ISCI, Volume *16*

Supplemental Information

Tracking the Leader: Gaze Behavior in Group Interactions

Francesca Capozzi, Cigdem Beyan, Antonio Pierro, Atesh Koul, Vittorio Murino, Stefano Livi, Andrew P. Bayliss, Jelena Ristic, and Cristina Becchio

SUPPLEMENTAL INFORMATION

Transparent Methods

Participants

We choose a sample size of 64 based on a power analysis on pilot data (power analysis: r = .4, α = .05, β = .05, Faul et al., 2007; see also the Supplemental Analyses below). All participants (44 females, 20 males; mean age = 21.59 years, age range = 19-29) were naïve to the purpose of the experiment, provided written informed consent, and were compensated with 8 Euros for their participation. The study was conducted in accordance with the ethical principles of the World Medical Association declaration of Helsinki 2008 and was approved by the Bio-Ethical Committee of the University of Torino.

Assessment of individual dispositions. Group composition and leadership designation were based on participants' individual dispositions to leadership and to leadership styles. Those dispositions were determined with reference to a larger pool of participants six months before the experiment, when voluntary students of the University of Torino (N = 274; 211 females, 63 males; mean age = 20.59, age range = 19-37) were asked to complete the Systematic method for the Multiple Level Observation of Groups (SYMLOG; Blumberg and Hare, 1999; Polley et al., 1988). The SYMLOG is a comprehensive instrument designed to evaluate individual dispositions along three bipolar dimensions: Dominance vs. Submissiveness; Acceptance vs. Non-acceptance of (Task Orientation of Established) Authority; Friendliness vs Unfriendliness. Based on the median split of the Dominance and Task Orientation scores, participants were identified as potential leaders or potential followers. Potential leaders were further subdivided into democratic leaders and autocratic leaders based on the median split of the Friendliness scores. This procedure was applied to male and female participants separately (see also Supplemental Analyses for the "Composition of groups").

Group composition. Participants were assigned to one of four-person groups, for a total of sixteen groups. Each experimental group was homogenous for gender (5 all-male groups; 11 all-female groups). Eight participants classified as leaders with a democratic leadership style and eight participants classified as leaders with an autocratic leadership style were randomly assigned as 'designated leaders' to one of the sixteen groups. Forty-eight of the potential followers were also randomly assigned to each group (see also Supplemental Analyses for the "Composition of groups").

Procedure

Setting. Four equidistant chairs were placed at the centre of an otherwise non-furnished room (Figure 1 B in the main text). The chairs were placed with a cross displacement, each chair being 127 cm distant from the centre (Figure 1 B.2 in the main text). Four AXISP1346 multi-view streaming cameras (1280x1024 pixels resolution, 20 frame per second frame rate) were fixed to the ceiling at a height of 210 m from the floor and at a distance of approximately 127 cm from the chairs and were used for individual video recording of the upper part of the body (head and shoulders) of each group member. These individual videos were used for VFOA modelling and visual behaviour features extraction. In order to create videos for use in the leadership perception manipulation check, a standard camera (440x1080 pixels resolution, and 25 frame per second frame rate) was placed at an approximate distance of 200 cm from the chairs to capture the whole scene.

Group task. Each group of participants was asked to complete one of two versions of a survival task ("Winter" or "Desert"; Johnson and Johnson, 1994). The task involved rank-ordering 12 ordinary items (e.g., a map, a mirror, a chocolate bar) based on their utility for group-surviving in a hostile environment. The use of pen paper was not allowed; the experimenter repeated the list of items twice before leaving the room. Participants were invited to contribute to the discussion allowing the leader to make a final decision (Johnson and Johnson, 1994). Performance scores were obtained by subtracting the rank given to each item from the optimal rank (established by survival experts). The final score was given by the sum of the absolute values of these differences.

Time-pressure manipulation. To manipulate situational conditions, a time-pressure manipulation was applied (Chirumbolo et al., 2004; De Grada et al., 1999; Kruglanski and Freund, 1983; Pierro et al., 2003). Groups assigned to the *high time-pressure* situation (n = 8) were instructed to perform the assigned task as quickly as possible, with a clear instruction that time was a critical demand to their task. Groups assigned to the *low time-pressure* situation (n =8) were instead encouraged to take their time to reach a decision with no specific time demand. In the high time-pressure condition, twenty minutes after the start of the session, the experimenter entered the room and urged participants to complete the task. In the low time-pressure condition, twenty minutes after the start of the session, the experimenter entered the room and invited participants to complete the task. In both cases, the maximum time allowed was thirty minutes. Post-hoc debriefing sessions confirmed the validity of the time-pressure manipulation, in that participants in the high time-pressure condition (but not those in the low time-pressure condition) reported the they had perceived time as a critical demand to their task.

Post-task questionnaires. At the end of the group task, group members completed three questionnaires: a report of satisfaction and stress, the General Leadership Impression (GLI) scale to measure leadership perception (Lord et al., 1984; Zaccaro et al., 1991), and the Implicit Followership Theories (IFT) scale to assess participants' personal assumptions about followership (Avolio et al., 2009; Sy, 2010).

Analysis of gaze behaviour

Visual Focus of Attention. To ensure an adequate sampling of observations, the videos of the 16 group interactions (N = 64 individual observations) were divided into 75 video-segments (N = 300 individual observations). The video-segmentation was performed such that all segments had approximately the same duration (i.e., between 4 and 5 minutes). In case the duration of a video was not divisible by 4 or 5, the remaining seconds were evenly distributed across the video-segments of that video. This operation resulted in an average duration of 5 minutes (range = 4-6), with the exception of one segment lasting 2 minutes due to processing issues, for an average of 5 segments per group (range = 2-6). The Constrained Local Model (CLM) was applied to each video-segment to detect and track facial landmarks (Cristinacce and Cootes, 2006). To model the frame-by-frame VFOA of each participant, participants' VFOAs were annotated by two annotators for 25600 randomly selected frames (400 frames for each video determined by the confidence level=90% and margin error=4%). Annotators were asked to determine whether a participant was looking at the person in front/at the right/at the left or at no-one. A total of 23000 frames (an average 359.4 per video with standard deviation of 46.54) were retained based on the agreement between the annotators. The annotated VFOA data were randomly divided into training and validation sets (100 repetitions) to learn the SVM model (radial basis kernel function, RBF) with varying kernel parameter. The cost function, the random under sampling, and the SMOTE methods were combined with SVM. For each video, the method performing the highest geometric mean of the detection rates was selected to classify the whole unlabelled head pose; this procedure was applied independently to each video-segment (Beyan et al., 2016). To reduce noise, a smoothing filter with a 5 frames window was applied to the VFOA result obtained from SVM predictions.

Leader classification analysis. Gaze features of interest were extracted and used as predictors for the leader classification analysis. A linear Support Vector Machine (SVM) with a leave-one-subject-out cross-validation scheme was utilized to solve the classification task (i.e., discriminating between leaders vs. followers based on visual features). Linear SVM was chosen to avoid overfitting while also ensuring an optimal function to separate the data classes (Ben-hur and Weston, 2010; Gokcen and Peng, 2002; Hsu et al., 2010). As the features were already indexed on a common scale, data were not rescaled prior to the analyses.

The classification performance was assessed based on the following criteria: classification accuracy (defined as the percentage of the number of individuals classified correctly over the total individuals), Kappa (i.e., proportion of correctly classified individuals after accounting for the probability of chance agreement), Sensitivity (i.e., true positive rate), Specificity (i.e., true negative rate), and F1 (i.e., the harmonic mean of Sensitivity and Specificity). Finally, we tested the statistical reliability of the classification results with a 1000-repetition permutation test, a non-parametric test which randomly rearranged 1000 times the labels of observed data points to calculate the distribution of the test statistic under the null hypothesis (Ojala and Garriga, 2010).

To estimate the relative contribution of gaze features for leadership prediction, we used a simplified Fisher criterion (F-score criterion). F-score provides a measure of how well a single feature at a time can discriminate between different classes. The higher the F-score, the better the discriminatory power of that feature. The F-score was computed in the same way as the classic Fisher criterion (Duda et al., 2012) (see Table 2 in the main text for the corresponding results).

All classification analyses were performed with the PredPsych package written in R (Koul et al., 2018).

Supplemental Analyses

Composition of groups

Figure S1 shows the distributions of the scores that the subjects of the participant pool obtained at each of the SYMLOG subscales (Polley et al., 1988). To select the potential leaders, we first selected participants with a score on the SYMLOG Dominance subscale higher than the median of the sample (females: Me = 4; males: Me = 6). Among these, we then selected the participants with a score on the SYMLOG Task Orientation subscale higher than the median of the sample (females: Me = 4; males: Me = 1). Finally, participants with a score on the Friendliness subscale higher than the median of the sample (Xi = Me + 1) were selected as democratic leaders (females: >18; males: >18), while participants with a score on the Friendliness subscale lower than the median of the sample (Xi = Me - 1) were selected as autocratic leaders (females: <15; males: <14). All other participants could be randomly selected as followers.

Based on participants' availability, we then composed the experimental groups. A series of independent t-tests on the SYMLOG scores of the participants of the experimental groups confirmed that autocratic (M = 13.00, SE = .85) and democratic (M = 21.63, SE = 1.03) leaders significantly differed on Friendliness, $t(14) = 6.458$, $p < .001$ (Dominance: $p = .858$, Task orientation: $p = .353$). Consistently, leaders overall significantly differed from the followers on Dominance ($p = .002$), Task orientation ($p < .001$), and Friendliness ($p = .012$). However, autocratic leaders showed a non-significant difference with the followers (M = 12.12, SE = 1.73) on Friendliness scores ($p = .489$). Taken together, these data show that the differentiation between democratic and autocratic leaders was consistent with the experimental design, although the specificity of autocratic leaders compared to the other participants was weaker than that of democratic leaders.

Manipulation checks

We checked the reliability of our procedure both for establishing leadership as well as for creating different leadership settings.

Leadership perception. As a manipulation check, we administered participants the General Leadership Impression (GLI) scale. GLI is a 5-item scale that asks participants to rate the other members of the group on their contribution to the group's overall effectiveness on the activity (Lord et al., 1984). The range of responses is 1 (nothing) to 5 (extreme amount). Individual GLI scores are calculated by averaging the ratings given by the other three group members. The higher the score on this scale, the higher the leadership perception. All participants, including the designated leaders, filled out the GLI. To obtain an additional independent measure of leadership perception, we also asked two independent observers to watch the videos of the group interactions and complete the GLI for each group member (ICC = .771, $p < .001$).

To test the efficacy of our procedures for establishing leadership, the GLI scores obtained by the designated leaders were compared to the average GLI scores obtained by the other group members with a split-plot ANOVA with role (2: leader, follower) as within-subject factor, and leadership style (2: autocratic, democratic) and situational condition (2: low time-pressure, high time-pressure) as between-subject factors. Designated leaders (M = 3.93, SE = .19) were perceived as showing higher leadership attitudes relative to other group members (M = 3.25, SE = .06) [$F(1,12) = 9.412$, $p = .010$, $\eta^2_p = .440$], with no apparent influence of leadership style ($p = .414$) or situational condition ($p = .774$). Similar results were obtained when considering the leadership perception scores obtained from independent observers [$F(1,12) = 14.944$, $p = .002$, $\eta^2_p = .555$; other p values ranging from .266 to .766]. Together, these data indicate that designated leaders were perceived as such, both by other group members and by external observers.

Leadership settings. To test the efficacy of our procedures in creating the different leadership settings (see Figure 1 A in the main text), we also assessed group performance, individual reports of satisfaction and stress across conditions.

To evaluate group task performances on the task, a univariate ANOVA was performed on the task scores, with leadership style (2: autocratic, democratic) and situational condition (2: low time-pressure, high time-pressure) as between-subject factors. A similar analysis was applied to the time of interactions. In line with previous evidence showing that situational fit improves group performance (Fiedler, 1971; Higgins, 2008; Strube and Garcia, 1981), analysis of task scores revealed that groups in high fit conditions (M = 41.5, SE = 2.89) performed

better than those in low fit conditions (M = 51.5, SE = 2.56) [$F(2,14) = 5.970$, $p = .031$, $\eta^2_p = .332$; ($t(14) = 2.585$, $p = .022$)].
Satisfaction and stress related to the group interaction were assessed with a 15-items ad hoc questionnaire. Satisfaction was assessed with reference to how they felt satisfied with group performance (2 items) and with individual performance (1 item) and how they felt a in positive mood in relation to the interaction (8 items). Stress was assessed with reference to how they felt pressured by time perception (1 item) and how they perceived a sense of fatigue (3 items). The range of responses was 1 (nothing) to 5 (extreme amount). Since followership attitudes have been associated with perception of relationship quality and satisfaction (Sy, 2010; Uhl-Bien et al., 2014), the results from the Implicit Followership Theories (IFT) scale were included in the analysis of satisfaction and stress as covariates. The IFT is an 18-item scale asks to participants to rate from 1 (not at all characteristic) to 10 (extremely characteristic) the extent to which they associate each adjective item to their idea of followers. The items describing positive attributes (for example, "hardworking", "enthusiastic") comprise the sub-scale follower Prototype, whereas items describing negative attributes (for example, "bad temper", "easy influenced") comprise the sub-scale follower Anti-prototype (Sy, 2010).
Two separate MANCOVAs were performed on satisfaction reports (group performance, individual performance, positive mood) and on stress reports (time perception and fatigue), with leadership style (2: autocratic, democratic) and situational condition (2: low time-pressure, high time-pressure) as between-subject factors. Scores on followership attitudes (sub-scales follower Prototype and follower Anti-prototype) were included as covariates. All significant main effects were explored with separate univariate ANOVAs.
Analyses revealed that participants' satisfaction reports were modulated by leadership style [$F(3,56) = 3.162$, $p = .032$, $\eta^2_p = .145$], by the fit between leadership style and situational condition [$F(3,56) = 3.188$, $p = .031$, $\eta^2_p = .146$], and by individual generally positive attributions to followers [$F(3,56) = 5.140$, $p = .003$, $\eta^2_p = .216$]. Specifically, participants in groups with democratic leaders (M = 3.77, SE = .14) reported higher satisfaction in relation to group performance, as compared to participants in groups with autocratic leaders (M = 3.26, SE = .14) [$F(1,59) = 6.467$, $p = .014$, $\eta^2_p = .099$]. These effects were modulated by the fit with the situational condition [$F(1,59) = 5.795$, $p = .019$, $\eta^2_p = .089$], such that the satisfaction for group performance was higher with democratic leaders (M = 4.11, SE = .20) than with autocratic leaders (M = 3.08, SE = .21) only in the low time-pressure condition ($p = .001$), and that in groups with democratic leaders, the satisfaction was higher under low time-pressure, as compared to high time-pressure (M = 3.44, SE = .20) ($p = .026$). Additionally, scores on positive mood were also modulated by the fit between leadership style and situational condition [$F(1,59) = 4.140$, $p = .046$, $\eta^2_p = .066$], in that participants in groups with democratic leaders reported higher positive mood after the interaction (M = 6.617, SE = .24) than participants in groups with autocratic leaders (M = 5.698, SE = .25) only under low time-pressure ($p = .012$). Both the reports of satisfaction for group performance and of positive mood were modulated by individual generally positive attributions to followers, such that the higher positive attributions were associated to higher satisfaction for group performance [$F(1,59) = 8.018$, $p = .006$, $\eta^2_p = .120$; B = .272] and to higher positive mood [$F(1,59) = 14.634$, $p < 0001$, $\eta^2_p = .199$; B = .444].
Analyses on *stress* reports revealed that participant in groups with autocratic leaders (M = 3.406, SE = .24) perceived time as not sufficient to a greater extent than participants in groups with democratic leaders (M = 4.125, SE = .23) [$F(2,57) = 3.193$, $p = .048$, $\eta^2_p = .101$)].
Taken together, these data suggest that the participants' stress, as indexed by time perception, and participants' satisfaction, as indexed by the satisfaction for the group performance and the positive mood associated with the interaction, were modulated by the leadership style and, partially, by the fit between the leadership style and situational condition. Specifically, democratic leaders were associated with higher satisfaction, especially in the low time-pressure condition. These results are in line with previous evidence showing that a general preference for democratic leadership is affected by its potential inadequacy to respond to task demands in situations that require straightforward solution strategies (Foels et al., 2000; Gastil, 1994; Kruglanski et al., 2006). All in all, manipulation check analyses confirmed the reliability of our procedure both for establishing leadership as well as for creating different leadership settings.

Control analysis on speaking time

To exclude that our results were simply driven by the proportion of speaking time, we applied MVCC to verify whether gaze behaviour identifies leaders regardless of speaking time. To do so, we trained a linear SVM to discriminate leaders based on gaze patterns recorded during the video-segments in which the leader spoke the most (39 video-segments, corresponding to 156 individual observations), and then tested it on the video-segments in which a non-leader member spoke the most (19 video-segments, corresponding to 76 individual observations). Seventeen video-segments had to be excluded from this analysis due to technical problems in audio recording. With an accuracy of 78%, cross-classification performance was well above the .50 chance level (95% CI = .67, .86; Kappa = 0.26; Sensitivity = 0.63; Specificity = 0.79; F1 = .37; $p < .001$). Training on the video-segments in which a non-leader member spoke the most and testing on the video-segments in which the leader spoke the most led to a cross-classification accuracy of 85% (95% CI = .78, .90; Kappa = .53; Sensitivity = .83; Specificity = .85; F1 = .61; $p = .002$). These results indicate that leadership-related gaze patterns are similar regardless of speaking time.

Data and software availability

Gaze behaviour data are available as Supplemental Information (provided as supplemental Excel table). The original videos of the group interactions are available from the repository of the Istituto Italiano di Tecnologia, https://www.iit.it/research/lines/pattern-analysis-and-computer-vision/pavis-datasets/574-leadership-corpus-dataset. The code to generate the VFOA is available from the Authors VM and CBeyan upon reasonable request. The code for classification analyses is available in Koul et al., 2018.

Supplemental References

Avolio, B.J., Walumbwa, F.O., Weber, T.J., 2009. Leadership: Current theories, research, and future directions. Annu. Rev. Psychol. 60, 421–449. doi:10.1146/annurev.psych.60.110707.163621

Ben-hur, A., Weston, J., 2010. A user's guide to Support Vector Machine, in: Carugo, O., Eisenhaber, F. (Eds.), Data Mining Techniques for the Life Sciences. Methods in Molecular Biology 609. Humana Press, pp. 223–239. doi:10.1007/978-1-60327-241-4

Blumberg, H.H., Hare, A.P., 1999. Sociometry applied to organizational analysis: A review. Int. J. Action Methods 52, 15.

Chirumbolo, A., Livi, S., Mannetti, L., Pierro, A., Kruglanski, A.W., 2004. Effects of need for closure on creativity in small group interactions. Eur. J. Pers. 18, 265–278. doi:10.1002/per.518

Cristinacce, D., Cootes, T., 2006. Feature detection and tracking with constrained local models, in: Chantler, M., Trucco, M., Fisher, B. (Eds.), BMVC 2006 - Proceedings of the British Machine Vision Conference 2006. Edinburgh, pp. 929–938. doi:10.5244/C.20.95

De Grada, E., Kruglanski, A.W., Mannetti, L., Pierro, A., 1999. Motivated cognition and group Interaction: Need for closure affects the contents and processes of collective negotiations. J. Exp. Soc. Psychol. 35, 346–365. doi:10.1006/jesp.1999.1376

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification, 2nd ed. John Wiley & Sons.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39, 175–191. doi:10.3758/BF03193146

Fiedler, F.E., 1971. Validation and extension of the contingency model of leadership effectiveness: A review of empirical findings. Psychol. Bull. 76, 128–148. doi:10.1037/h0031454

Foels, R., Driskell, J.E., Mullen, B., Salas, E., 2000. The effects of democratic leadership on group member satisfaction: An integration. Small Gr. Res. 31, 676–701. doi:10.1177/104649640003100603

Gastil, J., 1994. A definition and illustration of democratic leadership. Hum. Relations 47, 953–975. doi:10.1177/001872679404700805

Gokcen, I., Peng, J., 2002. Comparing linear discriminant analysis and support vector machines, in: Yakhno, T. (Ed.), Advances in Information Systems. Springer-Verlag Berlin Heidelberg, Izmir, Turkey, pp. 104–113. doi:10.1007/3-540-36077-8_10

Higgins, E.T., 2008. Regulatory fit, in: Shah, J.Y., Gardner, W.L. (Eds.), Handbook of Motivation Science. Guilford Press, pp. 356–372.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2010. A practical guide to support vector classification, Technical report, Department of Computer Science, National Taiwan University. doi:10.1177/02632760022050997

Johnson, D.W., Johnson, F.P., 1994. Joining Together: Group Theory and Group Skills, 5th ed. Allyn & Bacon.

Kruglanski, A.W., Freund, T., 1983. The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. J. Exp. Soc. Psychol. 19, 448–468. doi:10.1016/0022-1031(83)90022-7

Kruglanski, A.W., Pierro, A., Mannetti, L., De Grada, E., 2006. Groups as epistemic providers: Need for closure and the unfolding of group-centrism. Psychol. Rev. 113, 84–100. doi:10.1037/0033-295X.113.1.84

Lord, R.G., Foti, R.J., De Vader, C.L., 1984. A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. Organ. Behav. Hum. Perform. 34, 343–378. doi:10.1016/0030-5073(84)90043-6

Ojala, M., Garriga, G.C., 2010. Permutation tests for studying classifier performance. J. Mach. Learn. Res. 11, 1833–1863. doi:10.1109/ICDM.2009.108

Polley, R.B., Hare, A.P., Stone, P.J., 1988. The SYMLOG practitioner: Applications of small group research. Praeger Press.

Strube, M.J., Garcia, J.E., 1981. A meta-analytic investigation of Fiedler's contingency model of leadership effectiveness. Psychol. Bull. 90, 307–321. doi:10.1037/0033-2909.90.2.307

Sy, T., 2010. What do you think of followers? Examining the content, structure, and consequences of implicit followership theories. Organ. Behav. Hum. Decis. Process. 113, 73–84. doi:10.1016/j.obhdp.2010.06.001

Uhl-Bien, M., Riggio, R.E., Lowe, K.B., Carsten, M.K., 2014. Followership theory: A review and research agenda. Leadersh. Q. 25, 83–104. doi:10.1016/j.leaqua.2013.11.007

Zaccaro, S.J., Foti, R.J., Kenny, D.A., 1991. Self-monitoring and trait-based variance in leadership: An investigation of leader flexibility across multiple group situations. J. Appl. Psychol. 76, 308–315. doi:10.1037/0021-9010.76.2.308

Supplemental Figures

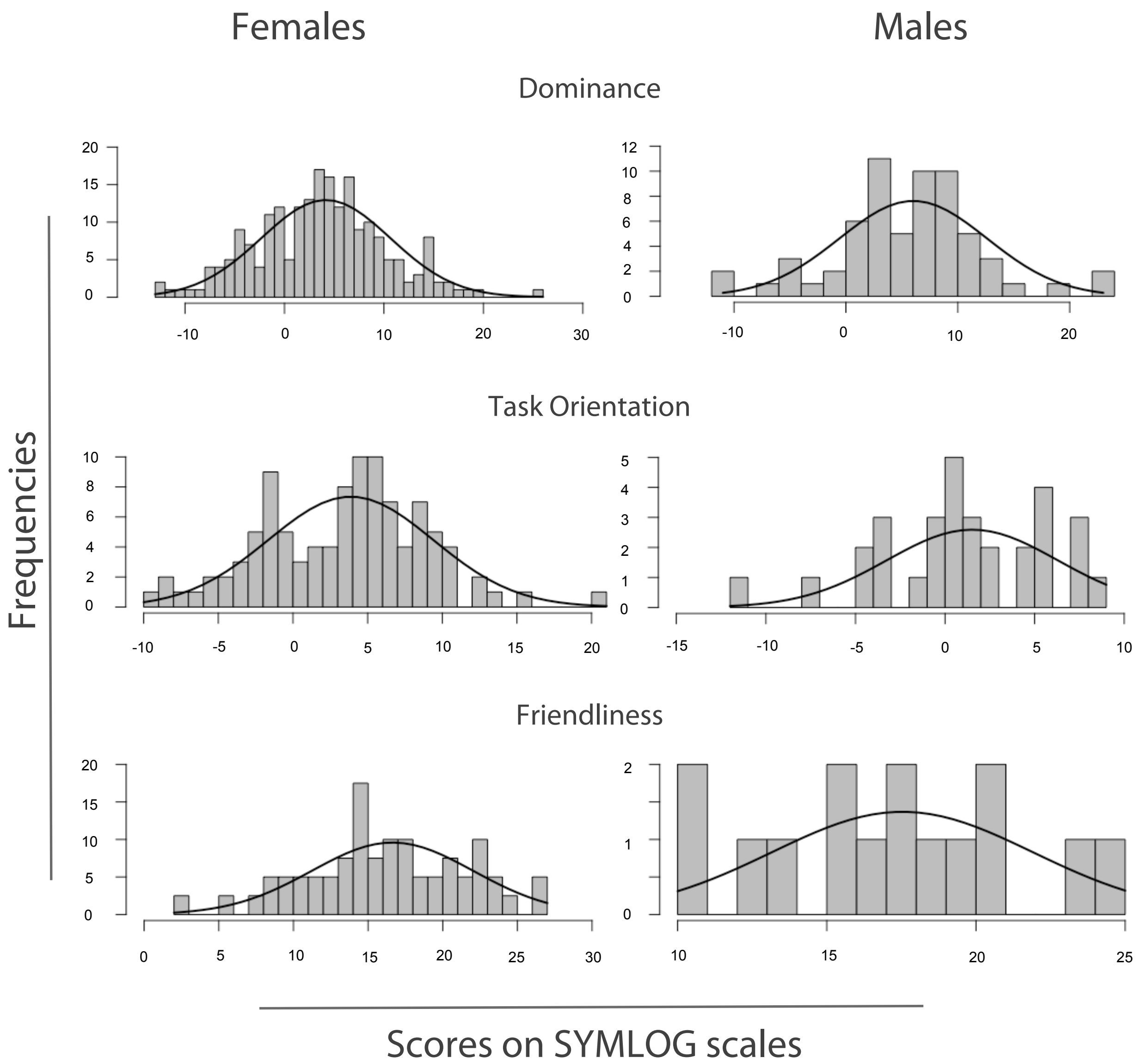

Figure S1: Distributions of the scores that the subjects of the participant pool obtained at each of the SYMLOG subscales (Dominance, Task Orientation, Friendliness), separately for female and male participants.