-	75	60	75	64	77
CAS		59	-				
IMI		-					59

B. RELIABILITY RESULTS (GLBS)

We calculated Greatest Lower Bounds (GLB) to establish the reliability of our questionnaires. GLBs are reported in Table 5. All of the values are over 0.7, indicating the reliability of our questionnaires.

TABLE 5: GLBs obtained for each questionnaire

		GLB values for each questionnaire					
		Baseline	Task 1	Task 2	Task 3	Task 4	Task 5
GEQ	Social	.882	.734	.713	.750	.830	.872
	Task	.920	.856	.723	.792	.702	.818
W&C	Warmth	.988	-				.996
	Competence	.996	-				.995
Leadership		-	.989	.954	.931	.990	.992
CAS		.882	-				
IMI		-					.909

ACKNOWLEDGMENT

First, we would like to thank Casa Paganini who hosted the data collection for the warm welcome and help, Valerio Bioglio for the ghost voices used to give instructions, Erik Bakke for his valuable technical support and Beatrice Biancardi for her feedback on the questionnaire analysis. Special thanks to Simone Ghisio and Roberto Salgoleo for the technical support and the fruitful discussions on motion capture technologies. We also thank all the volunteers who contributed their free time to take part in these recordings.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," Image and Vision Computing, vol. 27, no. 12, pp. 1743–1759, 2009.
[2] M. Marks, J. Mathieu, and S. Zaccaro, "A temporally based framework and taxonomy of team processes," The Academy of Management Review, vol. 26, no. 3, pp. 356–376, 2001.
[3] S. W. J. Kozlowski, "Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations," Organizational Psychology Review, vol. 5, no. 4, pp. 270–299, 2015.
[4] R. Reiter-Palmon, T. Sinha, J. Gevers, J.-M. Odobez, and G. Volpe, "Theories and models of teams and groups," Small Group Research, vol. 48, no. 5, pp. 544–567, 2017.
[5] S. W. J. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," Psychological Science in the Public Interest, vol. 7, no. 3, pp. 77–124, 2006.
[6] J. Mathieu, M. Travis Maynard, T. Rapp, and L. Gilson, "Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future," Journal of Management, vol. 34, no. 3, pp. 410–476, 2008.
[7] R. Rico, M. Sánchez-Manzanares, F. Gil, and C. Gibson, "Team implicit coordination processes: A team knowledge-based approach," The Academy of Management Review, vol. 33, no. 1, pp. 163–184, 2008.
[8] N. Lehmann-Willenbrock, H. Hung, and J. Keyton, "New frontiers in analyzing dynamic group interactions: Bridging social and computer science," Small Group Research, vol. 48, no. 5, pp. 519—-531, 2017.
[9] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 816–832, 2012.
[10] S. W. J. Kozlowski and G. T. Chao, "The dynamics of emergence: Cognition and cohesion in work teams," Managerial and Decision Economics, vol. 33, no. 5–6, pp. 335–354, 2012.
[11] L. Rosh, L. R. Offermann, and R. Van Diest, "Too close for comfort? Distinguishing between team intimacy and team cohesion," Human Resource Management Review, vol. 22, no. 2, pp. 116–127, 2012.
[12] J. C. Magee and L. Z. Tiedens, "Emotional ties that bind: The roles of valence and consistency of group emotion in inferences of cohesiveness and common fate," Personality and Social Psychology Bulletin, vol. 32, no. 12, pp. 1703–1715, 2006.

[13] D. Levine, D. Buchsbaum, K. Hirsh-Pasek, and R. M. Golinkoff, "Finding events in a continuous world: A developmental account," Developmental psychobiology, vol. 61, no. 3, pp. 376–389, 2018.
[14] E. Salas, R. Grossman, A. M. Hughes, and C. W. Coultas, "Measuring team cohesion: Observations from the science," Human Factors, vol. 57, no. 3, pp. 365–374, 2015.
[15] C. W. Coultas, T. Driskell, C. Shawn Burke, and E. Salas, "A conceptual review of emergent state measurement," Small Group Research, vol. 45, no. 6, pp. 671–703, 2014.
[16] K. Lewin, "Field theory and experiment in social psychology: Concepts and methods," American Journal of Sociology, vol. 44, no. 6, pp. 868–896, 1939.
[17] K. Lewin, "Behavior and development as a function of the total situation," in Manual of child psychology, pp. 791–844, John Wiley & Sons Inc, 1946.
[18] L. Festinger, S. Schachter, and K. W. Back, Social pressures in informal groups; a study of human factors in housing. Harper, 1950.
[19] A. J. Lott and B. E. Lott, "Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables," Psychological Bulletin, vol. 64, no. 4, pp. 259–309, 1965.
[20] K. W. Back, "Influence through social communication," The Journal of Abnormal and Social Psychology, vol. 46, no. 1, pp. 9–23, 1951.
[21] A. Van Bergen and J. Koekebakker, ""group cohesiveness" in laboratory experiments," Acta Psychologica, vol. 16, pp. 81–98, 1959.
[22] A. V. Carron, W. N. Widmeyer, and L. R. Brawley, "The development of an instrument to assess cohesion in sport teams: The group environment questionnaire," Journal of Sport Psychology, vol. 7, no. 3, pp. 244–266, 1985.
[23] J. Severt and A. Estrada, "On the function and structure of group cohesion," in Team Cohesion: Advances in Psychological Theory, Methods and Practice, vol. 17, pp. 3–24, Emerald Group Publishing Limited, 2015.
[24] K. Dion, "Group cohesion: From field of forces to multidimensional construct," Group Dynamics: Theory, Research, and Practice, vol. 4, no. 1, pp. 7–26, 2000.
[25] D. J. Beal, R. R. Cohen, M. J. Burke, and C. L. McLendon, "Cohesion and performance in groups: A meta-analytic clarification of construct relations," Journal of Applied Psychology, vol. 88, no. 6, pp. 989–1004, 2003.
[26] K. A. Bollen and R. H. Hoyle, "Perceived cohesion: A conceptual and empirical examination," Social Forces, vol. 69, no. 2, pp. 479–504, 1990.
[27] J. Griffith, "Measurement of group cohesion in u.s. army units," Basic and Applied Social Psychology, vol. 9, no. 2, pp. 149–171, 1988.
[28] D. Katz, "The functional approach to the study of attitudes," Public Opinion Quarterly, vol. 24, no. 2, pp. 163–204, 1960.
[29] A. V. Carron and L. Brawley, "Cohesion: Conceptual and measurement issues," Small Group Research, vol. 31, pp. 89–106, 2000.
[30] B. Acton, M. Braun, and R. Foti, "Built for unity: assessing the impact of team composition on team cohesion trajectories," Journal of Business and Psychology, pp. 1–16, 2019.
[31] M. T. Braun, S. W. J. Kozlowski, T. A. R. Brown, and R. P. DeShon, "Exploring the dynamic team cohesion–performance and coordination–performance relationships of newly formed teams," Small Group Research, 2020.
[32] F. M. Leo, I. González-Ponce, T. García-Calvo, D. Sánchez-Oliva, and E. Filho, "The relationship among cohesion, transactive memory systems, and collective efficacy in professional soccer teams: A multilevel structural equation analysis.," Group Dynamics: Theory, Research, and Practice, vol. 23, no. 1, pp. 44–56, 2019.
[33] N. C. Magpili and P. Pazos, "Self-managing team performance: A systematic review of multilevel input factors," Small Group Research, vol. 49, no. 1, pp. 3–33, 2018.
[34] R. Martens, D. M. Landers, and J. W. Loy, "Sport cohesiveness questionnaire," Unpublished manuscript, University of Illinois, Champaign, IL, 1972.
[35] J. R. Ball and A. V. Carron, "The influence of team cohesion and participation motivation upon performance success in intercollegiate ice hockey," Canadian Journal of Applied Sport Sciences, vol. 1, no. 4, pp. 271–275, 1976.
[36] D. Yukelson, R. Weinberg, and A. Jackson, "A multidimensional group cohesion instrument for intercollegiate basketball teams," Journal of Sport and Exercise Psychology, vol. 6, no. 1, pp. 103–117, 1984.
[37] A. V. Carron, L. R. Brawley, and W. N. Widmeyer, "The measurement of cohesiveness in sport groups," Advances in sport and exercise psychology measurement, vol. 23, no. 7, pp. 213–226, 1998.
[38] L. R. Brawley, A. V. Carron, and W. N. Widmeyer, "Assessing the cohesion of teams: Validity of the group environment questionnaire," Journal of Sport and Exercise Psychology, vol. 9, no. 3, pp. 275–294, 1987.
[39] A. V. Carron, "Cohesiveness in sport groups: Interpretations and considerations," Journal of Sport psychology, vol. 4, no. 2, pp. 123–138, 1982.
[40] W. N. Widmeyer, L. R. Brawley, and A. V. Carron, The measurement of cohesion in sport teams: The Group Environment Questionnaire. Sports Dynamics, 1985.
[41] R. R. Grand and A. V. Carron, "Development of a team climate questionnaire," in Proceedings of the Annual Conference of the Canadian Society for Psychomotor Learning and Sport Psychology, Edmonton, Alberta, pp. 217–229, 1982.
[42] M. A. Eys, J. Hardy, A. V. Carron, and M. R. Beauchamp, "The relationship between task cohesion and competitive state anxiety," Journal of Sport and Exercise Psychology, vol. 25, no. 1, pp. 66–76, 2003.
[43] S. A. Kozub and C. J. Button, "The influence of a competitive outcome on perceptions of cohesion in rugby and swimming teams," International Journal of Sport Psychology, vol. 31, no. 1, pp. 82–95, 2000.
[44] S. A. Kozub and J. F. McDonnell, "Exploring the relationship between cohesion and collective efficacy in rugby teams," Journal of sport behavior, vol. 23, no. 2, pp. 120–129, 2000.
[45] F. Li and P. Harmer, "Confirmatory factor analysis of the group environment questionnaire with an intercollegiate sample," Journal of Sport and Exercise Psychology, vol. 18, no. 1, pp. 49–63, 1996.
[46] P. J. Sullivan, S. E. Short, and K. M. Cramer, "Confirmatory factor analysis of the group environment questionnaire with co-acting sports," Perceptual and Motor Skills, vol. 94, no. 1, pp. 341–347, 2002.
[47] J.-P. Heuzé and P. Fontayne, "Questionnaire sur l'ambiance du groupe: A french-language instrument for measuring group cohesion," Journal of Sport and Exercise Psychology, vol. 24, no. 1, pp. 42–67, 2002.
[48] W. Boughattas and N. Kridis, "Validation transculturelle d'une mesure de cohésion au sein d'une population tunisienne. "questionnaire d'ambiance de groupe (heuzé et fontayne, 2002)"," Pratiques Psychologiques, vol. 22, no. 3, pp. 301–315, 2016.
[49] G. Andreaggi, C. Robazza, and L. Bortoli, "Coesione sociale e sul compito negli sport di squadra: il "group environment questionnaire"," Giornale Italiano di Psicologia dello Sport, vol. 2, pp. 19–23, 2000.
[50] K. S. Courneya, "Understanding readiness for regular physical activity in older individuals: An application of the theory of planned behavior," Health psychology, vol. 14, no. 1, pp. 80–87, 1995.
[51] P. A. Estabrooks and A. V. Carron, "The physical activity group environment questionnaire: An instrument for the assessment of cohesion in exercise classes," Group Dynamics, vol. 4, no. 3, pp. 230–243, 2000.
[52] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker, "Language style matching as a predictor of social dynamics in small groups," Communication Research, vol. 37, no. 1, pp. 3–19, 2010.
[53] L. Giraldo and K. Passino, "Dynamic task performance, cohesion, and communications in human groups," IEEE Transactions on Cybernetics, vol. 46, no. 10, pp. 2207—2219, 2016.
[54] S. Ghosh, A. Dhall, N. Sebe, and T. Gedeon, "Predicting group cohesiveness in images," in 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, 2019.
[55] U. Kubasova, G. Murray, and M. Braley, "Analyzing verbal and nonverbal features for predicting group performance," in Proc. Interspeech 2019, pp. 1896–1900, ISCA, 2019.
[56] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," IEEE Transactions on Multimedia, vol. 12, no. 6, pp. 563–575, 2010.
[57] 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 Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 206—-215, Association for Computing Machinery, 2017.
[58] H. Hung and G. Chittaranjan, "The idiap wolf corpus: exploring group behaviour in a competitive role-playing game," in Proceedings of the 18th ACM international conference on Multimedia, pp. 879–882, 2010.
[59] D. Olguin and A. Pentland, "Sensor-based organisational design and engineering," International Journal of Organisational Design and Engineering, vol. 1, no. 1, pp. 69–97, 2010.
[60] Y. Zhang, J. Olenick, C.-H. Chang, S. Kozlowski, and H. Hung, "Teamsense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors,"

Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 3, pp. 1–22, 2018.

[61] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 273–291, 2014.

[62] L. F. Borja, J. Azorin-Lopez, and M. Saval-Calvo, "A compilation of methods and datasets for group and crowd action recognition," International Journal of Computer Vision and Image Processing (IJCVIP), vol. 7, no. 3, pp. 40–53, 2017.

[63] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social interaction capture," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 1, pp. 190–204, 2019.

[64] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," IEEE Transactions on Affective Computing, pp. 1–1, 2018.

[65] X. Alameda-Pineda, R. Subramanian, E. Ricci, O. Lanz, and N. Sebe, "Chapter 14 - salsa: A multimodal dataset for the automated analysis of free-standing social interactions," in Group and Crowd Behavior for Computer Vision (V. Murino, M. Cristani, S. Shah, and S. Savarese, eds.), pp. 321–340, Academic Press, 2017.

[66] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–4, 2009.

[67] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305–317, 2005.

[68] L. Chen, R. T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang, "Vace multimodal meeting corpus," in Machine Learning for Multimodal Interaction. MLMI 2005. Lecture Notes in Computer Science (S. Renals and S. Bengio, eds.), vol. 3869, pp. 40–51, Springer Berlin Heidelberg, 2006.

[69] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), 2018.

[70] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. 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 Transactions on Affective Computing, pp. 1–1, 2018.

[71] A. Čereković, "An insight into multimodal databases for social signal processing: acquisition, efforts, and directions," Artificial Intelligence Review, vol. 42, no. 4, pp. 663–692, 2014.

[72] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," Int'l. Conf. on Methods and Techniques in Behavioral Research, 2005.

[73] S. Stashevsky and M. Koslowsky, "Leadership team cohesiveness and team performance," International Journal of Manpower, vol. 27, no. 1, pp. 63–74, 2006.

[74] G. Freedman and M. Flanagan, "From dictators to avatars: Furthering social and personality psychology through game methods," Social and personality psychology compass, vol. 11, no. 12, p. e12368, 2017.

[75] E. Redcay and L. Schilbach, "Using second-person neuroscience to elucidate the mechanisms of social interaction," Nature Reviews Neuroscience, vol. 20, no. 8, pp. 495–505, 2019.

[76] E. Van Dijk, C. K. W. De Dreu, and J. Gross, "Power in economic games," Current opinion in psychology, vol. 33, pp. 100–104, 2020.

[77] C. Bonillo, T. Romão, and E. Cerezo, "Persuasive games in interactive spaces: The hidden treasure game," in Proceedings of the XX International Conference on Human Computer Interaction, pp. 1–8, 2019.

[78] R. Niewiadomski, M. Mancini, T. Baur, G. Varni, H. Griffin, and M. Aung, "MMLI: Multimodal multiperson corpus of laughter in interaction," in Human Behavior Understanding, pp. 184–195, Springer International Publishing, 2013.

[79] D. Codish and G. Ravid, "Personality based gamification-educational gamification for extroverts and introverts," in Proceedings of the 9th CHAIS Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era, vol. 1, pp. 36–44, 2014.

[80] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work?–a literature review of empirical studies on gamification," in 2014 47th Hawaii international conference on system sciences, pp. 3025–3034, IEEE, 2014.

[81] O. Nov and O. Arazy, "Personality-targeted design: theory, experimental procedure, and preliminary results," in Proceedings of the 2013 conference on Computer supported cooperative work, pp. 977–984, 2013.

[82] B. Mullen and C. Copper, "The relation between group cohesiveness and performance: An integration," Psychological Bulletin, vol. 115, no. 2, pp. 210–227, 1994.

[83] R. Grossman, Z. Rosch, D. Mazer, and E. Salas, "What matters for team cohesion measurement? A Synthesis," Research on Managing Groups and Teams, vol. 17, pp. 147–180, 2015.

[84] R. T. Keller, "Predictors of the performance of project groups in R&D organizations," Academy of Management Journal, vol. 29, no. 4, pp. 715–726, 1986.

[85] N. Bendermacher, "Beyond alpha: Lower bounds for the reliability of tests," Journal of Modern Applied Statistical Methods, vol. 9, no. 1, p. 11, 2010.

[86] D. McNeish, "Thanks coefficient alpha, we'll take it from here.," Psychological Methods, vol. 23, no. 3, p. 412, 2018.

[87] G. Y. Peters, "The alpha and the omega of scale reliability and validity: why and how to abandon cronbach's alpha," European Health Psychologist, vol. 16, no. S, p. 576, 2014.

[88] W. Revelle and R. E. Zinbarg, "Coefficients alpha, beta, omega, and the glb: Comments on sijtsma," Psychometrika, vol. 74, no. 1, p. 145, 2009.

[89] K. Sijtsma, "On the use, the misuse, and the very limited usefulness of cronbach's alpha," Psychometrika, vol. 74, no. 1, p. 107, 2009.

[90] J. M. F. Ten Berge and G. Sočan, "The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality," Psychometrika, vol. 69, no. 4, pp. 613–625, 2004.

[91] I. Trizano-Hermosilla and J. M. Alvarado, "Best alternatives to cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements," Frontiers in Psychology, vol. 7, 2016.

[92] S. A. Carless and C. De Paola, "The measurement of cohesion in work teams," Small Group Research, vol. 31, no. 1, pp. 71–88, 2000.

[93] M. D. Michalisin, S. J. Karau, and C. Tangpong, "Top management team cohesion and superior industry returns: An empirical study of the resource-based view," Group & Organization Management, vol. 29, no. 1, pp. 125–140, 2004.

[94] J. I. Aragonés, L. Poggio, V. Sevillano, R. Pérez-López, and M.-L. Sánchez-Bernardos, "Measuring warmth and competence at inter-group, interpersonal and individual levels / medición de la cordialidad y la competencia en los niveles intergrupal, interindividual e individual," International Journal of Social Psychology, vol. 30, no. 3, pp. 407–438, 2015.

[95] S. Fiske, A. Cuddy, and P. Glick, "Universal dimensions of social cognition: Warmth and competence," Trends in cognitive sciences, vol. 11, no. 2, pp. 77–83, 2007.

[96] E. Menesini, F. Tassi, and A. Nocentini, "The competitive attitude scale (CAS): a multidimensional measure of competitiveness in adolescence," Journal of Psychology & Clinical Psychiatry, vol. 9, no. 3, pp. 240–244, 2018.

[97] I. J. Roseman, "A model of appraisal in the emotion system," Appraisal processes in emotion: Theory, methods, research, pp. 68–91, 2001.

[98] I. Roseman and C. Smith, "Appraisal theory," Appraisal processes in emotion: Theory, methods, research, pp. 3–19, 2001.

[99] I. J. Roseman, "Appraisal in the emotion system: Coherence in strategies for coping," Emotion Review, vol. 5, no. 2, pp. 141–149, 2013.

[100] F. Gerpott, N. Lehmann-Willenbrock, S. Voelpel, and M. Vugt, "It's not just what is said but also when it's said: A temporal account of verbal behaviors and emergent leadership in self-managed teams," The Academy of Management Journal, vol. 62, no. 3, pp. 717–738, 2018.

[101] K. Lanaj and J. R. Hollenbeck, "Leadership over-emergence in self-managing teams: The role of gender and countervailing biases," Academy of Management Journal, vol. 58, no. 5, pp. 1476–1494, 2015.

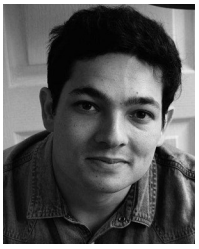
[102] E. J. McClean, S. R. Martin, K. J. Emich, and C. T. Woodruff, "The social consequences of voice: An examination of voice type and gender on status and subsequent leader emergence," Academy of Management Journal, vol. 61, no. 5, pp. 1869–1891, 2018.

[103] E. McAuley, T. Duncan, and V. V. Tammen, "Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A

confirmatory factor analysis," Research Quarterly for Exercise and Sport, vol. 60, no. 1, pp. 48–58, 1989.

[104] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the use of exploratory factor analysis in psychological research.," Psychological methods, vol. 4, no. 3, p. 272, 1999.

[105] F. P. Morgeson, D. S. DeRue, and E. P. Karam, "Leadership in teams: A functional approach to understanding leadership structures and processes," Journal of Management, vol. 36, no. 1, pp. 5–39, 2010.



LUCIEN MAMAN received both his MSc degree in Software Engineering for Technical Computing from Cranfield University, U.K. and his engineering diploma from ESTIA, France in 2017. After working for 2 years as a video engineer at Grabyo, U.K., he is now working towards his PhD degree at the LTCI, Télécom Paris, Institut Polytechnique de Paris, France. His research interests include: social signal processing, emergent states, cohesion, machine learning and affective computing.

ELEONORA CECCALDI is a Computer Science Phd student at CasaPaganini-InfoMus, DIBRIS, University of Genoa. Her background is Cognitive Psychology and her main areas of expertise and interest include event segmentation, unitizing, cognitive models of emotions and social interaction. She is member of the Italian Association for Cognitive Sciences.

NALE LEHMANN-WILLENBROCK is Full Professor and department head of Industrial/Organizational Psychology and leads the TeamLab at the University of Hamburg, Germany. Previously, she was an Associate Professor at the University of Amsterdam, The Netherlands. She holds a PhD in Psychology from Technical University Braunschweig (2012). She investigates team processes and leader-follower dynamics during organizational meetings and other interaction settings using pattern analytical methods and promotes interdisciplinary collaborations that bridge social and computer science. She currently serves as Associate Editor at Small Group Research.

LAURENCE LIKFORMAN-SULEM is graduated in engineering from ENST-Bretagne (Ecole Nationale Superieure des Telecommunications) in 1984, received her PhD from ENST-Paris in 1989 and her HDR (Habilitation a Diriger des Recherches) from Pierre & Marie Curie University in 2008. She is an Associate Professor at Telecom ParisTech in the IDS (Image Data Signal) Department since 1991 where she serves as a senior instructor in Pattern Recognition and handwriting recognition. She chaired the program committee of CIFED held in Fribourg, Switzerland, in 2006, the program committees of two DRR Conferences (Document Recognition and Retrieval) held in 2009 and 2010 in San Jose, California, and the ASAR (Arabic and derived Script Analysis and Recognition) workshop held in 2017 in Nancy, France.

MOHAMED CHETOUANI is the head of the PIRoS (Perception, Interaction et Robotique Sociales) research team at the Institute for Intelligent Systems and Robotics (CNRS UMR 7222), Sorbonne University. He is currently a Full Professor in signal processing and machine learning for human-machine interaction. His activities cover social signal processing, social robotics and interactive machine learning with applications in psychiatry, psychology, social neuroscience and education. He is the Deputy Director of the Laboratory of Excellence SMART Human/Machine/Human Interactions In The Digital Society. Since 2018, he is the coordinator of the ANIMATAS H2020 Marie Sklodowska Curie European Training Network. He is the local co-chair of IEEE ICRA 2020 (Paris) and Program co-chair of ICMI 2020 (Utrecht). Since 2020, he is the President of the Ethical Advisory Board of Sorbonne University.

GUALTIERO VOLPE received the M.Sc. degree in computer engineering in 1999 and the Ph.D. in electronic and computer engineering in 2003 from the University of Genoa, Italy. Since 2014, he is an Associate Professor at DIBRIS, University of Genoa. His research interests include intelligent and affective human-machine interaction, social signal processing, sound and music computing, modeling and real-time analysis of expressive content, and multimodal systems.

GIOVANNA VARNI is an Associate Professor at LTCI, Télécom Paris, Institut polytechnique de Paris, France. She received her PhD from the Università degli Studi di Genova (Genova, Italy) in 2009. She mainly investigates on Social Signal Processing (SSP) and Human Computer Interaction. She was involved in several EU FP7-FP6 projects, and she is currently PI of the national French project ANR JCJC GRACE (2019-2022) on the automated analysis of cohesion in small groups of humans.