Training Computational Models of Group Processes without Groundtruth: the Self- vs External Assessment's Dilemma

LUCIEN MAMAN, LTCI, Télécom Paris, Institut polytechnique de Paris, France
GUALTIERO VOLPE, DIBRIS, Univeristà degli Studi di Genova, Italy
GIOVANNA VARNI, LTCI, Télécom Paris, Institut polytechnique de Paris, France

Supervised learning relies on the availability and reliability of the labels used to train computational models. In research areas such as Affective Computing and Social Signal Processing, such labels are usually extracted from multiple self- and/or external assessments. Labels are, then, either aggregated to produce a single groundtruth label, or all used during training, potentially resulting in degrading performance of the models. Defining a "true" label is, however, complex. Labels can be gathered at different times, with different tools, and may contain biases. Furthermore, multiple assessments are usually available for a same sample with potential contradictions. Thus, it is crucial to devise strategies that can take advantage of both self- and external assessments to train computational models without a reliable groundtruth. In this study, we designed and tested 3 of such strategies with the aim of mitigating the biases and making the models more robust to uncertain labels. Results show that the strategy based on weighting the loss during training according to a measure of disagreement improved the performances of the baseline, hence, underlining the potential of such an approach.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Group Dynamics; Multimodal Interaction; Cohesion; Self and External Assessment

ACM Reference Format:
Lucien Maman, Gualtiero Volpe, and Giovanna Varni. 2022. Training Computational Models of Group Processes without Groundtruth: the Self- vs External Assessment's Dilemma. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536220.3563687

1 INTRODUCTION

Training supervised computational models implies the availability of labels usually provided by multiple human raters. Collecting these labels is a long and costly process that is not without some issues. Aggregating labels from multiple raters (e.g., via majority voting, average, and so on) to produce a single groundtruth label, or training a model directly by exploiting all the labels could result in poor model performance. To tackle the problem of building a groundtruth from multiple annotators, Wang and colleagues recently proposed a novel agreement learning framework composed of 2 streams [27]. The first stream is a classifier that learns from all the annotators while the second one produces regularization information to the classifier based on the estimated agreement between annotators. Such a framework helps the classifier to tune its decision by incorporating agreement information.

Research areas such as Social Signal Processing [26] and Affective Computing [20] face further issues: traditionally, indeed, labels can be gathered at different times (online vs. offline), by different raters (the participants to the data collection – self-assessment – or external raters), and with different tools (questionnaires vs. coding schemes). The time at which labels are collected matters [8]. Online labels are produced while the interaction

ICMI '22 Companion, November 7–11, 2022, Bengaluru, India
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India*, https://doi.org/10.1145/3536220.3563687.

occurs, and they can be obtained by using thinking-aloud approaches (e.g., [9]); offline labels are collected after the interaction occurred, that is in a retrospective fashion (e.g., [10]). As reported in [25], self-assessments and external assessments have biases. Self-assessments might indeed be over-optimistic since participants tend to provide ratings towards socially desirable characteristics. External assessments reflect the behavior that people adopt toward others, without necessarily corresponding to their true internal state [24]. Both these assessments may also be affected by the ability, sensitivity to the construct under study, and honesty of the raters. These latter issues are less relevant for expert and trained raters. Concerning the tools to gather labels, self-assessments grounds on questionnaires including items expressed in the first person (e.g., the "Group Environment Questionnaire" [6] for cohesion) and/or exploiting a referent-shift composition model (e.g., for group potency [11]). In the latter case, a construct is assessed by each participant taking the group as the referent of the assessment [7]. Questionnaires for external assessments can use a referent-shift composition model as well (e.g., for assessing collective efficacy [28]) or coding schemes decomposing interactions in meaningful small segments (e.g., ACT4Teams [13] for cohesion and TRAWIS [5] for transactive knowledge and knowledge exchange).

A major challenge of Social Signal Processing is dealing with groups [26]. Investigating groups makes the complexity of such issues increase dramatically, especially because multiple self- and external assessments are available for the same samples. Since these assessments can catch different nuances of group processes, one can expect that using them together could be beneficial to mitigate the biases they introduce and to solve the cases in which the groundtruth is uncertain (i.e., labels from self- and external assessments differ) during the training of supervised computational models. This topic, however, is still under-investigated. Most of the computational studies on group processes rely on external assessments only (e.g., [12, 19, 21]) as they can be collected *a posteriori* whereas only a few datasets also provide self-assessments of group processes (e.g., group emotion in AMIGOS [17] or cohesion in GAME-ON [14]). Only a very few studies exploited both types of assessments in making their models, but without combining the labels together. For example, in [23] and [4], the authors show that self- and external perceptions of social stress are both close but significantly different and they explore how these different assessments impacted the classification performances of a Support Vector Machine (SVM). In both cases, they reached excellent performances (i.e., F1-scores of 0.90 ±0.01 and 0.87 ±0.02 for external and self-assessments, respectively).

In this study, we present 3 strategies to address uncertain groundtruth at different stages of a computational model pipeline. Thereon, experiments on a dataset on group cohesion are run.

2 STRATEGIES

2.1 Training models with labels extracted from both self- and external assessments (S1)

The first strategy can be applied when scores are available for both the self- as well as the external assessments for each sample. An exemplary case is when both use questionnaires. All the scores are first concatenated in a vector, next they are aggregated in a single value by means of some functional, and finally, the resulting score is mapped to a label. Without any *a priori* knowledge of the process producing the scores, the functional is supposed to apply equal footing. This strategy is suited to both binary and multiclass classification and regression too.

2.2 Training models with the most reliable labels according to Social Sciences' insights (S2)

The second strategy can be applied when labels reflect a positive and a negative outcome in the group process. Positive group outcomes are associated with positive labels, negative group outcomes are associated with negative ones. Labels from both self- and external assessments are obtained independently by using questionnaires and/or annotation schemes. Based on these labels, the aim is to define the most reliable one according to Social Sciences' insights on the group process of study. Three cases can occur: (1) both assessments provide the same label; (2) the label extracted from self-assessments is positive, while the label from external assessments is negative; and

(3) the label extracted from self-assessments is negative, while the label obtained form external assessments is positive. In the first case, the label is retained as is. In the second case, the label produced by external assessments is retained. We ground this choice based on Vinciarelli and Mohammadi's study [25] stating that when persons assess themselves, they tend to provide ratings towards socially desirable characteristics (e.g., taking leadership in the group). Thus, we select the negative label from external assessments to limit such a bias. In the third case, the negative label produced by self-assessments is retained as external raters usually do not have the full context and outcomes of the interaction, hence, having no information about the success or failure of a particular task. According to Mullen and Copper [18], performance, indeed, has a stronger effect on group processes than the group processes themselves on performance. Furthermore, Boone and colleagues [3] show that failure has a stronger effect on group processes than success. Successes, indeed, only maintain the level of a particular group process without necessarily positively impacting it. Hence, because of the negative impact of failure and the limited positive impact of success on the dynamics of various group processes, we select the negative label computed from the self-assessments, relying on group members' feelings and knowledge about the task's success and context. This strategy, as presented here, applies to binary classification, but it can be possibly extended and refined to deal with multiclass classification when labels reflect the intensity of the positive and negative outcomes.

2.3 Training models by weighting the loss function according to disagreement in the labels (S3)

The third strategy can be applied during training by acting on the loss function. It consists of weighting the loss of the samples for which labels differ, according to the amount of disagreement between the scores computed from both types of assessments. The higher the disagreement is, the smaller the sample is weighted. In that way, we let the model learn from every sample while limiting the impact of uncertain labels. This strategy can be applied when scores are available. To compute the amount of disagreement, we used the Bhattacharyya distance [2]. Such a distance is in $[0, +\infty[$ and measures the similarity between 2 distributions (here the scores from self- and external assessments), as in Equation 1:

$$D(p, q) = -\ln(BC(p, q)) \tag{1}$$

where BC is the Bhattacharyya coefficient for the scores computed from self- and external assessments. For each assessment, scores are aggregated in a histogram, $p(x)$ and $q(x)$ for self-and external assessments respectively. The 2 histograms must have the same number of bins. The BC coefficient is computed as in Equation 2:

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{2}$$

where X is the discrete random variable expressing the score, and x is a particular outcome.
Finally, the weight of the sample j is computed as $w_j = \frac{1}{1-D(p(j),q(j))}$. For the sake of clarity, Equation 3 shows how the loss function is modified in the case of a binary classification with a cross-entropy loss:

$$L = -\frac{1}{N} \sum_{j=1}^{N} w_j \left(t_j log(p_j) + (1 - t_j) log(1 - p_j) \right) \tag{3}$$

where N is the number of samples, w_j is the weight computed for sample j, t_j is the binary label of sample j, and p_j is the probability that sample j has label t_j.

3 EXPERIMENTAL SETTINGS

In this study, we explicitly focus on cohesion, a multidimensional affective group emergent state (i.e., a dynamic construct that characterizes properties of the group and that results from the interactions among group members [16]) that can be defined as the tendency of a group to stick together to pursue goals and/or affective

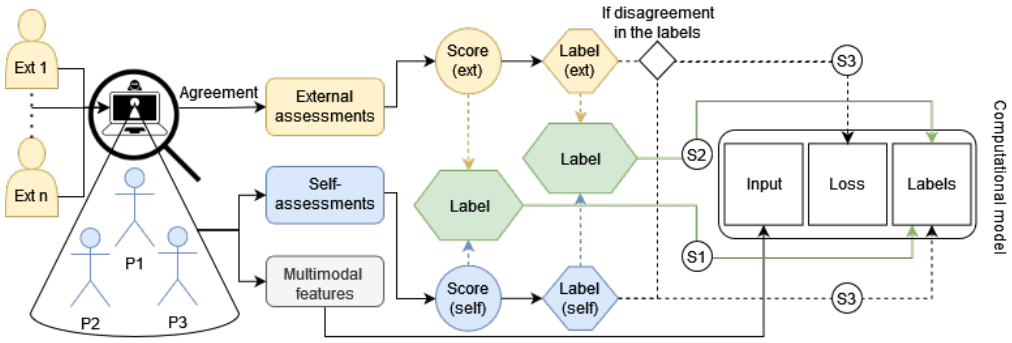


Fig. 1. The 3 strategies applied onto a computational model training pipeline. S1 uses both types of ratings to produce a label. S2 is based on already computed labels and leverages Social Sciences' insights to select the most appropriate one. S3 applies weights to the loss only when assessments differ. Strategies are applied independently and cannot be combined.

needs [6]. In particular, we take into account the Social and Task dimensions of cohesion following Severt and Estrada's theoretical framework [22].

3.1 Dataset

We used audio features and motion capture data from the GAME-ON dataset [14] that makes available 15 group interactions in a social game scenario over 5 tasks. These tasks were designed for eliciting variations of Social and Task cohesion. Before and after each task, self-assessments of cohesion were collected from every member of each group, using a modified version of the Group Environment Questionnaire (GEQ) [6]. GAME-ON also provides external assessments of cohesion. For each group, at least 3 raters evaluated cohesion for each task, using a version of the GEQ with all first-person singular questions transformed into third-person plural. For 12 out of 15 groups, a *good* inter-rater agreement was reached for both the Social and Task dimensions, using $ICC(2,k)$ with a consistency definition. For the remaining 3 groups, a *good* ICC was obtained for only one dimension (i.e., Social cohesion for 2 of the groups and Task cohesion for the third group). We anyway decided to keep them in the analysis.

3.2 Computational model

To evaluate the 3 strategies, we adopted a slightly modified version of the "from Individual to Group" (fItG) model [15] predicting the no decrease / decrease of Social and Task cohesion, in a multilabel fashion, between pairs of consecutive tasks (e.g., the first and the second tasks)[1]. It was fed with 91 multimodal nonverbal individual and group features extracted on consecutive non-overlapping windows of 20s. processes them differently, in distinct modules. This means that individual features are first processed and then combined with the group features to learn a group behavior representation. Each module comprises fully connected and LSTM layers, hence taking the time dependencies between the time windows into account. This model trained on self-assessments only was chosen as the baseline.

[1]We modified the original fItG to take into account that external raters did not have any prior information about the baseline cohesion score of the groups. Concretely, here, we do not predict variations of cohesion between the baseline and the first task.

3.3 Implementation of the strategies

Labels were built from the self- and external ratings as described in [15], that is by computing, for each dimension, the mean rank differences of the cohesion scores, for 2 consecutive tasks as in Equation 4:

$$GS_{T_x} = \frac{1}{n}\sum_{i=1}^{n}\left(rank_{Tx}^{(i)} - rank_{Tx-1}^{(i)}\right) \tag{4}$$

with GS_{T_x} being the group score computed for the transition between the tasks T_x and T_{x-1} ($x \in \{2,3,4,5\}$), n being the number of raters (i.e., set to 3 in case of self-assessments, and set to 3 or 4 for external assessment depending on the number of external raters for the group), and $rank^{(i)}$ being the rank corresponding to the associated GEQ score given by rater i. Finally, this score was binarized: a value equal to 0 is assigned when the group score is negative (i.e., a decrease in cohesion occurred), whereas a value equal to 1 is assigned when the group score is 0 or positive (i.e., no change or an increase in cohesion occurred).

The first strategy exploits the labels computed as mentioned above. The second strategy takes no decrease as the positive outcome and decrease as the negative outcome. The third strategy uses self-assessments as groundtruth and, in case of disagreement between the labels extracted from self- and external assessments, computes the histogram of the group scores on 6 bins for each dimension to determine the weight to be used within the following loss function:

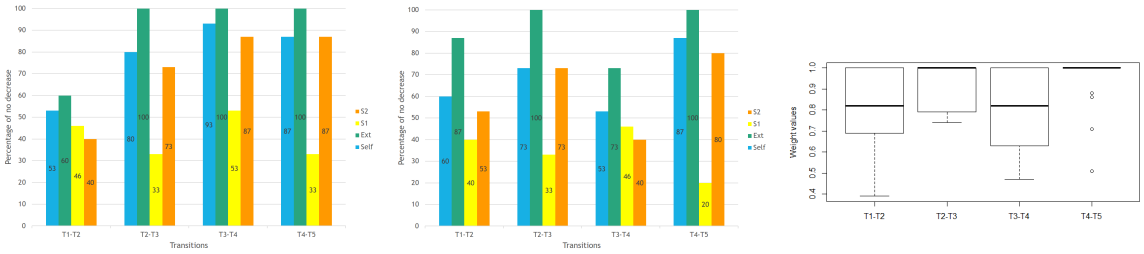
$$L = -\frac{1}{N}\sum_{d\in D}\sum_{t\in T}\sum_{n=1}^{N} w_{t,n}\left(t_{d,t,n}log(p_{d,t,n}) + (1-t_{d,t,n})log(1-p_{d,t,n})\right) \tag{5}$$

with $D = \{Social, Task\}$ (i.e., both dimensions predicted in a multilabel setting), $T = \{T1-T2, T2-T3, T3-T4, T4-T5\}$ (i.e., all the transitions predicted in a multitask setting) and N, the number of samples. Here, the weight $w_{t,n}$ is only dependent on the transition t and the sample n as we extracted, for each t and n, the Bhattacharyya distances from both Social and Task scores' distributions and we took their average. Also, $t_{d,t,n}$ is the binary label associated to sample n for dimension d in transition t whilst $p_{d,t,n}$ is the probability that sample n of dimension d in transition t has label $t_{d,t,n}$.

Applying strategies S1 and S2 resulted in different labels' distributions over all the tasks with respect to those ones obtained from self- and external assessments taken on their own (see Figure 2(a) and Figure 2(b)). By applying S3, we obtained per-transition weights distributed as displayed in Figure 2(c). These were computed based on the average of the Social and Task Bhattacharyya distances: $W_2 = 0.82 \pm 0.21$, $W_3 = 0.91 \pm 0.11$, $W_4 = 0.81 \pm 0.19$ and $W_5 = 0.93 \pm 0.14$. As Figure 2(c) shows, weights have a small impact on transition T4 - T5 (most samples are given a weight close to 1), whereas the impact is stronger on transitions T1 - T2 and T3 - T4.

4 TRAINING AND EVALUATION

We trained multiple versions of the fItG model, each implementing one of the 3 strategies. A nested Leave-One-Group-Out (LOGO) cross-validation was carried out to account for the high diversity of groups. We split the 15 groups into training, validation, and test sets with 10, 4, and 1 group(s), respectively. Then, we augmented the training set as follows: 1) adding Gaussian noise (x4 the training set); 2) permuting the order of the group members (x6 the training set). In total, we obtained 240 groups (i.e., $4 \times 6 \times 10$). Models were trained up to 500 epochs with a fixed learning rate of 0.001 and the weights of the models were updated at every mini-batch composed of 4 groups. Every 10 epochs, they were evaluated on the validation set. In that way, we determined the optimal number of epochs based on the performances across the 4 transitions. Then, we retrained the models (merging both training and validation sets) that were obtained by using the optimal number of epochs. The first and the second strategies used a weighted binary cross entropy loss for the prediction of Social and Task dynamics of cohesion for each transition that takes the distribution of the labels into account (i.e., a larger weight



(a) Social cohesion (b) Task cohesion (c) Weights per transition

Fig. 2. Distribution of the no decrease labels, for each transition and for Social (Figure 2(a)) and Task (Figure 2(b)) cohesion using self- and external assessment and strategies S1 and S2. Box plots of the weights per transition obtained by applying S3 (Figure 2(c)).

is applied for the under-represented classes). The third strategy applied a further weight as previously described. As in [15], each model was run over 15 random seeds. Finally, we averaged the performances over these seeds. For each dimension, performances were evaluated using the averaged F1-score over the transitions.

We assessed potential significant differences between the performances of the 3 strategies through a computationally intensive randomization test. In detail, we performed a k-sample permutation test. This is a non-parametric test avoiding the independence assumption between the results being compared and that is suitable for non-linear measures such as F1-score [29]. The significance level α was at 0.05. Since we ran multiple comparisons between the models, a post-hoc analysis was carried out using pairwise permutations with an FDR-adjusted p-value [1]. All the models were developed and trained using Python 3.7 and Tensorflow 2.6 on NVIDIA V100 GPUs.

5 RESULTS

We compare the performances between the baseline model trained with the self-assessments only (i.e., fItG_Self) and the models trained with each strategy applied (i.e., fItG_S1, fItG_S2, and fItG_S3). A transition-by-transition analysis is also carried out. Table 1 summarizes the results (average performances obtained by each model over the 15 seeds).

As for the prediction of the Social dimension of cohesion, fItG_Self achieves an averaged F1-score of 0.69 ±0.02, while fItG_S1 obtains 0.53 ±0.04, fItG_S2 reaches 0.67 ±0.04, and fItG_S3 attains 0.74 ±0.04. Statistical analysis shows that there are significant differences in performances between models (p = .001). A post-hoc analysis shows that fItG_S3 significantly outperformed all of the other ones (p = .002 for all the comparisons). No significant difference is found between fItG_Self and fItG_S2. They, however, both obtained significantly better performances than fItG_S1 (p = .002 for both comparisons). Looking at the per-transition performances, we ran 4-sample permutation tests, showing significant differences between the models (i.e., p = .002 for transition T1 - T2, and p = .001 for T2 - T3, T3 - T4, and T4 - T5, respectively). Post-hoc analyses revealed that fItG_S3 obtains better performances for each of the transitions except T4 - T5 with respect to fItG_S1 (p = .006 in transition T1 - T2, p = .018 in T2 - T3, and p = .002 in T3 - T4) and fItG_S2 (p = .050 in transition T1 - T2, p = .018 in T2 - T3, and p = .002 in T3 - T4), but it only significantly improves the performances in transition T1 - T2 (p = .006) with respect to fItG_Self (see Table 1).

About the prediction of the Task dimension of cohesion, fItG_Self obtains an averaged F1-score of 0.61 ±0.03, fItG_S1 reaches 0.62 ±0.07, fItG_S2 attains 0.55 ±0.06, and fItG_S3 achieves an averaged F1-score of 0.64 ±0.04. Statistical analysis shows significant differences between the models' performances (p = .001). A post-hoc analysis shows that fItG_S3 significantly outperforms fItG_Self, fItG_S1, and fItG_S2 (p = .002, for the 3 comparisons).

Table 1. Summary of the averaged overall and per-transition F1-scores obtained for each dimension by fItG_Self, fItG_S1, fItG_S2, and fItG_S3. Significantly higher F1-score(s), for a particular transition and dimension, are highlighted in bold.

	F1-score±std							
	Social				**Task**			
	fItG_Self	**fItG_S1**	**fItG_S2**	**fItG_S3**	**fItG_Self**	**fItG_S1**	**fItG_S2**	**fItG_S3**
T1 - T2	0.43 ±0.10	0.46 ±0.07	0.51 ±0.10	**0.60 ±0.12**	**0.54 ±0.10**	**0.59 ±0.08**	0.48 ±0.11	**0.60 ±0.11**
T2 - T3	**0.66 ±0.06**	0.56 ±0.10	0.59 ±0.10	**0.67 ±0.06**	0.56 ±0.11	**0.65 ±0.12**	0.57 ±0.11	**0.70 ±0.10**
T3 - T4	**0.85 ±0.03**	0.50 ±0.14	0.75 ±0.05	**0.87 ±0.02**	0.56 ±0.09	0.58 ±0.13	0.46 ±0.14	0.54 ±0.08
T4 -T5	**0.81 ±0.04**	0.59 ±0.12	**0.86 ±0.06**	**0.83 ±0.05**	**0.77 ±0.06**	0.64 ±0.14	0.69 ±0.07	**0.74 ±0.03**
Average	0.69 ±0.02	0.53 ±0.04	0.67 ±0.04	**0.74 ±0.04**	0.61 ±0.03	0.62 ±0.07	0.55 ±0.06	**0.64 ±0.04**

fItG_S2, however, significantly under-performs compared to fItG_Self and fItG_S1 ($p = .002$, for both comparisons). Concerning the per-transition analysis, we performed 4-sample permutation tests showing that statistically significant differences exist in transitions T1 - T2 ($p = .012$), T2 - T3 ($p = .004$), and T4 - T5 ($p = .002$). Post-hoc analysis reveals that fItG_S3 outperforms fItG_Self in transition T2 - T3 ($p = .018$). No significant differences are found with fItG_Self for transitions T1 - T2 and T4 - T5 despite the significant differences with the other models. No significant differences between strategies and fItG_Self were found in transition T3 - T4.

6 DISCUSSION AND CONCLUSIONS

The application of S3 achieved the best results in terms of an increase in model performances in our experimental settings suggesting that intervening on the loss function is a viable approach for taking into account the contribution of self- and external assessments. Such results highlight that changing the labeling distributions with the aim of defining a "true" unbiased label (see S1 and partially S2) may not be beneficial for the model. Group processes are, indeed, complex to assess and disagreements might also exist between group members, making it particularly hard to compute a trustworthy label. The overall improvement of performance observed by applying S3 suggests that potential disagreements should be taken into account during training, preventing it from giving too much importance to uncertain samples. Analysis of per transition performances shows that such an increase in performances concerns some specific transitions. These transitions are characterized by poor performances using self-assessments only, meaning that separating classes is not trivial. This confirms that S3 helps the model to simplify class separation. While this work is, to the best of our knowledge, the first attempt at leveraging both types of assessments, our strategies remain to be tested on different datasets and group processes. Further labeling strategies can be envisaged too. Moreover, in this study we did not train and test a model using external assessments only which could provide additional insights. Unfortunately, the distributions of labels computed from external assessment did not allow us to proceed because of several transitions (i.e., T2 - T3, and T4 - T5 for the Task dimension, and T2 - T3, T3 - T4, and T4 - T5 for the Social dimension) for which all of the labels belonged to the same class (i.e., no decrease of cohesion). Also, further investigation on S3 could help defining an optimal measure of disagreement.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[2] Anil Bhattacharyya. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society* 35 (1943), 99–109.

[3] Kathy S Boone, Patricia Beitel, and Jolynn S Kuhlman. 1997. The effects of the win/loss record on cohesion. *Journal of Sport Behavior* 20, 2 (1997), 125–134.

[4] Nadège Bourvis, Aveline Aouidad, Michel Spodenkiewicz, Giuseppe Palestra, Jonathan Aigrain, Axel Baptista, Jean-Jacques Benoliel, Mohamed Chetouani, and David Cohen. 2021. Adolescents with borderline personality disorder show a higher response to stress but a lack of self-perception: Evidence through affective computing. *Progress in Neuro-psychopharmacology and Biological Psychiatry* 111 (2021), 110095.

[5] Elisabeth Brauner. 2018. TRAWIS: Coding transactive knowledge and knowledge exchange. (2018), 575–582.

[6] Albert V Carron, W Neil Widmeyer, and Lawrence R Brawley. 1985. The Development of an Instrument to Assess Cohesion in Sport Teams: The Group Environment Questionnaire. *Journal of Sport Psychology* 7, 3 (1985), 244–266.

[7] David Chan. 1998. Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of applied psychology* 83, 2 (1998), 234–246.

[8] Sidney K D'Mello, Scotty D Craig, and Art C Graesser. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology* 4, 3-4 (2009), 165–187.

[9] Sidney K. D'Mello, Scotty D. Craig, Jeremiah Sullins, and Arthur C. Graesser. 2006. Predicting Affective States Expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. *Int. J. Artif. Intell. Ed.* 16, 1 (2006), 3–28.

[10] AC Graesser, Bethany McDaniel, Patrick Chipman, Amy Witherspoon, Sidney D'Mello, and Barry Gholson. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th annual meetings of the cognitive science society*. 285–290.

[11] Richard A Guzzo, Paul R Yost, Richard J Campbell, and Gregory P Shea. 1993. Potency in groups: Articulating a construct. *British journal of social psychology* 32, 1 (1993), 87–106.

[12] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[13] Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L. Meinecke. 2018. *The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme.* Cambridge University Press, Chapter 21, 422–431.

[14] Lucien Maman, Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Laurence Likforman-Sulem, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. 2020. GAME-ON: A Multimodal Dataset for Cohesion and Group Analysis. *IEEE Access* 8 (2020), 124185–124203.

[15] Lucien Maman, Laurence Likforman-Sulem, Mohamed Chetouani, and Giovanna Varni. 2021. Exploiting the Interplay between Social and Task Dimensions of Cohesion to Predict Its Dynamics Leveraging Social Sciences. In *Proceedings of the 23rd International Conference on Multimodal Interaction*. 16–24.

[16] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *The Academy of Management Review* 26, 3 (2001), 356–376.

[17] J. A. Miranda Correa, M. K. Abadi, N. Sebe, and I. Patras. 2018. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* (2018), 479–493.

[18] Brian Mullen and Carolyn Copper. 1994. The relation between group cohesiveness and performance: An integration. *Psychological Bulletin* 115, 2 (1994), 210–227.

[19] Philipp Matthias Muller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *Proceedings of the 21st International Conference on Multimodal Interaction*. ACM Digital Library, 274–278.

[20] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[21] Dairazalia Sanchez-Cortes, Oya Aran, Dinesh Babu Jayagopi, Marianne Schmid Mast, and Daniel Gatica-Perez. 2013. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces* 7, 1 (2013), 39–53.

[22] Jamie B Severt and Armando X Estrada. 2015. On the Function and Structure of Group Cohesion. In *Team Cohesion: Advances in Psychological Theory, Methods and Practice*. Vol. 17. Emerald Group Publishing Limited, 3–24.

[23] Michel Spodenkiewicz, Jonathan Aigrain, Nadège Bourvis, Séverine Dubuisson, Mohamed Chetouani, and David Cohen. 2018. Distinguish self-and hetero-perceived stress through behavioral imaging and physiological features. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 82 (2018), 107–114.

[24] James S Uleman, S Adil Saribay, and Celia M Gonzalez. 2008. Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology* 59 (2008), 329–360.

[25] A. Vinciarelli and G. Mohammadi. 2014. A Survey of Personality Computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291.

[26] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[27] Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lun Lam, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2021. AgreementLearning: An End-to-End Framework for Learning with Multiple Annotators without Groundtruth. *arXiv preprint arXiv:2109.03596*

(2021).

[28] Carl B Watson, Martin M Chemers, and Natalya Preiser. 2001. Collective efficacy: A multilevel analysis. *Personality and social psychology bulletin* 27, 8 (2001), 1057–1068.

[29] Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. *The 18th International Conference on Computational Linguistics (COLING)* 2 (2000), 947–954.