**Supplementary Material**

**Social anxiety and concordance in emotional responses across levels of evaluative threat**



[1] Note that a secondary visualization of the community structure of the non-social network is not provided as this network exhibited concordance.

## Deviations from Preregistered Analytic Plan

- We elected to focus on a subset of the planned analyses rather than all of the planned analyses to make the present study more focused and comprehensible (this followed feedback we received on an earlier draft that there was far too much for one paper). A secondary manuscript will present the analyses that were not the focus of this paper. Note that, because we ran all of the planned analyses, some of the results presented in this study reflect the results of the full analyses. For example, the threshold level that was identified as being optimal across networks was optimal across all 10 of the original planned networks. Additionally, the BH-corrected *p*-values are corrected for 10 absolute concordance tests and 4 relative concordance tests rather than the 5 absolute and 2 relative concordance tests presented here.
- The preregistered analytic plan mistakenly noted that a network would be considered to exhibit absolute concordance if its optimal modularity value is *lower* than that of a sample of random networks with no community structure by definition. This should state that a network will be considered concordant if its optimal modularity value *could have likely arisen from the null distribution* of a sample of random networks with no community structure by definition (as is noted elsewhere in the preregistration). The latter criterion is the one applied in the analyses presented here.
- The Benjamini-Hochberg correction was applied separately to the absolute concordance tests and the relative concordance tests, rather than to all of the tests together.
- In addition to *qgraph,* the *MDSnet* and *PROCRUSTESnet* functions from the *networktools* package (Jones, 2022) were used to visualize the networks.
- The behavioral ratings were reverse-coded so that the direction of the Likert scale matched that of the affective and cognitive variables (i.e., very good to very poor as opposed to very poor to very good).
- Regarding the behavioral coding, we initially planned to follow the guidelines outlined by Fydrich et al. (1998), who suggested that training proceed until, across raters, none of the ratings based on any of the five behavioral anchors have a difference of more than one rating point. Though the coding team met this criterion for the dyadic experiences, the criterion was adjusted slightly for the group experiences. Specifically, the full team coded group experiences until there was a difference of no more than two

rating points on a single behavioral anchor during one (out of three total) two-minute section of the conversation.

- The preregistration noted that we would consider using data from the Empatica E4 worn on participants' right wrists in the event that a participant had over 40% of data points within a phase missing or removed (due to outliers/artifacts) from the left wristband data *and* that participant had less than 40% of data points missing or removed from the corresponding phase as collected via the right wristband. This did not end up being relevant, as any participant missing substantial amounts of data from their left wristband also had similar patterns of missingness in their right wristband data. Accordingly, to maintain consistency and reduce complexity, we only analyzed data from the E4 wristbands worn on participants' left wrists.
- The preregistration noted that we would use a two-step process to determine whether or not a network exhibited concordance. Specifically, we reported that a network would only be said to exhibit absolute concordance if (1) we could reject the null hypothesis from the permutation test and (2) if it appeared concordant based on a visual inspection. This is because we wanted to confirm that the network lacked community structure due to being densely connected rather than completely disconnected. However, this step ultimately ended up being redundant with the thresholding procedure we used, and thus irrelevant. We selected the threshold level by choosing the highest threshold that still produced a fully connected graph (i.e., no nodes are completely disconnected from the others) across all networks. Thus, by default, if one of the networks exhibited statistical concordance based on the permutation test, it would be impossible for it to be concordant due to it being a completely disconnected graph because the thresholding step ensured this would not be the case. Accordingly, we still visualized each network and inspected it to better understand network-specific patterns of associations, but did not use the visual inspection step to determine whether networks exhibited concordance.

## Rationale for Physiological Outlier Detection and Removal

**Tonic SCL.** Typical tonic SCL values range from 1-40 μSiemens, with most usually falling between 2-20μS (Braithwaite et al., 2013; Dawson et al., 2017). However, these ranges reflect SCL values collected via wet electrodes (i.e., electrodes prepared with a gel that enhances the electrical conductivity of the skin) secured to the palm or fingertips, whereas most ambulatory psychophysiological recording devices (such as the Empatica E4) use dry electrodes on the wrist and yield much lower SCL values (i.e., < 1μS at rest; Kleckner et al., 2021). Based on previous research indicating that tonic SCL data as low as 0.1μS collected from the E4 can be potentially meaningful (i.e., likely reflective of a combination of both signal and noise; Kleckner et al., 2021), we set 0.1μS as a lower bound outlier and removed any data points falling below this value. Further, any participants who had over 40% of their tonic SCL data points removed within a given phase of an experience had that task block set to missing because their average tonic SCL for that experience could not be reliably estimated. Taking this approach, 11.94% of 136,670 data points were removed, yielding 120,350 data points that were included in analyses. See Supplementary Material F for a sensitivity analysis examining how using a more stringent outlier detection threshold (i.e., < 1μS) impacted results.

**Skin Temperature.** Human skin temperature, as measured via the wrist, has a normal range that varies from approximately 24℃ to 36℃ as a function of ambient temperature (Leonov et al., 2007). Accordingly, we removed any individual data points falling below 24℃ or above 36℃ from the dataset prior to conducting analyses. Taking this approach, 1.3% of 414,050 data points were removed, all of which were temperature readings over 36℃ that came from one participant.

**Heart Rate.** In adults, heart rate can vary substantially depending on cardiovascular fitness, activity (e.g., exercise vs. rest), and stress level (Centers for Disease Control and Prevention, 2022; Ostchega et al., 2011; Prior & La Gerche, 2012). Accordingly, we defined the feasible HR range for this sample as 30-200bpm and removed any data points that fell outside of this range. Additionally, as an additional artifact detection step, we inspected each participant's range in HR for large, sudden increase or decreases (i.e., over 60bpm) in HR that occurred over a period of less than 2 minutes. Taking this approach, one participant's HR data was excluded from all experiences (6.16% of 110,851 total data points removed).

## References


Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*(1), 1017–1034.

Centers for Disease Control and Prevention. (2022, June 3). *Target Heart Rate and Estimated Maximum Heart Rate*. https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm

Dawson, M. E., Schell, A. M., & Filion, D. L. (2017). The electrodermal system. In *Handbook of psychophysiology, 4th ed* (pp. 217–243). Cambridge University Press.

Fydrich, T., Chambless, D. L., Perry, K. J., Buergener, F., & Beazley, M. B. (1998). Behavioral assessment of social performance: A rating system for social phobia. *Behaviour Research and Therapy*, *36*(10), 995–1010. https://doi.org/10.1016/S0005-7967(98)00069-2

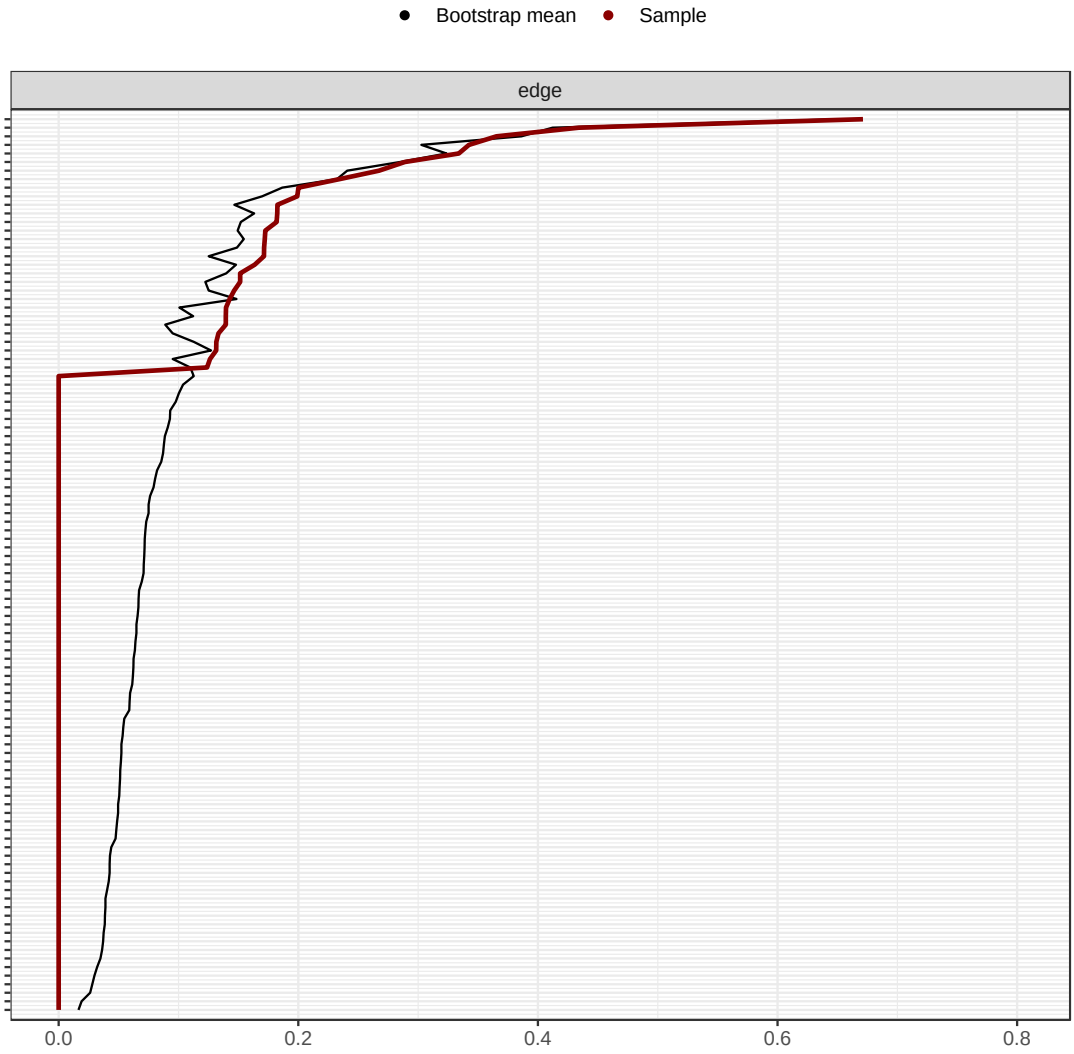
Jones, P. (2022). *networktools: Tools for Identifying Important Nodes in Networks* (1.5.0). https://CRAN.R-project.org/package=networktools

Kleckner, I. R., Feldman, M. J., Goodwin, M. S., & Quigley, K. S. (2021). Framework for selecting and benchmarking mobile devices in psychophysiological research. *Behavior Research Methods*, *53*(2), 518–535. https://doi.org/10.3758/s13428-020-01438-9

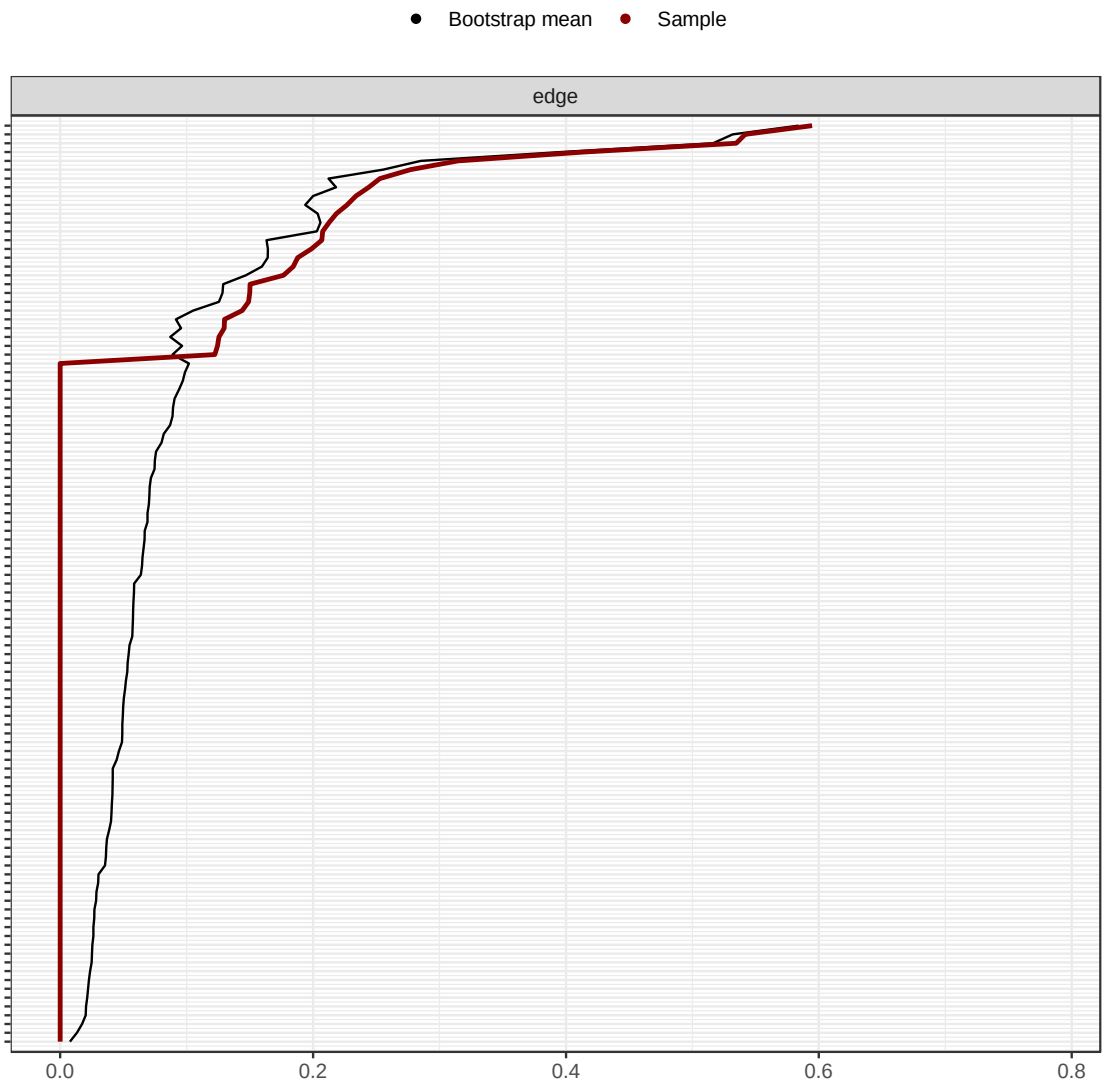
Leonov, V., Vullers, R. J. M., & Centre, H. (2007). Thermoelectric Generators on Living Beings. *Proceedings of the 5th European Conference on Thermoelectrics*, 47–52.

Ostchega, Y., Porter, K. S., Hughes, J., Dillon, C. F., & Nwankwo, T. (2011). *Resting Pulse Rate Reference Data for Children, Adolescents, and Adults: United States, 1999-2008* (No. 41; National Health Statistics Reports, pp. 1–17). National Center for Health Statistics.

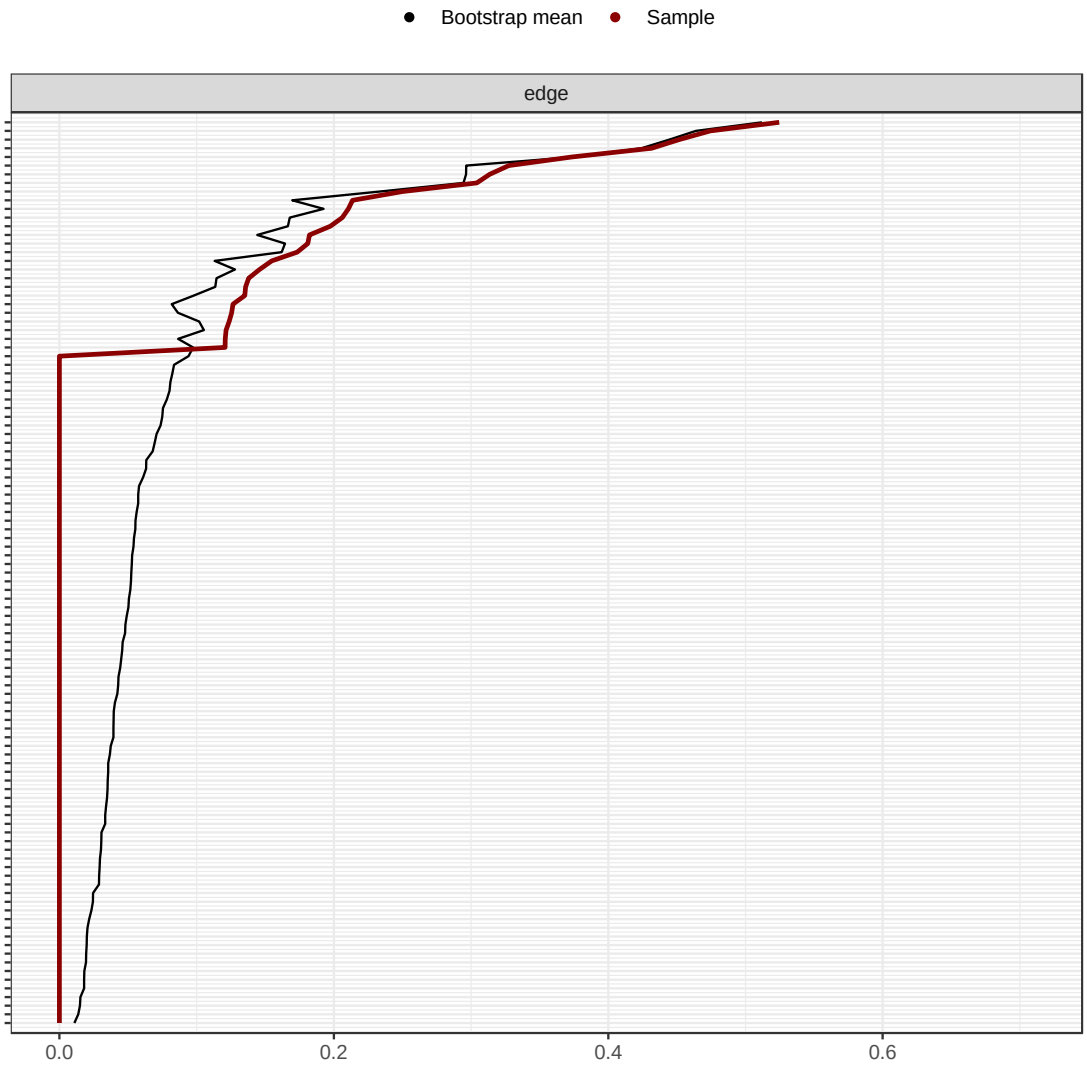
Prior, D., & La Gerche, A. (2012). The athlete's heart. *Heart (British Cardiac Society)*, *98*, 947–955. https://doi.org/10.1136/heartjnl-2011-301329



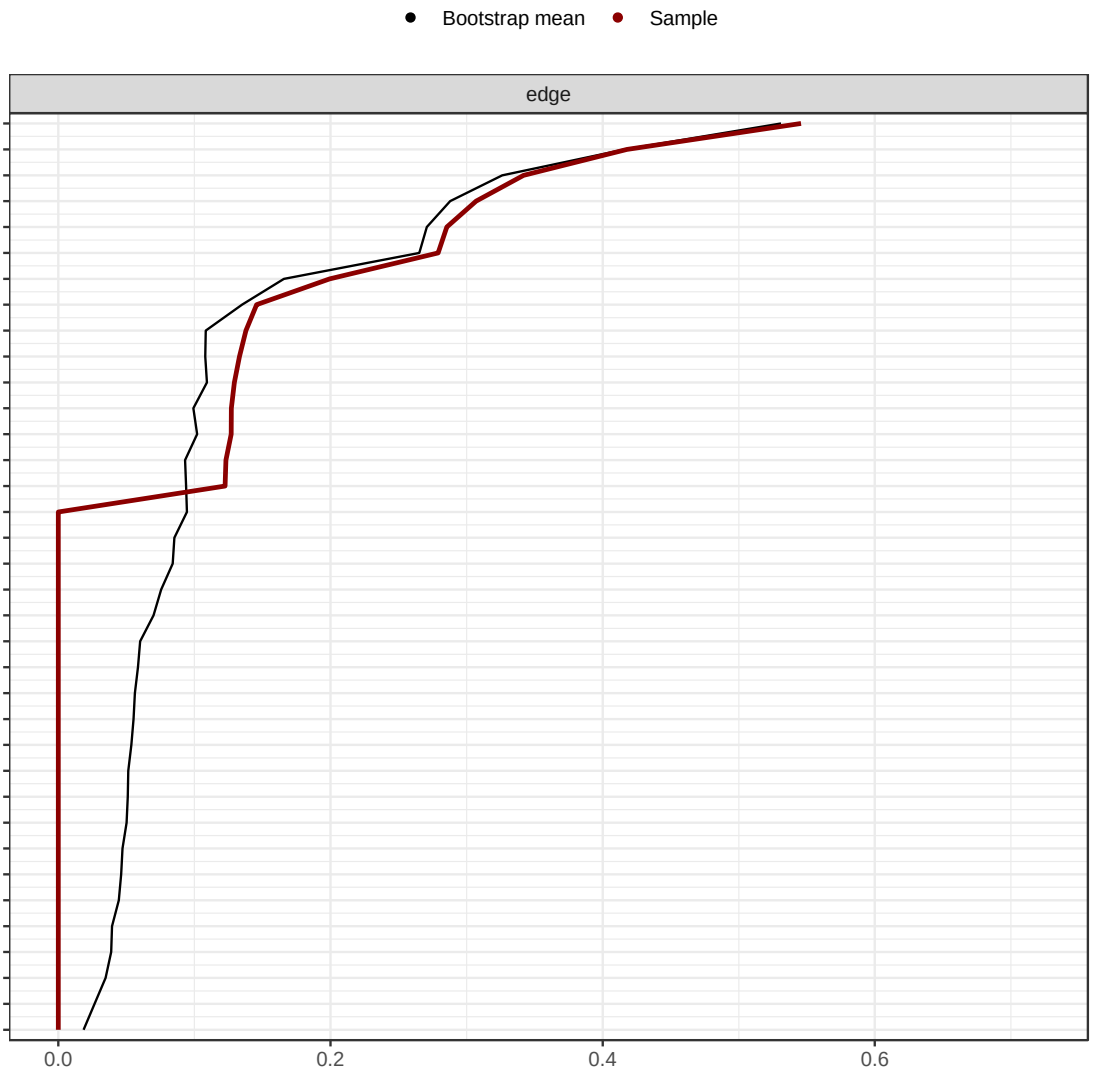

**Supplementary Figure 1.** Nonparametric bootstrapping results with 1,000 samples for the explicit evaluation network. Each notch on the y-axis corresponds to each of the possible edges (i.e., associations between emotion response indicators) in the network, and the x-axis corresponds to the weight of each edge (i.e., strength of the association). For the explicit evaluation network, there are 105 possible undirected edges between the 15 emotion response indicators included in the network. Note that the majority of edge weights from the sample are zero (i.e., no edge exists between those two indicators) due to the thresholding procedure, which involved setting all edges < 0.12 to zero.



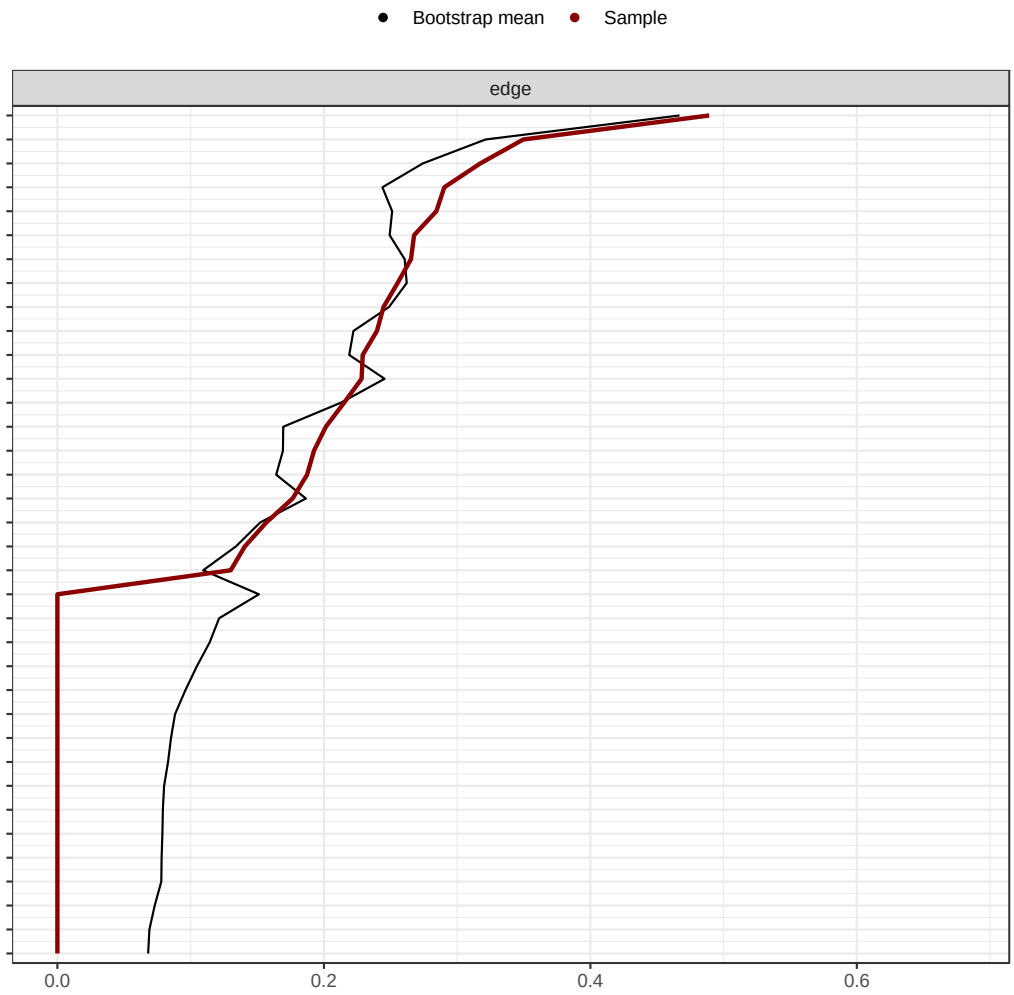

**Supplementary Figure 2.** Nonparametric bootstrapping results with 1,000 samples for the non-explicit evaluation network. Each notch on the y-axis corresponds to each of the possible edges (i.e., associations between emotion response indicators) in the network, and the x-axis corresponds to the weight of each edge (i.e., strength of the association). For the non-explicit evaluation network, there are 105 possible undirected edges between the 15 emotion response indicators included in the network. Note that the majority of edge weights from the sample are zero (i.e., no edge exists between those two indicators) due to the thresholding procedure, which involved setting all edges < 0.12 to zero.



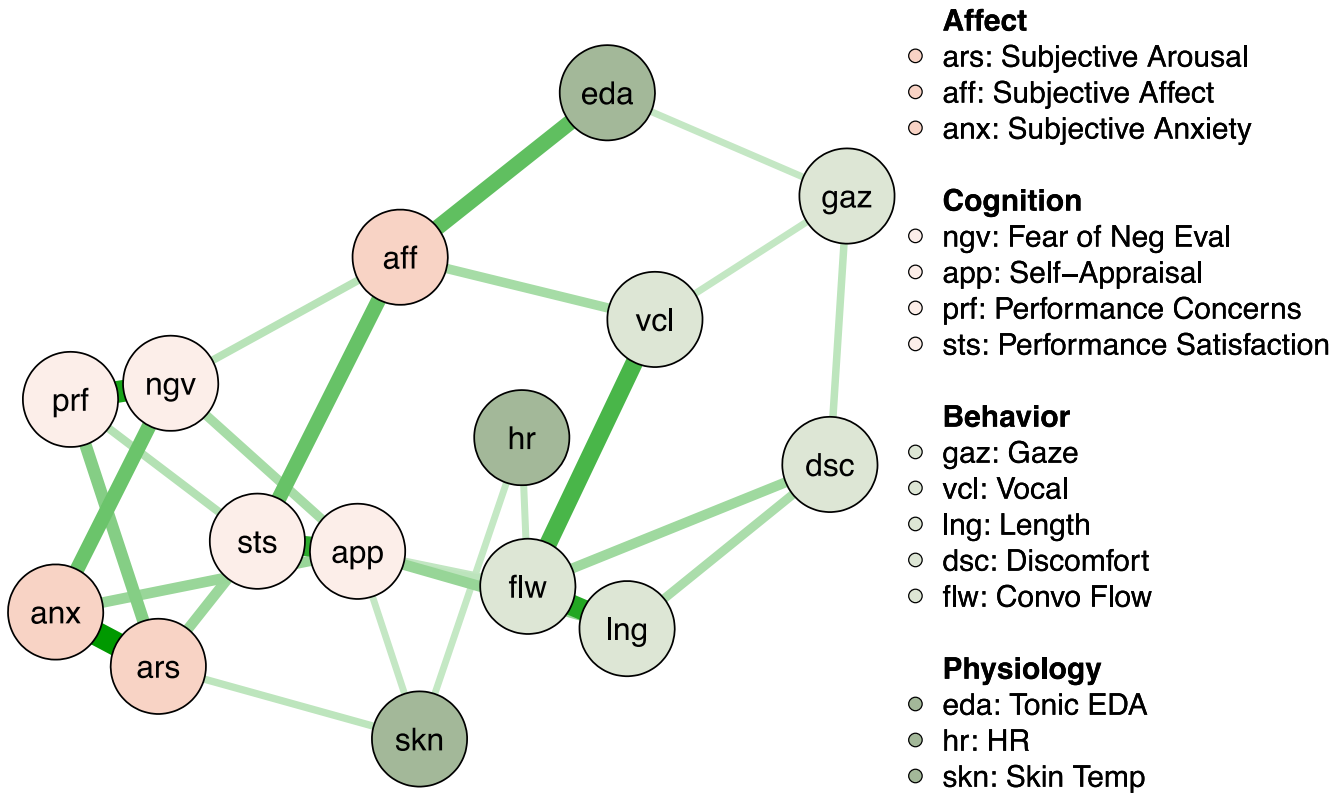

**Supplementary Figure 3.** Nonparametric bootstrapping results with 1,000 samples for the full social experience network. Each notch on the y-axis corresponds to each of the possible edges (i.e., associations between emotion response indicators) in the network, and the x-axis corresponds to the weight of each edge (i.e., strength of the association). For the full social experience network, there are 105 possible undirected edges between the 15 emotion response indicators included in the network. Note that the majority of edge weights from the sample are zero (i.e., no edge exists between those two indicators) due to the thresholding procedure, which involved setting all edges < 0.12 to zero.



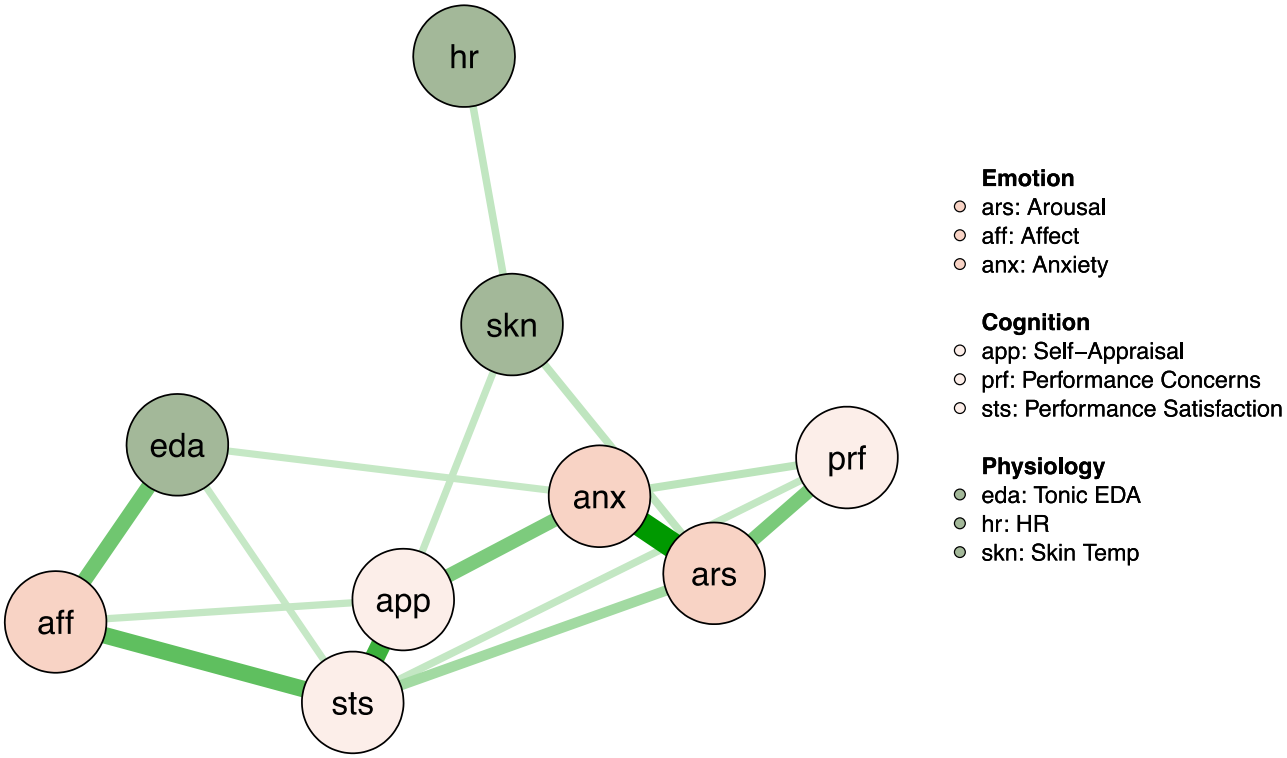

**Supplementary Figure 4.** Nonparametric bootstrapping results with 1,000 samples for the reduced social experience network. Each notch on the y-axis corresponds to each of the possible edges (i.e., associations between emotion response indicators) in the network, and the x-axis corresponds to the weight of each edge (i.e., strength of the association). For the reduced social experience network, there are 36 possible undirected edges between the 9 emotion response indicators included in the network. Note that the majority of edge weights from the sample are zero (i.e., no edge exists between those two indicators) due to the thresholding procedure, which involved setting all edges < 0.12 to zero.



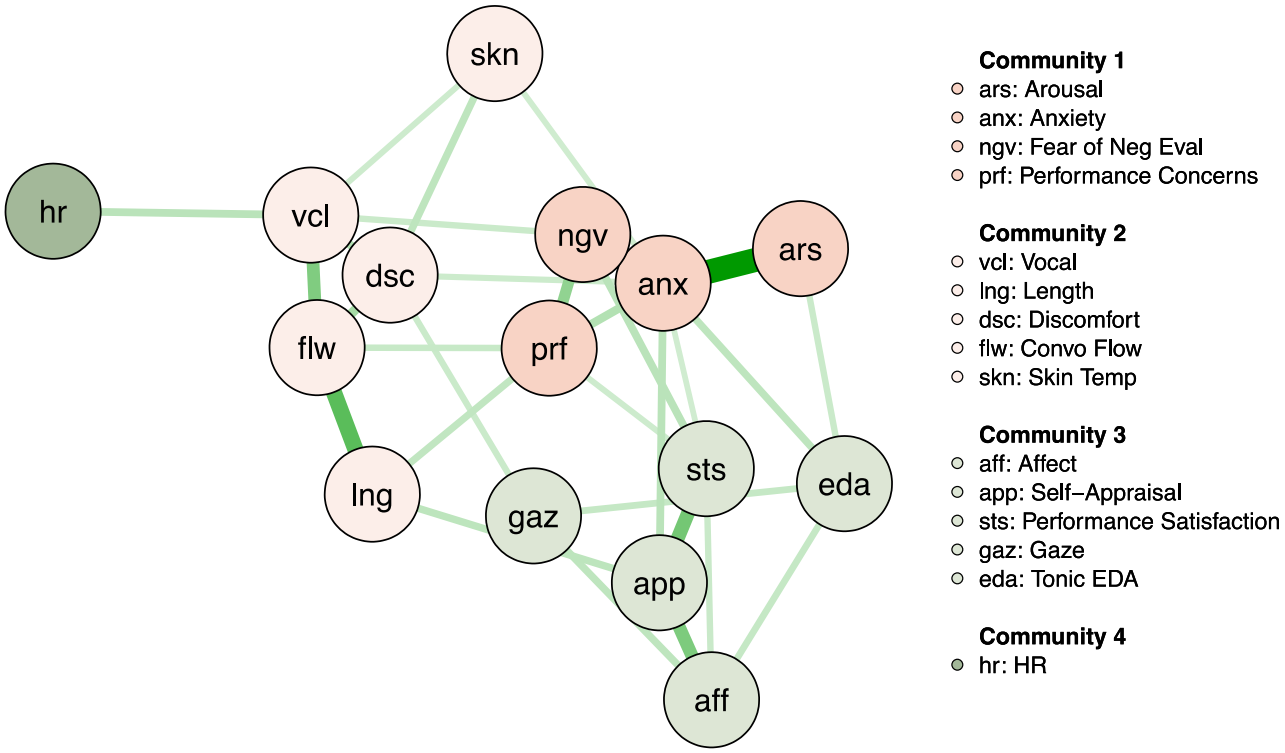

**Supplementary Figure 5.** Nonparametric bootstrapping results with 1,000 samples for the non-social network. Each notch on the y-axis corresponds to each of the possible edges (i.e., associations between emotion response indicators) in the network, and the x-axis corresponds to the weight of each edge (i.e., strength of the association). For the non-social experience network, there are 36 possible undirected edges between the 9 emotion response indicators included in the network. Note that the majority of edge weights from the sample are zero (i.e., no edge exists between those two indicators) due to the thresholding procedure, which involved setting all edges < 0.12 to zero.



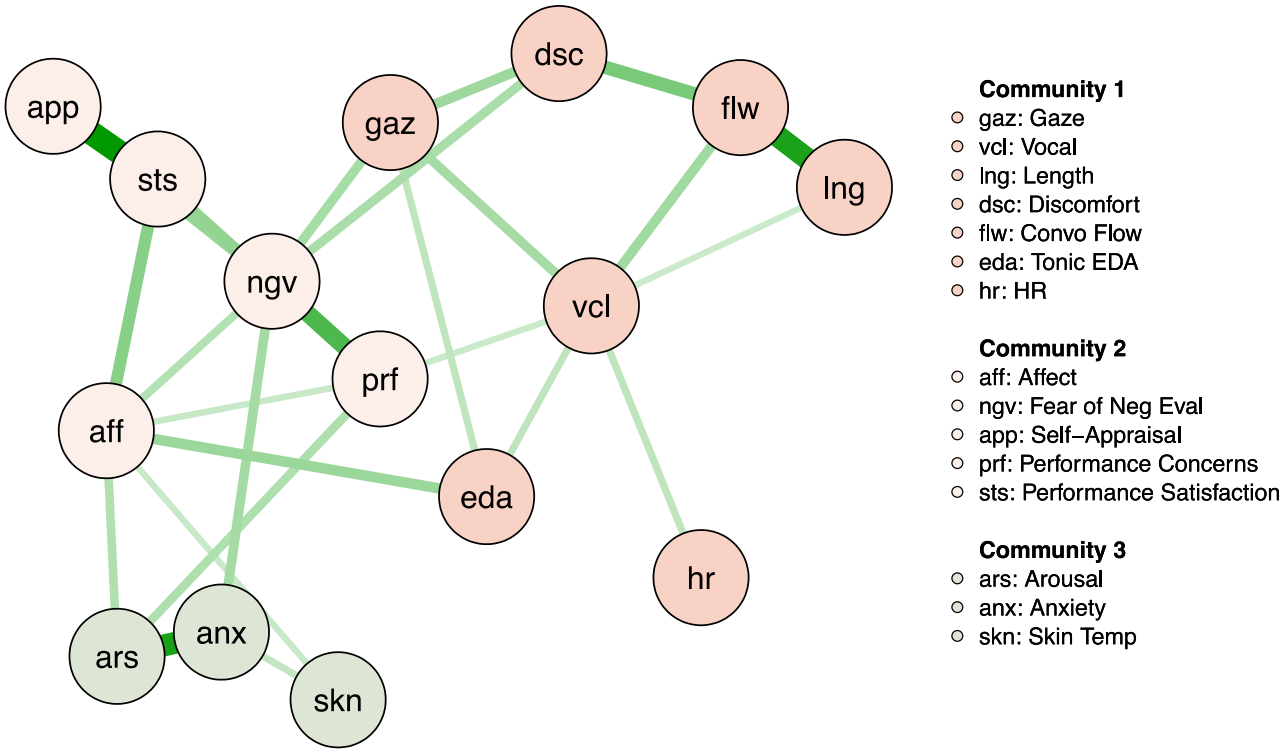

**Supplementary Figure 6.** Partial-$\tau$ network with absolute thresholding (0.12) for all four social experiences (group explicit evaluation, group non-explicit evaluation, dyad explicit evaluation, dyad non-explicit evaluation). Edges represent partial rank correlations and nodes represent emotional response indicators. Darker/thicker edges indicate stronger partial rank correlations. Node color corresponds to type of emotion response component (i.e., affective, cognitive, physiological, behavioral). The network is visualized using ordinal MDS with repulsion = 0.9 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights in the full social network ranged from 0.121 *(heart rate – skin temperature)* to 0.525 *(subjective arousal – subjective anxiety).* This network is discordant (modularity = 0.363), but there are relatively strong within-group associations (e.g., between arousal/anxiety and fear of negative evaluation/performance concerns) as compared to between-group associations, which may be indicative of concordance within but not across emotional response categories.



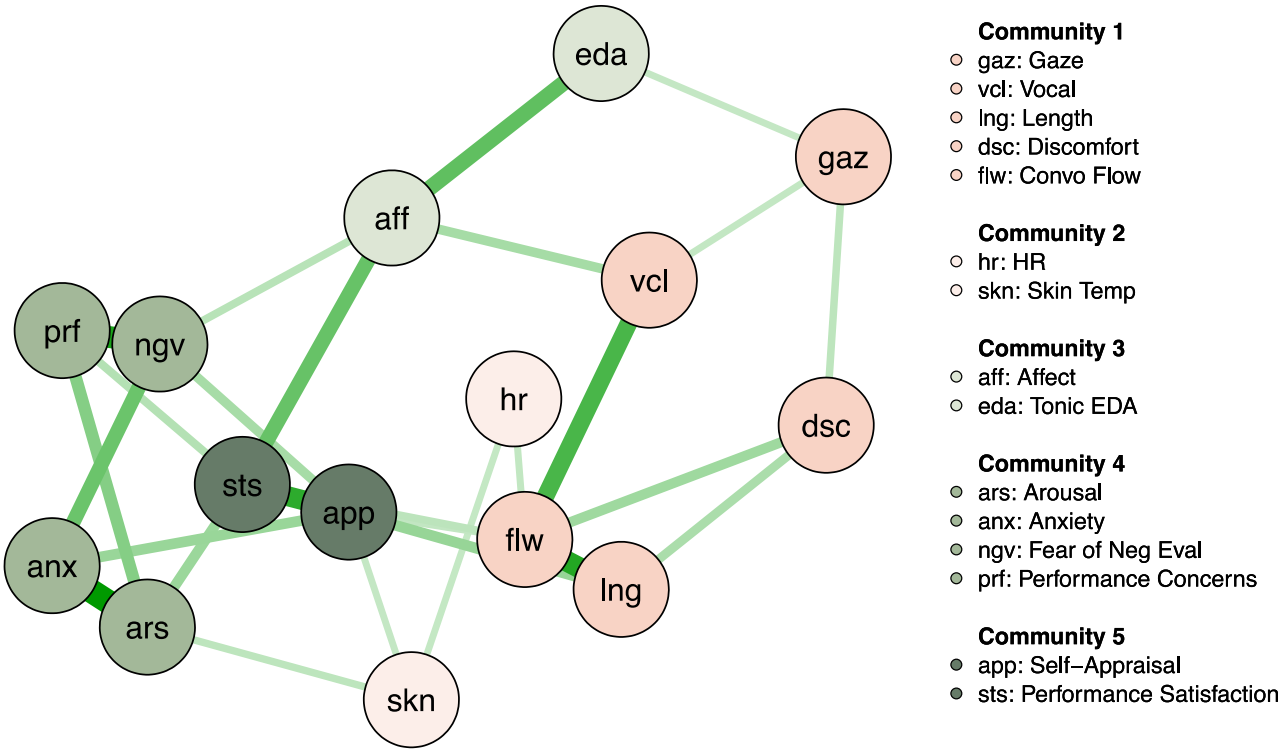

**Supplementary Figure 7.** Partial-$\tau$ network with absolute thresholding (0.12) for the reduced social experience network. Behavioral nodes and the fear of negative evaluation cognitive node are excluded from this network. Edges represent partial rank correlations and nodes represent emotional response indicators. Darker/thicker edges indicate stronger partial rank correlations. Node color corresponds to type of emotion response component (i.e., affective, cognitive, physiological, behavioral). The network is visualized using ordinal MDS with repulsion = 0.7 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights ranged from 0.123 *(performance satisfaction – tonic SCL)* to 0.546 *(subjective arousal – subjective anxiety).* Like the full social experience network, this network is discordant (modularity = 0.273), but the communities do not map as clearly onto the emotional response categories. However, the associations among the cognitive and emotional nodes, which were both assessed via self-report, are generally higher than those among physiological nodes or between physiological and non-physiological nodes.



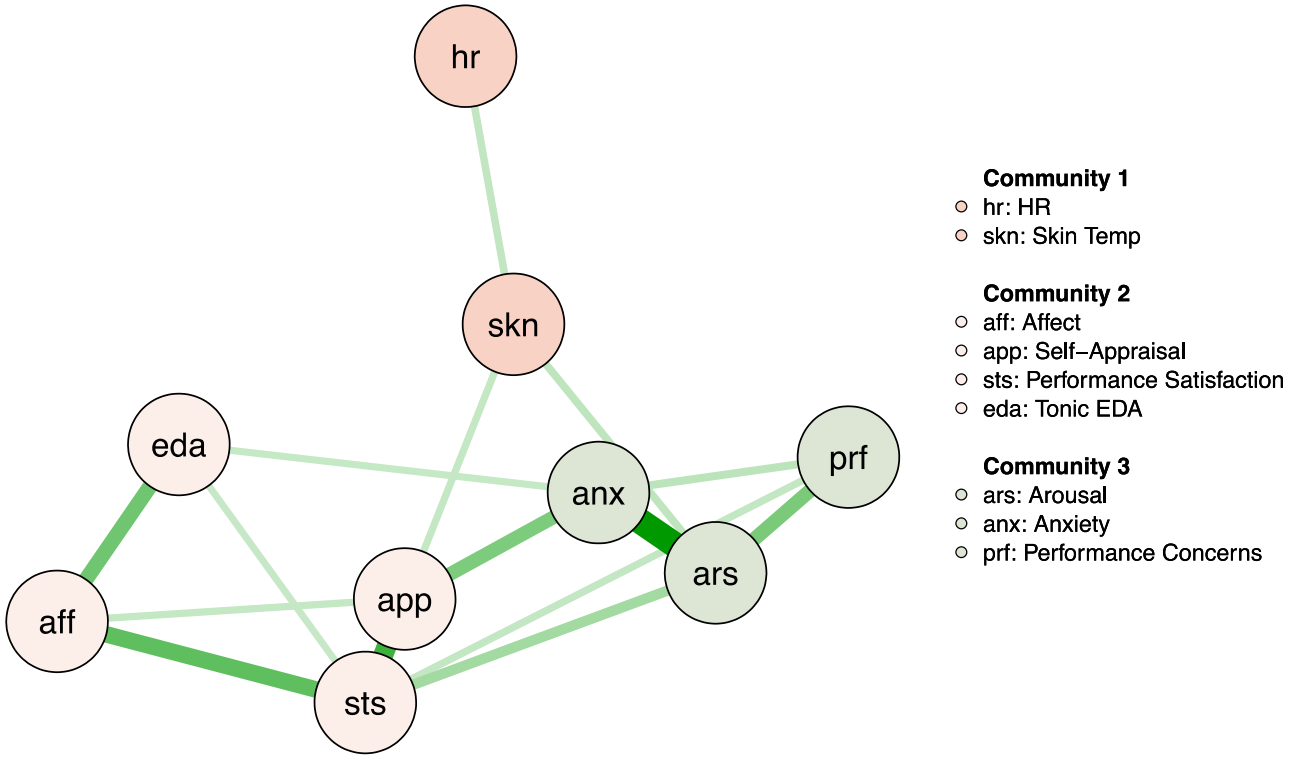

**Supplementary Figure 8.** Partial-$\tau$ network with absolute thresholding (0.12) for the explicit evaluation network. Edges represent partial rank correlations and nodes represent emotional response indicators. Thicker edges indicate stronger partial rank correlations. Node color node color corresponds to *walktrap*-identified community membership. The network is visualized using ordinal MDS with repulsion = 0.7 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights ranged from 0.124 *(anxiety – performance satisfaction)* to 0.671 *(anxiety – arousal).*




**Supplementary Figure 9.** Partial-$\tau$ network with absolute thresholding (0.12) for the non-explicit evaluation network. Edges represent partial rank correlations and nodes represent emotional response indicators. Thicker edges indicate stronger partial rank correlations. Node color node color corresponds to *walktrap*-identified community membership. The network is visualized using ordinal MDS with repulsion = 0.8 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights ranged from 0.122 *(performance concerns – performance satisfaction)* to 0.595 *(self-appraisal – performance satisfaction).*




**Supplementary Figure 10.** Partial-$\tau$ network with absolute thresholding (0.12) for all four social experiences (group explicit evaluation, group non-explicit evaluation, dyad explicit evaluation, dyad non-explicit evaluation). Edges represent partial rank correlations and nodes represent emotional response indicators. Thicker edges indicate stronger partial rank correlations. Node color node color corresponds to *walktrap*-identified community membership. The network is visualized using ordinal MDS with repulsion = 0.8 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights in the full social network ranged from 0.121 *(heart rate – skin temperature)* to 0.525 *(arousal – anxiety).*




**Supplementary Figure 11.** Partial-$\tau$ network with absolute thresholding (0.12) for the reduced social experience network. Behavioral nodes and the fear of negative evaluation cognitive node are excluded from this network. Edges represent partial rank correlations and nodes represent emotional response indicators. Thicker edges indicate stronger partial rank correlations. Node color node color corresponds to *walktrap*-identified community membership. The network is visualized using ordinal MDS with repulsion = 0.7 to minimize node overlap. Nodes that are closer together in space tend to be more strongly correlated than nodes that are father apart. Edge weights ranged from 0.123 *(performance satisfaction – tonic SCL)* to 0.546 *(arousal – anxiety).*

**Supplementary Table 1.** Results of listwise deletion analyses for Hypotheses 1a-b

| Network | Interpretation | Modularity | Original p-value | BH-Adjusted p-value* | Comparison with Main Analyses |
|---|---|---|---|---|---|
| Explicit Evaluation | Discordant | 0.344 | 0.001 | 0.005 | Consistent |
| Non-explicit Evaluation | Concordant | 0.241 | 0.044 | 0.088 | Inconsistent |
| Full Social | Concordant | 0.269 | 0.122 | 0.169 | Inconsistent |
| Reduced Social | Concordant | 0.286 | 0.021 | 0.053 | Inconsistent |
| Non-social | Concordant | 0.120 | 0.135 | 0.169 | Consistent |

Total $N = 26$ for the listwise deletion analyses. An absolute threshold level of 0.13 (compared to 0.12 for the main analyses) was applied to the listwise deletion networks given that this was the highest threshold across networks that resulted in fully connected graphs.
* = The BH-adjusted p-values reflect a correction for 10 tests rather than the five tests presented here. We chose to keep the correction at the same level because we already ran the planned full set of 10 absolute concordance tests (see preregistration) before ultimately deciding to reduce the scope of this paper and focus on only five of those tests.

**Supplementary Table 2.** Results of listwise deletion analyses for Hypotheses 2a-b

| Comparison | $D_{obs}$ (Absolute Value) | Original p-value | BH-Adjusted p-value* | Comparison with Main Analyses |
|---|---|---|---|---|
| Explicit Evaluation vs. Non-explicit Evaluation | 0.104 | 0.090 | 0.450 | Consistent |
| Social vs. Non-Social | 0.166 | 0.200 | 0.500 | Consistent |

* = The BH-adjusted p-values reflect a correction for four tests rather than the two tests presented here. We chose to keep the correction at the same level because we already ran the planned full set of four relative concordance tests (see preregistration) before ultimately deciding to reduce the scope of this paper and focus on only two of those comparisons.

**Supplementary Table 3.** Results of stringent EDA analyses for Hypotheses 1a-b

| Network | Total *N* | Interpretation | Modularity | Original p-value | BH-Adjusted p-value* | Comparison with Main Analyses |
|---|---|---|---|---|---|---|
| Explicit Evaluation | 10 | Discordant | 0.238 | < .001 | < .001 | Consistent |
| Non-explicit Evaluation | 9 | Concordant | 0.142 | 0.223 | 0.364 | Inconsistent |
| Full Social | 11 | Discordant | 0.249 | < .001 | < .001 | Consistent |
| Reduced Social | 11 | Concordant | 0.053 | 0.319 | 0.393 | Inconsistent |
| Non-social | 7 | Concordant | 0.109 | 0.255 | 0.364 | Consistent |

An absolute threshold level of 0.17 (compared to 0.12 for the main analyses) was applied to the stringent EDA networks given that this was the highest threshold across networks that resulted in fully connected graphs.
* = The BH-adjusted p-values reflect a correction for 10 tests rather than the five tests presented here. We chose to keep the correction at the same level because we already ran the planned full set of 10 absolute concordance tests (see preregistration) before ultimately deciding to reduce the scope of this paper and focus on only five of those tests.

**Supplementary Table 4.** Results of stringent EDA analyses for Hypotheses 2a-b

| Comparison | $D_{obs}$ (Absolute Value) | Original p-value | BH-Adjusted p-value* | Comparison with Main Analyses |
|---|---|---|---|---|
| Explicit Evaluation vs. Non-explicit evaluation | 0.096 | 0.048 | 0.137 | Consistent |
| Social vs. Non-Social | 0.056 | 0.510 | 0.510 | Consistent |

* = The BH-adjusted p-values reflect a correction for four tests rather than the two tests presented here. We chose to keep the correction at the same level because we already ran the planned full set of four relative concordance tests (see preregistration) before ultimately deciding to reduce the scope of this paper and focus on only two of those comparisons.