Author: Behrens, F.
Title: Physiological synchrony in the context of cooperation: Theoretical and methodological considerations
Issue Date: 2020-10-28

Appendices

Appendix A

Supplementary material for Chapter 2

Appendix A1

Table A.S1

*Descriptive statistics of the participants of the current study. This information was asked in an online questionnaire at the end of the experiment (*N = 73*)*

Descriptive	% (count)
Gender:	
Female	.69 (51)
Nationality:	
Dutch	.75 (55)
Non-Dutch	.25 (18)
Highest completed education:	
High school	.52 (38)
Applied University	.12 (9)
University	.36 (26)
Number of siblings:	
No siblings	.14 (10)
1 sibling	.44 (32)
2 siblings	.26 (19)
3 or more siblings	.16 (12)

Appendix A2

Descriptive statistics of the Liking and Knowing scale ratings

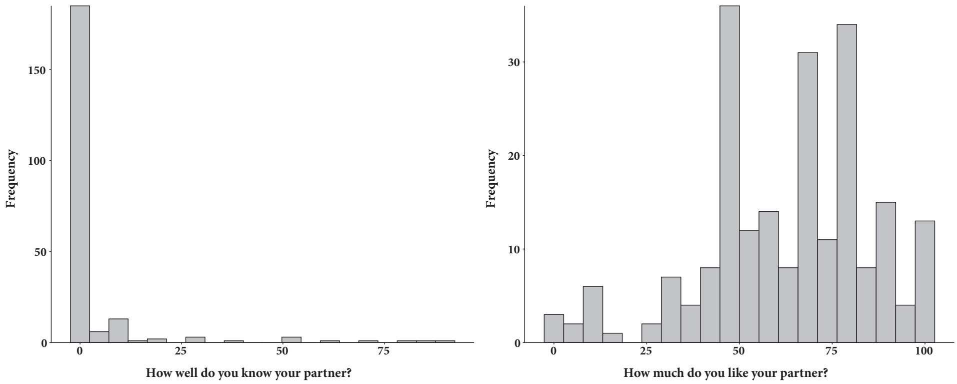

Figure A.S1. Distribution of ratings for the questions "How well do you know your partner?" (left) and "How much do you like your partner?" (right) on a scale from 0 to 100.

Table A.S2

Descriptive statistics of how well participants know (Knowing) and how much they liked their partner (Liking). Each participant rated three different partners

Descriptive	Knowing	Liking
Median	1	70
MAD	1	16
Maximum	90	100
Minimum	0	0
N	219	219
Missing	3	3

Note. MAD = median absolute deviation; *N* = sample size.

A

Table A.S3
Correlation between liking the partner and prosocial behavior

Game	Spearman's rho	95%-*CI*
Prisoner's Dilemma	.10	-.14–.33
Extended PD	.28	.03–.50
Rope Pull	.09	-.15–.32
Tangram	-.04	-.27–.20
Hidden Profile	.09	-.15–.33
Egg Hunt	.14	-.13–.40

Note. CI = confidence interval.

In Table A.S3, the Spearman's rho correlations between a person's rating of the partner and her/his prosocial behavior towards that person are shown. None of the outcomes of the games was reasonably correlated with how much participants liked each other.

Appendix A3

Alternative Prosocial Behavior Measures for the Tangram Game

In the following we provide descriptive statistics on possible alternative prosocial measures for the Tangram game on a dyadic level.

Do people perform better together than alone?

We here present three different measures to answer this question: (1) *max*, the number of completed puzzles from the participant with the highest number in the individual condition; (2) *sum*, the sum of the completed puzzles of both participants in the individual condition; (3) *mean*, the mean of the completed puzzles of both participants in the individual condition. For all measures, we calculated the difference score by subtracting the number of completed puzzles in the cooperative condition minus the corresponding measure (max, sum, or mean). A positive value indicates that people performed better together than alone. The descriptive statistics of the difference scores are presented in Table A.S4 and the distribution of the difference scores are displayed in Figure A.S2–4.

Table A.S4

Descriptive statistics of the difference score between the joint performance in the cooperative condition minus the maximum, sum, and mean performance of the two participants in the individual condition

Descriptive	Max	Sum	Mean
Median	-1.0	-3	0
MAD	1.0	1.0	1.0
Maximum	3.0	1	4
Minimum	-5.0	-6	-3
N	36	36	36
Missing	0	0	0

Note. MAD = median absolute deviation; *N* = sample size.

A

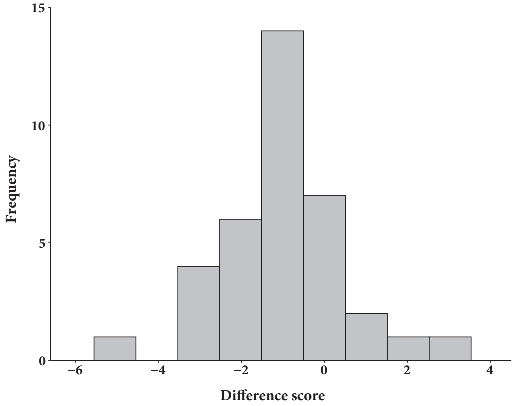

Figure A.S2. Distribution of the difference score between the joint performance in the cooperative condition minus the maximum performance of the two participants in the individual condition.

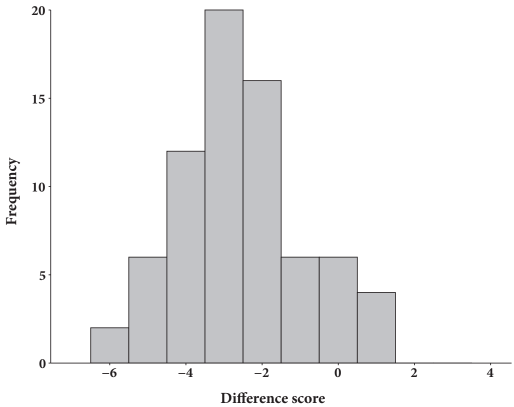

Figure A.S3. Distribution of the difference score between the joint performance in the cooperative condition minus the sum of performance of the two participants in the individual condition.

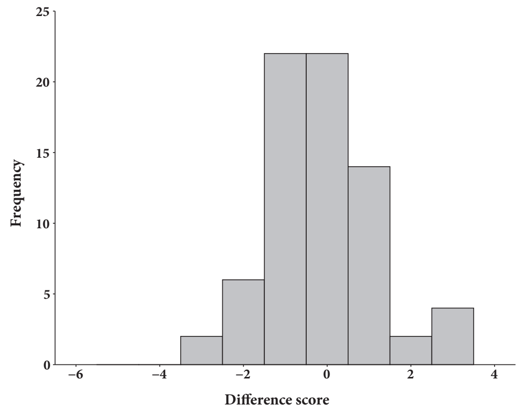

Figure A.S4. Distribution of the difference score between the joint performance in the cooperative condition minus the mean of performance of the two participants in the individual condition.

Does the difference in skills between participants influence the joint performance?

The Spearman's rho is .27 with *95% CI* (.04, .48) suggesting that there is a weak relationship between the mismatch in people's ability to perform the Tangram game (difference in completed puzzles between participants in the individual condition) and their performance together in the cooperative condition.

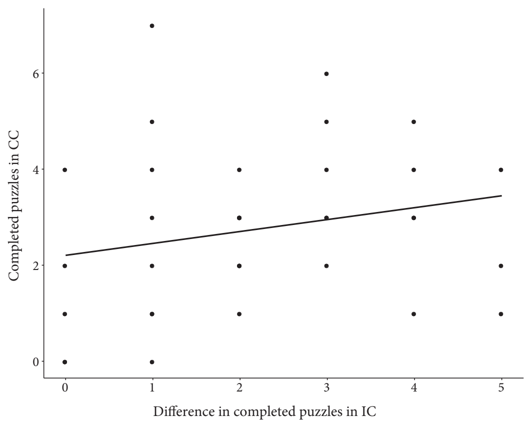

Figure A.S5. Scatterplot of the relation between the joint performance in the cooperative condition and the difference in completed puzzles between participants in the individual condition.

A

Appendix A4

Comparison of cooperation (success) rates between social dilemma game variants

In a follow-up analysis, we compared the cooperation rates between the Prisoner's Dilemma game, its extended version, and the Rope Pull game. The matched-pairs rank biserial correlation r is reported as a measure of effect size (Kerby, 2014). The largest difference was observed between the Prisoner's Dilemma and Rope Pull game ($Z = 4.82$, $r = .75$, $p < .001$) with higher cooperation rates in the latter. Also compared to the extended Prisoner's Dilemma game, the cooperation rate was higher in the Rope Pull game ($Z = 4.10$, $r = .62$, $p < .001$). Finally, participants cooperated more in the extended compared to the original Prisoner's Dilemma game ($Z = 2.49$, $r = .39$, $p = .013$).

Furthermore, we investigated not only the cooperation rate on the individual, but also on the dyadic level to see whether the willingness to cooperate translated into successful cooperation. Therefore, we calculated the proportion of how many times a dyad successfully cooperated from the ten trials in the three games. In the Prisoner's Dilemma and Rope Pull games, mutual cooperation occurred if both participants cooperated. In the extended Prisoner's Dilemma game, we looked at the mean of joint points participants received ranging from four (both fully defect) to six points (both fully cooperate). To make measures comparable, we transformed the joint points into proportions. The descriptive statistics are shown in Table A.S5 and visualized in Figure A.S6. Interestingly, the median of the Prisoner's Dilemma game substantially dropped compared to the cooperation rate at the individual level (from .60 to .30). In the other two games, mutual cooperation rates also decreased compared to cooperation rates on the individual level, but to a lesser extent (extended Prisoner's Dilemma: .68 versus .60; Rope Pull: .90 versus .80). Comparing the three games, participants succeeded least often in the Prisoner's Dilemma game and most often in the Rope Pull game (extended PD versus PD: $Z = 5.19$, $r = 0.52$, $p < .001$; RP versus PD: $Z = 5.61$, $r = 0.54$, $p < .001$; RP versus extended PD: $Z = 1.69$, $r = 0.21$, $p = .091$). This was also evident when looking at the proportion of trials participants cooperated successfully: in the Pull Rope game, almost half of the dyads mutually cooperated in all trials (.42), this proportion dropped to only .05 (= two dyads) in the Prisoner's Dilemma games. On the other hand, in the original Prisoner's Dilemma game, a quarter of the dyads (.25) did not successfully cooperate in any of the trials. In the extended Prisoner's Dilemma game, the success of mutual cooperation was evenly distributed throughout the spectrum.

Table A.S5

Descriptive statistics of the proportion of mutual cooperation for the games of the "social dilemma games" component

Game	Mean	Median	SD	MAD	Maximum	Minimum	Range
Prisoner's Dilemma	.35	.30	.31	.25	1.00	.00	1.00
Extended Prisoner's Dilemma	.63	.60	.23	.18	1.00	.17	.83
Rope Pull	.69	.80	.34	.20	1.00	.00	1.00

Note. SD = standard deviation; MAD = median absolute deviation.

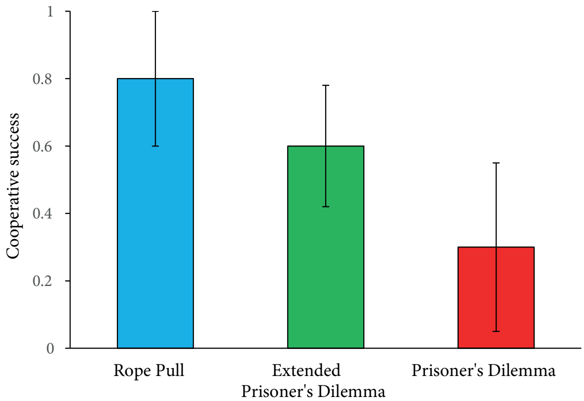

Figure A.S6. Median (± Median Absolute Deviation) of the cooperative success in the three variants of the social dilemma game.

A

In sum, on both the individual (willingness) and dyadic (success) level, cooperation was highest for the Rope Pull game and lowest for the classical Prisoner's Dilemma game. Intriguingly, in the extended Prisoner's Dilemma game, cooperation was higher compared to the original version, which is consistent with our previous studies (classical Prisoner's Dilemma game: .60 in current study and .58 in Behrens and Kret (2019); extended Prisoner's Dilemma game: .67 in the current study and .71 in Behrens et al. (2019)). Similarly, more choice options also yielded more successful cooperation, again, replicating our previous studies (classical Prisoner's Dilemma: .30 in the current study and .35 in Behrens and Kret (2019); extended Prisoner's Dilemma: .60 in the current study and .70 in Behrens et al. (2019)). The difference in cooperative success was mainly driven by a substantial proportion of dyads that always failed to cooperate (one-sided cooperation and mutual defection) when playing the classical Prisoner's Dilemma (where participants only had a dichotomous choice). Although a small majority was willing to cooperate (.60), dyads barely succeeded in *mutual* cooperation (.30). This was considerably less the case in the extended version (.67 versus .60). Our findings indicate that if people are given the option to indicate *how much* instead of *whether* they would like to work together, they are more inclined to and more successful at doing so. Thus, cooperation can be boosted by giving people multiple choices.

The results also demonstrate that the willingness and success of cooperation were considerably higher when participants played the Rope Pull game compared to both Prisoner's Dilemma games. One element that is incorporated in the former, but not the latter games is that participants received continuous feedback. Participants could adjust their behavior during a trial in response to the direct feedback of the rope. Research has indeed shown that making information about an interaction partner's decisions available to a participant facilitates cooperation (Behrens & Kret, 2019; Bixenstine & Wilson, 1963; Jorgenson & Papciak, 1981; Monterosso et al., 2002; Tedeschi et al., 1968). Another potential factor contributing to the discrepancy between the Rope Pull task and the Prisoner's Dilemma games is related to the payoff structures. In the Rope Pull task, the outcome of mutual defection and one-sided cooperation was the same (no reward), whereas in the Prisoner's Dilemma games, mutual defection led to a higher outcome (two points) than one-sided cooperation (one point). Therefore, if a participant predicts that the other person will defect, the preferred option in the latter two games is to defect as well, which is likely to elicit mutual defection in the subsequent rounds. However, if a participant predicts that the other person will defect in the Rope Pull game, she will receive no reward independent of whether she will defect or cooperate herself. Consequently, cooperation is wise because it might trigger the other to reciprocate in the next round. This shift towards mutual cooperation due to the payoff structure might therefore have inflated the cooperation rate in the Rope Pull game. We therefore argue that the greater willingness and success of cooperation in the Rope Pull game compared to the Prisoner's Dilemma games can be explained by: (1) receiving moment-to-moment feedback about the partner's intentions through pulling the rope, and (2) the payoff structure that gives no benefits to mutual defection over one-sided cooperation.

Appendix B

Supplementary material for Chapter 3

Appendix B1

Exploratory analysis of personality traits on cooperative behavior

This section includes exploratory analyses on how different personality traits (a) influence a person's willingness to cooperate and (b) modulate the effect of face-to-face contact on a person's willingness to cooperate.

There was great variation between dyads and players in how individuals were influenced by the face-to-face manipulation, as suggested by the large variances of the random effects of Dyad and Dyad * Player in the main analyses (see Result section of Chapter 3). In an attempt to explain these differences, we investigated how personality traits influenced the experimental manipulations on participants' willingness to cooperate. Participants completed questionnaires about their empathy level (IRI), social anxiety (LSAS), emotion recognition ability (Reading the Mind in the Eyes task), and social value orientation (SVO). For an overview of the descriptive statistics of these questionnaires in our sample and the correlations between them, see Tables B.S1-S2. For the IRI, we looked at the total score and the subscales perspective taking (PT), empathic concern (EC), fantasy scale (FS), and personal distress (PD) separately. Also, the two subscales anxiety and avoidance of the LSAS were analyzed individually. For both questionnaires, the mean was calculated per subscale per person. The SVO classifies individuals into four categories: prosocial, individualistic, competitive and no classification (Van Lange, 2000). The sample size for the latter three categories was too small to constitute a group, which is why we decided to combine them into the category "non-prosocial" (prosocial $n = 73$, non-prosocial $n = 33$ [consisting of 2 competitive, 15 individualistic, and 16 unclassified participants]). The performance for the Reading the Mind in the Eyes task was calculated based on the mean accuracy level (0 = incorrect, 1 = correct).

We concentrated the analysis on participants' own decisions rather than on the joint outcome as we expected individual characteristics to mainly influence individual decisions. Furthermore, we laid our main focus on the face-to-face manipulation and not on the feedback conditions as we were mainly interested in how personality traits would influence interpersonal communication rather than how people differ in their use of explicit, objective feedback.

The effect of nonverbal communication was moderated by the characteristics of the participants (Table B.S3). In particular, we observed a significant interaction between Face condition and SVO, IRI (total score and subscales PT and EC), Reading the Mind in the Eyes task, and LSAS anxiety scale (p's $\leq .002$). To disentangle these interaction effects and for the ease of interpretation, we median-splitted participants on the IRI, Reading the Mind in the Eyes task and LSAS anxiety scales. For these questionnaires, there was a significant difference between Face conditions for participants scoring low on each of the scales, but not for individuals having high scores (p's $\leq .04$; Figure B.S1 a-c). By visual inspection, it can be seen that participants who were less empathic and who had more difficulties to read another person's mind, generally cooperated less than people scoring high on these measures. On top of that, they were more strongly influenced by the experimental manipulation and were even less willing to cooperate in the face-blocked than the face-to-face condition. For the anxiety scale of the LSAS, people who scored below the median were again more influenced by the experimental conditions. In this case, low socially anxious people were more willing to cooperate when facing the other person compared to when not and high socially anxious people were generally less cooperative, regardless of the face condition. For the SVO, prosocial individuals cooperated more in the face-to-face compared to face-blocked condition, but non-prosocials were unaffected by the face manipulation (Figure B.S1d).

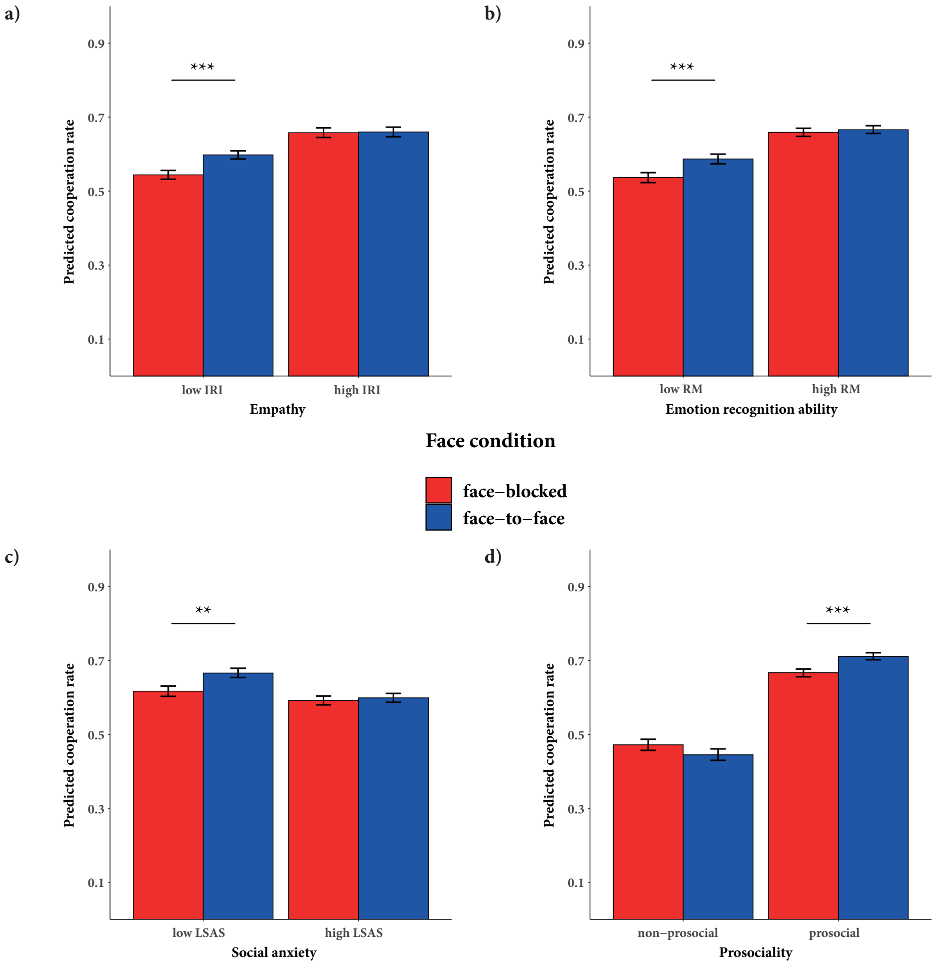

Figure B.S1. Predicted mean cooperation rate per Face condition moderated by (a) empathy (IRI total score), (b) emotion recognition abilities (RM = Reading the Mind in the Eyes task), (c) social anxiety (LSAS anxiety subscale), and (d) prosociality (SVO); ** $p < .01$, *** $p < .001$.

Table B.S1

Descriptive statistics of the personal characteristic questionnaires

Questionnaire	Mean	SD	Max	Min	$N_{Missing}$
RM	0.70	0.09	0.92	0.47	4
LSAS anxiety *	17	10	49	1	13
LSAS avoidance *	11	9	45	0	13
IRI Total *	101	15	134	66	13
IRI PT *	28	5	40	13	13
IRI EC *	27	5	39	11	13
IRI FS *	26	6	39	9	13
IRI PD *	20	6	34	6	13
Prosocial (SVO)	73 (68.9%)				

Note. * based on the summed score per subject.

Table B.S2

Pearson's correlation matrix with the RM, LSAS and IRI questionnaires

	RM	LSAS anxiety	LSAS avoidance	IRI Total	IRI PT	IRI EC	IRI FS
RM							
LSAS anxiety	0.160						
LSAS avoidance	0.12	0.70 ***					
IRI Total	0.071	0.308 **	0.17				
IRI PT	-0.108	0.101	0.071	0.510 ***			
IRI EC	0.038	0.213 *	0.10	0.731 ***	0.279 **		
IRI FS	0.068	0.071	-0.06	0.683 ***	0.099	0.289 **	
IRI PD	0.160	0.422 ***	0.33 **	0.710 ***	0.087	0.423 ***	0.310 **

Note. * $p < 0.05$; ** $p < 0.005$; *** $p < 0.001$; RM = Reading the Mind in the Eyes task; LSAS = Liebowitz Social Anxiety Scale; IRI = Interpersonal Reactivity Index; PT = Perspective Taking; EC = Empathic Concern; FS = Fantasy Scale; PD = Personal Distress.

Table B.S3

Descriptive statistics of model parameters regressing the interaction between each personality trait and the Face condition against the willingness to cooperate (defect = 0, cooperate = 1)

Effects	Test statistics	*p*-value
SVO * Face	$F(1, 9335) = 18.85$	**< .001**
Split by SVO		
Face (prosocials)	$B = .34$, $SE = .07$, CI (.21, .48)	**< .001**
Face (non-prosocials)		.086
IRI total * Face	F (1, 9088) = 11.22	**.001**
median-split IRI total		
Face (high IRI)		.345
Face (low IRI)	$B = .38$, $SE = .07$, CI (.24, .53)	**< .001**
IRI PT * Face	F (1, 9088) = 21.66	**< .001**
Median-split IRI PT		
Face (high IRI PT)		.465
Face (low IRI PT)	$B = .49$, $SE = .09$, CI (.33, .66)	**< .001**
IRI EC * Face	F (1, 9088) = 24.87	**< .001**
Median-split IRI EC		
Face (high IRI EC)		.052
Face (low IRI EC)	$B = .49$, $SE = .08$, CI (.34, .64)	**< .001**
IRI FS * Face		.213
IRI FS main effect		.986
IRI PD * Face		.040
IRI PD main effect		.507
RM * Face	F (1, 9870) = 10.22	**.001**
Median-split RM[+]		
Face (high RM)		.839
Face (low RM)	$B = .37$, $SE = .08$, CI (.20, .53)	**< .001**
LSAS anxiety * Face	F (1, 9088) = 12.26	**< .001**
Median-split LSAS		
Face (high LSAS anxiety)		.080
Face (low LSAS anxiety)	$B = .26$, $SE = .09$, CI (.09, .42)	**.002**
LSAS avoidance * Face		.311
LSAS avoidance main effect		.486

Note. P-values below the significance level of .005 are indicated in bold. SVO = Social Value Orientation; IRI = Interpersonal Reactivity Index; PT = Perspective Taking; EC = Empathic Concern; FS = Fantasy Scale; PD = Personal Distress; RM = Reading the Mind in the Eyes task; LSAS = Liebowitz Social Anxiety Scale. [+]random effect Dyad excluded because of the lack of enough variance.

Appendix B2

Exploratory analysis on the accuracy of predicting the partner's cooperative decision

In the current study, participants gave two responses per trial: whether they wanted to choose option A or B (corresponding to cooperating and defecting, respectively) and what they thought their partner chose. Based on this, we investigated whether individuals could read each other's intentions based on nonverbal cues only. To that extent, we conducted a one-sample t-test in the face-to-face / no feedback condition to compare the mean accuracy level to the level of chance. The results revealed that participants in this condition were not able to predict their partner's decisions ($M = .51$, $SD = .31$, $t(39) = .14$, $p = .890$).

Appendix B3

Exploratory analysis on the order effect of the Face condition on cooperative behavior

The order of the Face condition was counterbalanced between dyads. In an attempt to assure that the order did not affect the manipulation, we tested for the interaction between Face condition and the Face order. Surprisingly, this interaction was significant, $F(1, 10223) = 24.20$, $p < .001$ (see Figure B.S2). Specifically, the increase in the willingness to cooperate in the face-to-face compared to face-blocked condition was bound to those dyads who started the experiment in the face-to-face condition, $B = .42$, $SE = .07$, CI (.27, .56), $OR = 1.52$, $p < .001$, but not when they began in the face-blocked condition ($p = .180$). Similarly, the order of the Face condition influenced the successfulness of the joint outcome, $F(2, 5157) = 9.42$, $p < .001$. For all analyses of the study including the Face condition, we performed the analyses with and without including the interaction effect between Face condition and the Face order. None of the findings were influenced by it, so we only report the analyses without the interaction effect.

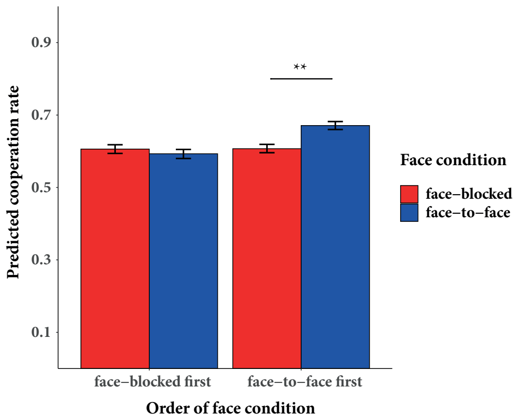

Figure B.S2. Predicted mean cooperation rate (± 2 SE) in the face-blocked and face-to-face condition moderated by the order of the condition (** $p < 0.005$).

B

Appendix B4

Descriptive statistics of participants' experiences during the study

Here, we present descriptive statistics of the experiences of the participants during the experiment. They filled out a visual analogue scale (VAS) after the first practice trails, after the first completed session, after the second practice trials and after the second session (the end of the experiment). Among others, they indicated how motivated they were, how much difficulty participants had to keep their attention to the task, how much they felt connected to their interaction partner and how anxious they felt. All questions were answered by setting a marker on a 10 cm long line ranging from "not at all" on the left to "very much" on the right. Additionally, participants completed the Positive And Negative Affect Schedule (PANAS) questionnaire before they started the game. Finally, participants filled out the Desire for Future Interaction scale (DFI) to indicate how much participants would like to meet their interaction partner again in different situations on a 5-point Likert scale. In Table B.S4, we present the mean and standard deviation of the VAS for each time point, the PANAS and the DFI.

Table B.S4

Descriptive statistics (mean and standard deviation) of the PANAS (Positive And Negative Affect Schedule), the VAS (Visual Analogue Scale), and the DFI (Desire of Future Interaction scale) participants completed before, during, and after the experiment, respectively

Question	After 1st practice trials	After 1st session	After 2nd practice trials	After 2nd session
How…do you feel at this moment?				
tense	2.5 (2.1)	1.8 (2.4)	1.9 (2.2)	1.4 (2.1)
awkward	3.0 (2.5)	2.2 (2.4)	2.8 (2.7)	1.5 (2.0)
shy	1.8 (2.0)	1.4 (1.9)	2.0 (3.4)	1.1 (1.7)
anxious	0.8 (1.1)	0.6 (1.0)	0.6 (1.1)	0.3 (0.6)
observed	4.5 (2.8)	3.8 (3.0)	3.7 (3.4)	2.8 (3.1)
Do you feel like the other sees right through you?	2.5 (2.4)	2.3 (2.4)	2.0 (2.5)	2.2 (2.7)
Do you feel connected to the other person?	7.4 (2.2)	7.0 (2.0)	7.5 (2.5)	7.3 (2.5)
How motivated are you to complete this task?	7.0 (1.9)	5.8 (2.8)	5.9 (2.6)	5.6 (2.7)
How difficult is it for you to keep your attention directed to the task?	2.2 (2.4)	5.0 (3.1)	4.5 (3.1)	5.7 (3.1)
Desire for Future Interaction (DFI)				3.21 (.62)
Positive And Negative Affect Schedule (PANAS)				
Positive subscale	3.22 (.51)			
Negative subscale	1.30 (.31)			

Note. The questions were answered by setting a mark on a 10 cm line, therefore the scale ranges from 0 to 10. The two PANAS subscales and the DFI were rated on a 5-point Likert scale before and after the experiment, respectively.

Appendix C

Supplementary material for Chapter 4

Appendix C1

Sensitivity Analysis

The sensitivity analysis has been proposed to be a valid post-hoc analysis in case an a priori power analysis has not been conducted before the study (Davis et al., 2018). In contrast to the traditional power analysis, where the relationship between power and sample size given a specified effect size is computed, the sensitivity analysis investigates the relationship between power and effect size given a particular sample size. The idea is to run simulation-based power analyses and detect the minimum true effect size that a study is sensitive enough to detect given a certain level of power (mostly, 80%) and a specific sample size.

The simulation-based sensitivity analysis includes the following steps that are repeated 1000 times: (i) simulate new data for the response variables based on the specified model (in our case, the full model shown in Table S2); (ii) refit the model to the new data; (iii) perform a statistical test on the effect of interest (in our case, the interaction effect between skin conductance level synchrony and Face condition). The assumption is that the effect of interest reflects the true population effect size, so every positive test is a true positive and every negative test is a false negative (i.e., a Type II error). Based on these results, the power can be directly calculated from the number of successes and failures (Green & Macleod, 2016). This power analysis is not only performed for the observed effect (in our case, the estimated interaction effect between skin conductance level synchrony and Face condition [.86], see Table S2), but also for a range of other effect sizes. Notice that the effect size is based on the scaled estimate of the model rather than a standardized effect size. For each effect size, the power to detect that effect (assuming that it is the true population effect size) is calculated resulting in the curve shown in Figure S1. The dashed line indicates the 80% power criterion and its associated true effect size (.70) that we can detect given our sample size. In other words, with our design, we would find a significant p-value in 80% of the cases if the true effect size was .70. The observed effect size of .86 is associated with a power of 89%, again assuming that the observed effect size reflects the true population effect size.

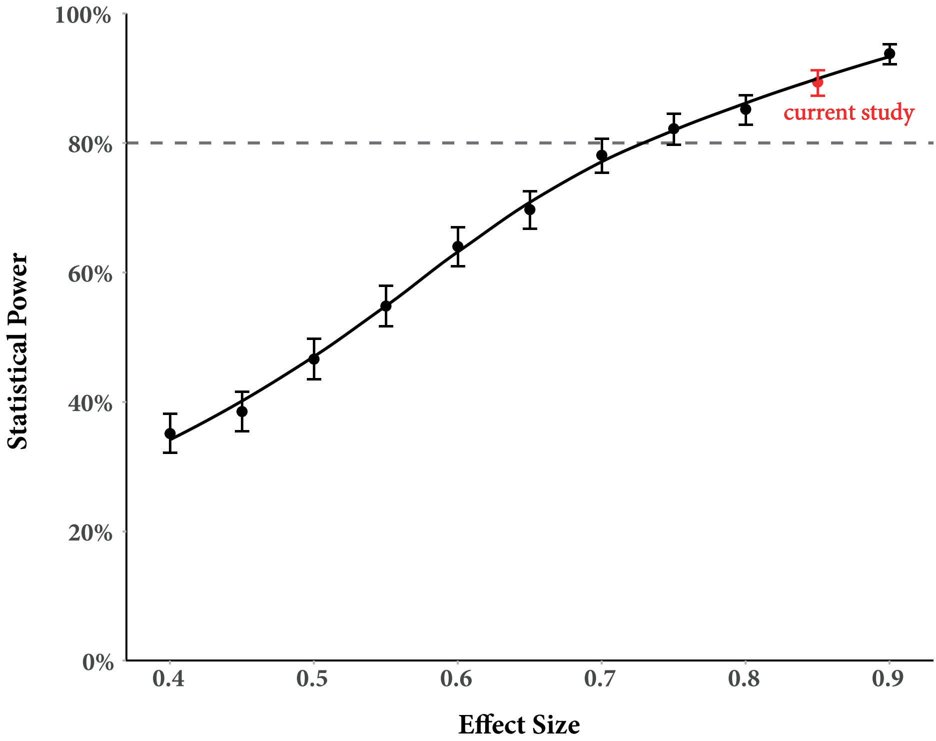

Figure C.S1. Simulation-based sensitivity analysis with statistical power as a function of different effect sizes. The observed effect size and associated power of the current study is marked in red. The dashed line marks an 80%-power threshold. The error bars reflect 95% confidence intervals. Notice that the effect sizes are based on the raw scale of the model and should not be interpreted following rule of thumb guidelines regarding the strength of the effect (e.g., Cohen's d).

C

Appendix C2

Information about the self-reported questionnaires

Table C.S1

Descriptive statistics of the self-reported questionnaires

Questionnaire	Mean	SD	Range	Theoretical range	$N_{missing}$
RM	26.05	3.77	12–34	0–37	1
LSAS	34.34	20.17	0–100	0–144	10
IRI	124.10	16.03	78–165	28–196	9
PANAS POS	30.30	6.99	11–46	10–50	1
PANAS NEG	13.44	3.42	9–25	10–50	1
DFI	3.17	.64	1.25–5	1–5	1
SVO (prosocial)	66.4%				9

Note. RM = Reading the Mind in the Eyes game; LSAS = Liebowitz Social Anxiety Scale; IRI = Interpersonal Reactivity Index; PANAS = Positive and Negative Affect Schedule; DFI = Desire for Future Interaction Scale.

Appendix C3

Quantification of physiological synchrony

Two methods that take non-stationarity into account are lagged windowed cross-correlation (Boker et al., 2002) and recurrence quantification analysis (Gates & Liu, 2016). The latter method is frequently used which has the advantage of having very few assumptions. However, the disadvantage is that it determines synchrony on a binary scale of moments being classified as either synchronized or not. The former method, albeit constraint by more assumptions, has the advantage of differentiating the degree of synchronization by quantifying it on a continuous (correlation) scale. Additionally, we feel that windowed cross-correlation is more intuitive to interpret. Consequently, we decided to apply this method which provides measures of the strength of synchrony and its variability.

The objective of the lagged windows-cross correlations analysis (Boker et al., 2002) is to calculate the strength of association between two time series while taking into account the non-stationarity of the signals and the lag between responses, that is, to consider the dynamics of a dyadic interaction. Specifically, the time series are segmented into smaller intervals, calculating the cross-correlation for each segment. This allows the means and variances to differ between segments accounting for non-stationarity. This is important as the level of synchrony may change during the experiment, sometimes having moments of strong synchronization while during other times responding less strongly to one another. Additionally, as the strength of association between two time points may differ depending on how far apart they are from each other, the segments are moved along the time series by an increment such that two adjacent segments overlap. Hence, segmenting the time series into smaller intervals and partially overlapping these intervals while moving along the time series provides a better estimate of the local strength of association between the physiological signals of two participants.

Besides the dynamics in the strength of synchronization during the course of the experiment, participants differ in how fast one might respond to a certain event or the other person. In other words, participants might not always be perfectly "in sync" whereby one participant might sometimes respond to the other person or vice versa introducing a delay between the responses of two individuals. To account for this, for each segment, the signals of the two participants are lagged in relation to one another. Specifically, the signal of participant 1 is kept constant while the signal of participant 2 is shifted more and more by a specified lag increment until a maximum lag is reached. Next, the same procedure is performed the other way around with participant 2 being kept constant. The maximum lag determines what is still considered synchrony. For example, if the maximum lag is four seconds, responses from two participants that are four seconds apart from each other are still considered synchronized. On the other hand, if one participant reacts to a certain event and the other participant shows a response 5 seconds later, it is not considered a response to the same event anymore and therefore does not count as synchrony. Based on this approach, there are four parameters that need to be determined: (1) the length of each segment, referred to the window size w_{max}; (2) the increment with which the segments are moved along the time series, the window increment w_{inc}; (3) the maximum with which two segments can be lagged

C

from one another, the maximum lag τ_{max}; and (4) the increment with which two segments are lagged from each other, the lag increment τ_{inc}. We determined the parameters following an extensive process by comparing previous studies using similar statistical methods, by looking at what is physiologically plausible given the time course of the physiological signals and by employing a data-driven bottom-up approach where we investigated how changing the parameters affected the outcomes using a different dataset. As expected, the absolute values of the synchrony measures varied depending on the parameters, but as supported by (McAssey et al., 2013), the relative results were not affected (e.g. a dyad manifesting relatively high synchrony showed such tendency for the different parameters). Based on these three factors, we set the parameters as follows: the window size was 8 seconds (160 samples), the window increment was 2 seconds (40 samples), the maximum lag was 4 seconds (80 samples) and the lag increment was 100ms (2 samples).

Calculating the cross correlations of each lag for each window segment generates a result matrix with each row representing one window segment and each column indicating a lag. The middle column represents the cross-correlation with a lag of zero, while the first and last column contain the cross-correlations for the maximum lag of participant 1 and 2. Hence, the number of columns in the result matrix is $(2^* \tau_{max} / \tau_{inc}) + 1$. The number of rows is given by $(N - w_{max} - \tau_{max})/w_{inc}$, with N being the number of observations in the whole time series.

Based on this result matrix, a so-called peak picking algorithm is applied. For each segment (i.e., each row in the matrix), the maximum cross-correlation across the lags is detected closest to the zero-lag (i.e., across all columns in a given row). If that maximum correlation is preceded and followed by smaller correlations, it is marked as a peak. For example, if participant 2 synchronizes with participant 1 with a lag of one second, the cross-correlations will become higher the closer the segments from the two participants are shifted towards the point where they are one second apart from each other. When the two signals are lagged by exactly one second the cross-correlation is highest (the peak). If the signals are lagged further away from each other, the cross-correlation decreases again. If, however, a peak cannot be detected, the algorithm assigns a missing value for that segment. This might be the case, for example, if people do not respond to an event or to each other (e.g., both participants wait and do nothing). The peak picking algorithm outputs a matrix with two columns, containing the value of the maximum cross-correlation (the peak) and the corresponding lag at which the peak cross-correlation is detected. The output has the same number of rows as the result matrix as it searches for a peak cross-correlation for each window segment.

Both the windowed cross-correlations and the peak picking algorithm are conducted four times per dyad, once for the heart rate responses and once for the skin conductance level responses for the face-to-face session and for the face-blocked condition resulting in $N_{dyads} * 4$ result and peak picking matrices. Finally, the mean of the peak cross-correlations of all window segments (i.e., all rows of the peak picking matrix) is calculated for both physiological measures per Face condition per dyad as the measure of synchrony and is grand-mean centered for the analysis predicting cooperative success.

Appendix C4

Model summary – main analysis

Table C.S2

Model summary of the multilevel linear regression analysis predicting cooperative success based on the level of synchrony in heart rate (HR) and skin conductance level (SCL) and their interaction with Face condition (face-blocked = 0; face-to-face = 1). Feedback condition (feedback no = 0; yes = 1) was included as a control variable and Dyad as a random intercept effect.

	Cooperative success				
Predictors	**Estimates**	**CI**	**t-value**	**df**	**p**
Intercept	5.07	4.85 – 5.29	46.65	49.31	**<0.001**
Feedback condition	0.20	-0.10 – 0.50	1.33	48.51	0.188
Face condition	0.10	0.06 – 0.13	5.47	2890.15	**<0.001**
HR synchrony	0.02	-0.64 – 0.67	0.05	2668.12	0.962
SCL synchrony	-0.01	-0.52 – 0.50	-0.04	2884.94	0.968
HR synchrony * Face condition	0.22	-0.28 – 0.72	0.86	2861.92	0.389
SCL synchrony * Face condition	0.86	0.34 – 1.38	3.24	2882.33	**0.001**
Random Effects					
σ^2	0.18				
$\tau_{00\ \text{Dyad}}$	0.28				
ICC	0.61				
N_{Dyad}	50				
Observations	2905				
Marginal R^2 / Conditional R^2	0.033 / 0.619				

Note. SCL = Skin Conductance Level; PPN = participant; CI = 95% confidence interval; σ^2 = residuals; $\tau_{00\ \text{Dyad}}$ = random intercept effect for Dyad; ICC = intraclass correlation.

C

Appendix C5

Control analysis – does arousal predict cooperative success?

In the current study we observed that physiological synchrony could predict cooperative success. One possible confound is that it is not the synchrony between two participants, but the co-occurrence of the arousal responses of the two individuals that drive these findings. For example, skin conductance levels might rise if a participant decides to cooperate due to the increased risk of being exploited. Similarly, if the other participant decides to cooperate as well, the same physiological reaction could be expected. Consequently, the responses of the two participants would highly correlate reflecting the individuals' decisions rather than an interpersonal process. To test this, we conducted a control analysis where cooperative success (the joint points won per trial) was regressed against the participants' skin conductance level and their interaction with Face condition (face-blocked = 0; face-to-face = 1). For the skin conductance level, we first standardized the responses per participant and then computed the mean skin conductance level per trial. Consistent with the model of the main analysis, we included the Feedback condition (feedback no = 0; yes = 1) as a control variable and Dyad as a random intercept effect. The model summary is displayed in Table C.S3 which shows that cooperative success could not be predicted by the arousal responses of the two individuals.

Table C.S3

Model summary of the control analysis (multilevel linear regression analysis) with participants' own skin conductance level (SCL PPN) and the interaction with Face condition (face-blocked = 0; face-to-face = 1) predicting cooperative success. Feedback condition (feedback no = 0; yes = 1) was added as a control variable and Dyad was included as a random intercept effect

	Cooperative success				
Predictors	**Estimates**	**CI**	**t-value**	**df**	**p**
Intercept	5.15	4.97 – 5.34	54.41	61.98	**<0.001**
Feedback condition	0.12	-0.14 – 0.38	0.91	60.95	0.362
Face condition	0.10	0.07 – 0.13	7.00	3566.66	**<0.001**
SCL PPN1	0.00	-0.02 – 0.03	0.13	3591.01	0.895
SCL PPN2	-0.02	-0.04 – 0.00	-1.64	3585.37	0.100
SCL PPN1 * Face condition	0.02	-0.02 – 0.06	1.12	3603.30	0.262
SCL PPN2 * Face condition	0.03	-0.01 – 0.06	1.39	3597.28	0.164
Random Effects					
σ^2	0.18				
$\tau_{00\ Dyad}$	0.27				
ICC	0.61				
N_{Dyad}	63				
Observations	3634				
Marginal R^2 / Conditional R^2	0.016 / 0.614				

Note. SCL = Skin Conductance Level; PPN = participant; CI = 95% confidence interval; σ^2 = residuals; $\tau_{00\ Dyad}$ = random intercept effect for Dyad; ICC = intraclass correlation.

C

Appendix C6

Control analysis – is the level of synchrony an artifact of the experimental set-up?

Because the heart rate and skin conductance level will always show a certain level of synchrony between participants due to the nature of the signals and the experimental set-up (Moulder et al., 2018), we conducted a control analysis to show that synchrony was elevated due to the interaction itself. Specifically, we compared the original dyads with newly generated dyads (Player 1 from $Dyad_i$ and Player 2 from $Dyad_{i+1}$). Because the trial length varied (there was no time restriction for making a decision), each trial was cut to the shorter trial of the newly generated dyad. Subsequently, the correlation between the responses of the two individuals was calculated per trial per dyad for heart rate and skin conductance level. Finally, we ran an independent t-test on the Fisher-Z-transformed correlation values between the original and the newly generated dyads. As a measure of effect size, we report Cohen's *d*. The results revealed that for both heart rate and skin conductance level synchrony, the level of synchrony was significantly higher in the original dyads compared to the newly generated dyads (HR: $t(3622.7) = 8.06$, $p < .001$, $d = .27$; SCL: $t(3015.5) = 4.38$, $p < .001$, $d = .15$). This indicates that the level of synchrony was due to the interaction rather than the experimental set-up of the study.

Appendix C7

Behavioral results

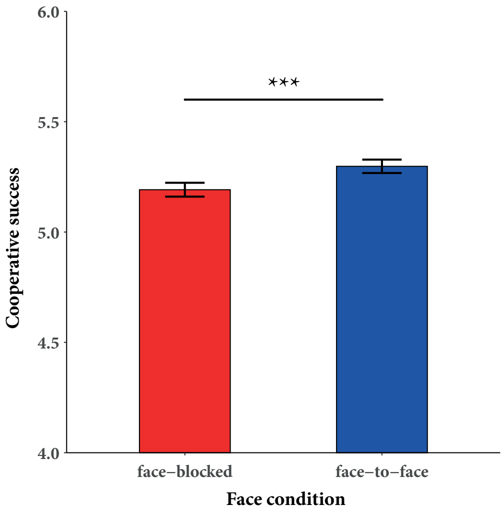

Figure C.S2. The cooperative success rate for the face-blocked and face-to-face conditions with error bars representing 95%-confidence intervals. * $p < .05$; ** $p < .01$; *** $p < .001$.