Physiological synchrony in the context of cooperation: Theoretical and methodological considerations

Behrens, F.

Chapter 5

Quantifying physiological synchrony through windowed cross-correlation analysis: Statistical and theoretical considerations

Abstract

Interpersonal synchrony is a widely studied phenomenon. A great challenge is to statistically capture the dynamics of social interactions with fluctuating levels of synchrony and varying delays between responses of individuals. Windowed Cross-Correlation analysis accounts for both characteristics by segmenting the time series into smaller windows and shifting the segments of two interacting individuals away from each other up to a maximum lag. Despite evidence showing that these parameters affect the estimated synchrony level, there is a lack of guidelines on which parameter configurations to use. The current study aimed to close this knowledge gap by comparing the effect of different parameter configurations on two outcome criteria: (1) the ability to distinguish synchrony from pseudosynchrony by means of surrogate data analyses and (2) the sensitivity to detect change in synchrony as measured by the difference between two within-subject conditions. Focusing on physiological synchrony, we performed these analyses on heartrate, skin conductance level, pupil size, and facial expressions data. Results revealed that a range of parameters was able to discriminate synchrony from pseudosynchrony. Window size was more influential than the maximum lag with smaller window sizes showing better discrimination. No clear patterns emerged for the second criterion. Integrating the statistical findings and theoretical considerations regarding the physiological characteristics and biological boundaries of the signals, we provide recommendations for optimizing the parameter settings to the signal of interest.

Based on: Behrens, F., Moulder, R. G., Boker, S. M., & Kret, M. E. (2020). Quantifying Physiological Synchrony through Windowed Cross-Correlation Analysis: Statistical and Theoretical Considerations. *bioRxiv*.

Introduction

During social interactions, humans tend to synchronize on different levels: They mimic postures (Ramseyer & Tschacher, 2011), facial expressions (Chartrand & Bargh, 1999) and align their level of physiological arousal (Feldman, Magori-Cohen, Galili, Singer, & Louzoun, 2011; Levenson & Gottman, 1983; Prochazkova et al., 2018). Although this synchrony comes naturally and without effort, it is a great challenge for social scientists to measure it statistically. The current paper addresses this issue and proposes a Windowed Cross-Correlation (WCC) analysis to investigate the dynamic changes in heartrate, skin conductance level, pupil size, and facial expression. Recommendations are provided on which parameter configurations to use to quantify synchrony of these four responses.

Synchrony is a multifaceted phenomenon evident on the behavioral, physiological, and neural level. Not surprisingly then, the causes and consequences of synchrony have been studied in a broad range of contexts investigating the dynamic nature of social interactions from clinical (Galazka et al., 2019; Wehebrink et al., 2018), developmental (de Klerk et al., 2018; Shih, Quiñones-Camacho, Karan, & Davis, 2019), evolutionary (Mancini et al., 2013; Palagi, Leone, Mancini, & Ferrari, 2009), neural (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004; Prochazkova et al., 2018), social (Behrens et al., 2019; Tarr, Launay, & Dunbar, 2016b), and cognitive (Kret, Fischer, & De Dreu, 2015; Kret & De Dreu, 2017) perspectives. Such fascination across disciplines has revealed the far-reaching scope of synchrony: it has been demonstrated in different species, it occurs from birth on, and it influences a variety of interpersonal processes such as marital quality, cooperative success between strangers and outcomes of therapeutic interactions (Behrens et al., 2019; Feldman et al., 2011; Kret, Tomonaga, & Matsuzawa, 2014; Levenson & Gottman, 1983; Ramseyer & Tschacher, 2011). Because of these implications and this wide interest, it is of particular importance to establish solid statistical methods to quantify synchrony.

A variety of methods have been proposed in previous literature to quantify synchrony including correlations, regressions, structural equation models and recurrence quantification analyses. These approaches differ in their assumptions, their operationalization of synchrony, and the type of synchrony they measure (for reviews, see Gates & Liu, 2016; McAssey et al., 2013; Schoenherr et al., 2018; Thorson, West, & Mendes, 2017). In the current article, we focus on continuous time series measures in dyads. For this type of data, it is important that the method captures responses that happen "in sync" (e.g., two individuals react simultaneously to an external event), but also responses that occur with a small time delay (e.g., one individual responds to another or at a different pace). Furthermore, the method needs to allow for changes in the level of synchrony as it will vary depending on the events happening in a conversation with moments of stronger and weaker synchrony. Moreover, we focus on the strength rather than the frequency of synchrony. Some methods first specify intervals of synchrony and subsequently compute the frequency of these intervals within a time series (Altmann, 2011). This method is particularly interesting for movement synchrony where people can either move or not. In the current study, on the other hand, we concentrate on physiological measures that constantly change, therefore categorizing intervals into synchronous and non-synchronous segments is difficult. Instead,

we are interested in obtaining a global estimate of the strength of synchrony in a conversation. A method that fulfills these different criteria is Windowed Cross-Correlation (WCC) analysis, the focus of the current study (Boker et al., 2002).

WCC analysis offers a neat method to account for dynamic changes in synchrony (Boker et al., 2002). This is achieved by extending a classical cross-correlation estimate by two aspects: windows and lags. Specifically, rather than calculating a correlation coefficient over the whole time series, the signals are broken into smaller overlapping segments or windows. Changes in synchronization can be captured because the degree to which two signals co-vary is estimated for each window separately. The lag is introduced to account for differences in the pace of individuals' responses to one another and to track the follow-lead relationship between them. It might be that at some point Person A responds to Person B and a moment later the pattern is reversed. Consequently, allowing for varying time lags can account for such dynamics. Although this method offers an advanced way to quantify synchrony in naturalistic settings, it does not come without a challenge: parameters need to be specified to tailor the analysis to the signal of interest. In the original paper by Boker and colleagues (2002), the authors advised on parameters using data from motor movements. To this date, there are no guidelines on which parameter settings are most suitable for physiological measures. The goal of the current paper is to close that knowledge gap.

WCC analysis requires the specification of four parameters that tailors the method to the signal of interest: window size, maximum lag, window increment, and lag increment (see Figure 1). Carefully choosing the right parameter settings is crucial, because these settings can substantially affect the outcome of the WCC analysis (Schoenherr et al., 2018). First, the window size determines the number of observations (i.e., data points) in each sliding window across the time series. The window should be small enough to be sensitive to changes in the degree of synchronization and the lead-follow relationship between individuals. Disregarding fluctuations within a large window might undermine the strength of association at certain moments. Here, the biological nature of the signal of interest and its time course are of particular importance. A relatively slow signal such as skin conductance requires a longer window than a fast signal such as facial expressions. Moreover, the window segments need to be small enough such that the assumption of stationarity is likely to hold (Boker et al., 2002). However, if the window size is too small, there are not enough data points left to provide reliable estimates of the relationship between the two segments. Whereas 50–70 values have been proposed as sufficient (Cappella, 1996), more recent work performing Monte-Carlo simulations recommends 65 to 250 values, depending on the strength of the correlation (Schönbrodt & Perugini, 2013). Given the high sampling rates incorporated in many psychophysiological measurement devices, this range should be fairly easy to accomplish, if the window size is not overly small. Decisions on the window size should be based on both statistical and theoretical considerations.

Second, the maximum lag indicates the maximum number of observations one window is shifted in relation to the other window and consequently determines the maximum lag two events are still considered reactions to one another. For example, if the maximum lag is three seconds, then if Person A smiles two seconds later in response to Person B, this would be captured with the three second window. However, if that smile occurs four seconds later, it would

not be considered a response to the smile of the other person anymore. If the maximum lag is too long, synchrony might be attributed to two unrelated events. However, if the maximum lag is chosen too small, then important delayed responses between two individuals are missed. Previous research suggests that the maximum lag between responses impacts on synchrony. Specifically, it has been shown that skin conductance responses within, but not beyond seven seconds correlate with the empathetic relationship between counselors and clients (Robinson, Herman, & Kaplan, 1982). The authors did not, however, directly compare whether the shorter latency could predict the relationship better than the longer latency. Additionally, although this study provides an indication that the maximum lag indeed matters, the categorization of latencies (responses between 0 and 7 sec compared to responses between 7 and 40 sec) does not allow for fine-grained conclusions about which maximum lag is optimal. To our knowledge, this is the only study investigating the impact of the maximum lag on synchrony. Thus, a systematic comparison of different maximum lags is needed to make well-informed decisions on this parameter.

Third, the window increment determines the size of the steps (i.e., the number of observations) when moving from one window segment to the next. If the increment is one, then the window is moved by one data point. If the window increment is the same size as the window size or greater, then adjacent windows are non-overlapping. Similarly, the fourth parameter, the lag increment, indicates how big the steps are between time lags. Both increment parameters regulate the resolution in terms of time lag and elapsed time. Ideally, the increment should be kept as small as possible to ensure the best resolution. However, at some point the estimates will stabilize and the limited additional information that can be added by increasing the resolution is not worth the increased computational time. Comparing it to sampling rates, if one aims to measure heartrate changes, a sampling rate of 1000 Hz gives a smooth signal. Increasing the sampling rate to 2000 Hz adds little information because the heartrate does not change this fast resulting in very similar heartrate signals using both sampling rates. Similarly, increasing the resolution of the increment of the moving windows and lags will eventually stabilize around a correlation estimate. The size of the increment will, of course, also depend on the sampling rate which represents the lower bound of possible increments. Therefore, setting the increment parameters for the windows and lags is a question of balancing the benefit of a better resolution and the drawback of increased computational time.

In order to determine the best parameter configurations, we used two criteria. The first criterion was the ability to discriminate synchrony from pseudosynchrony. Pseudosynchrony has been defined as "the amount of apparent and spurious synchrony between two individuals not engaged in information exchange with one another" (Moulder et al., 2018, p. 2). The reason for spurious synchrony is that the signals of interest are restricted in their patterns and how they can behave across contexts. For example, heartrate is constantly changing, decreasing and increasing depending on the person's inner state and environmental circumstances (i.e., participating in a study with the same procedure across dyads). However, the changes stay in a certain range causing recursiveness and commonality within and between heartrate measures. As a consequence, to determine whether synchrony exists between two time series, the null hypothesis is not zero as for standard null-hypothesis testing, but rather a fundamental value due to the similarities between the biological time series. It is therefore necessary to find an appropriate

comparison between the level of synchrony of individuals engaging in an interaction and the level of synchrony that occurs due to the nature of the signals. One way to account for pseudosynchrony is to perform a surrogate data analysis (Moulder et al., 2018). The idea is that the original time series is compared to the same time series where synchrony is destroyed while keeping all other properties constant. Specifically, the synchrony level from the original dyads engaging in an interaction is compared to the synchrony levels from newly generated dyads that never actually interacted. To generate these dyads, the time series from each participant is coupled with every other participant. That way it can be tested whether being in an interaction adds something over and beyond being in the same situation and investigating the same physiological measure. Therefore, being able to distinguish synchrony from pseudosynchrony offers an ideal criterion to test whether some parameter configurations are more sensitive to this distinction.

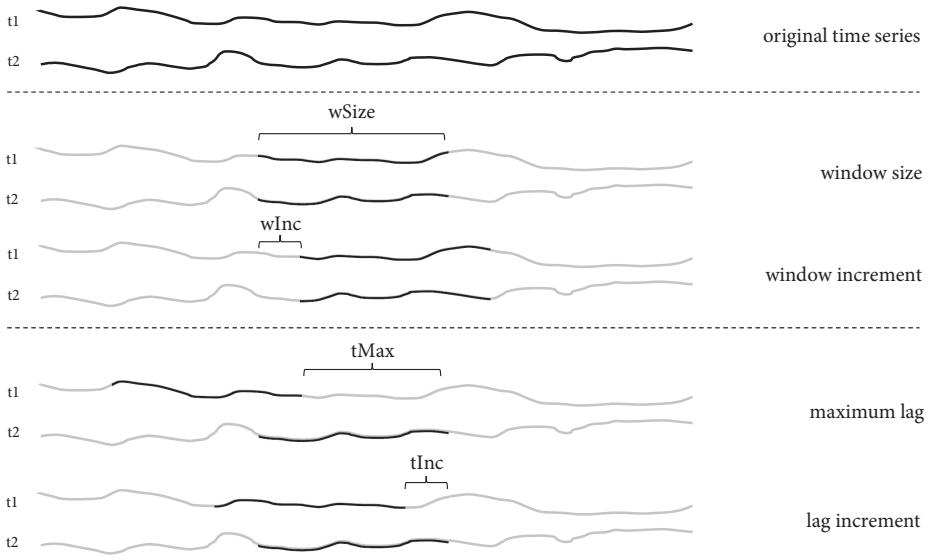

Figure 1. Schematic outline of the four parameters that are specified in the WCC analysis: window size (*wSize*), window increment (*wInc*), maximum lag (*tMax*), and lag increment (*tInc*). The abbreviations *tMax* and *tInc* originate from using "tau" (τ) to refer to the lags in the cross-correlation equation (see Equation 1).

The second criterion that is essential when it comes to research on synchrony is to be able to detect changes in synchrony. To study the underlying mechanisms of synchrony, its boundary conditions and individual differences, researchers are often interested in how synchrony changes in relation to experimental manipulations. For example, in a previous study, we observed that physiological synchrony promoted cooperative success, but only when partners could see each other and not when a cover prevented eye contact (our manipulation) (Behrens et al., 2019).

Another study investigated the effect of emotional salience during storytelling on pupil mimicry and showed that physiological coupling between the speaker and the listener was stronger during emotionally intense moments compared to less salient moments (Kang & Wheatley, 2017). Storytelling is particularly interesting because it is a uniquely human and universal activity creating social bonds between people (Smith et al., 2017). In Kang and Wheatley's (2017) study, listeners watched videos of speakers telling the story and therefore did not engage in an actual conversation. However, direct face-to-face interactions has been shown to affect synchrony levels (Behrens et al., 2019). Therefore, in the current study, two individuals engaged in face-to-face storytelling and completed baseline measures, silent moments of eye-contact. In line with the findings by Kang and Wheatley (2017), we expected higher levels of synchrony when people engaged in storytelling compared to the baseline measure. Ideally, the analysis that measures synchrony is sensitive to detect changes in synchrony between the two (within-subject) conditions.

The aim of the current study was to determine the best parameter configurations for the WCC analysis applied to different common physiological measures. The two criteria we used to decide on these configurations are (i) the ability to distinguish synchrony from pseudosynchrony and (ii) the sensibility to detect *changes* in synchrony (i.e., distinguish between two conditions). The reason to include two criteria is to investigate whether the purpose of the study (i.e., detect synchrony or change in synchrony) influences which parameters configurations are most suitable. We tested these criteria on data from dyadic interactions where two individuals told each other four stories. During the interaction, their heartrate, skin conductance level, pupil size, and contractions of the left zygomaticus major (a muscle associated with smiling) were measured. For a range of window sizes and maximum lags that were tailored to each signal, we calculated a measure of distance for the comparison (i) between the original dyads and newly generated surrogate dyads, and (ii) between intervals of storytelling and baseline measures in the original dyads. The window and lag increments were not systematically compared, but were adjusted as a function of the window size and maximum lag, respectively. Based on the outcome of these comparisons, we provide recommendations on which parameter configurations are best for detecting synchrony and change in synchrony for the four physiological measures. With these recommendations, we hope to help other researchers to make well-informed decisions in applying the WCC analysis and to increase the comparability of findings across studies.

Method

Participants

In total, 34 same-sex dyads participated in the study of which six dyads had to be excluded due to technical problems (dyads included in analysis: Female = 22 [78%]; M_{age} = 22.79; SD_{age} = 3.23; Dutch = 17 [30%]). Participants were recruited via the Leiden University online recruitment system, flyers distributed around the university building, and through personal contacts. In the latter case, participants were tested by a researcher they did not know. Individuals had normal or corrected-to-normal vision wearing contact lenses. Glasses were not compatible with the

eye-tracking glasses worn during the experiment. The duration of the study was about one hour and participants received two course credits or 6€, and chocolate for compensation. The study was approved by the local Psychology Ethics Committee of Leiden University (CEP19–0313/208).

Design

The design of the study is outlined in Figure 2. The study consisted of two parts. First, participants completed a breathing exercise where they were instructed to look at each other and synchronize their breathing. Second, participants engaged in storytelling with each participant telling a neutral and a positive story while the other participant was listening. Thus, participants told four stories in total with story 1 and story 3 always being told by participant 1 (sitting on the left side) and story 2 and story 4 being told by participant 2 (sitting on the right side). Story 1 & story 2 and story 3 & story 4 were of the same valence, with the order of starting with the neutral or positive story being counterbalanced between dyads. The breathing and storytelling parts were both preceded by a 2-min baseline measure where participants were instructed to relax and look at each other. After the second baseline measure and after each story, participants filled out the Positive And Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) to measure their current affect. Also, they rated each story with regard to its valence and intensity on a scale from 0 to 10. The PANAS and the story ratings are not discussed any further, but the descriptive statistics are provided in Appendix D2 (see Table D.S1).

Procedure

Upon arrival at the lab, participants were separated, received information about the study, and gave informed consent for participation. Afterwards, electrodes were attached to the torso, fingers, and face as preparation for the measurement of ECG, EDA, and EMG activity, respectively. Specifically, three electrodes were attached on the left and right side of the abdomen and on the thorax below the right collar bone to measure heartrate; two electrodes were attached to the non-dominant hand on the intermediate phalanges of the index and ring finger to measure skin conductance level; and three electrodes were attached to the left face on the zygomaticus major and behind the ear to measure facial expressions. The MP160 BIOPAC data acquisition system was used to record these measures at a sampling rate of 2000 Hz. After the preparation, participants filled out the Interpersonal Reactivity Inventory (IRI; Davis, 1980) and the Five Facet Mindfulness Questionnaire (FFMQ; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006) online. The descriptive statistics of both questionnaires can be found in Appendix D2 (see Table D.S1). Next, participants were seated on the same table and participants were asked to wear the eye-tracking device Tobii Pro Glasses 2 which were subsequently calibrated. Afterwards, the experimenters left the room and started the recordings of the physiological measures and the pre-recorded instructions that were provided via speakers. The experiment started with a 2-min baseline measure where participants were instructed to relax and look at each other (Baseline 1). Afterwards, the breathing exercise started where participants were again asked to look at each other, but this time synchronize their breathing for two minutes (not discussed in the current study). After this first

part of the experiment, participants had time to think of a neutral and positive personal story. When they were ready to begin, another 2-min baseline (Baseline 2) was taken and participants filled out the first PANAS which was provided on the table. Then Participant 1 (the individual at the left side of the table) started with the first story. Participants were instructed to talk for at least three minutes till they heard a beep and were requested to finish up. Afterwards, both participants filled out the PANAS and rated the story based on its valence and intensity on a scale between 0 and 10. Then the next story began. Participants took turns in telling them and filled out the PANAS and the rating after each story. At the end, participants put all filled out papers in an envelope, read the debriefing, and the experimenters removed the electrodes. Finally, individuals were paid and thanked for participation.

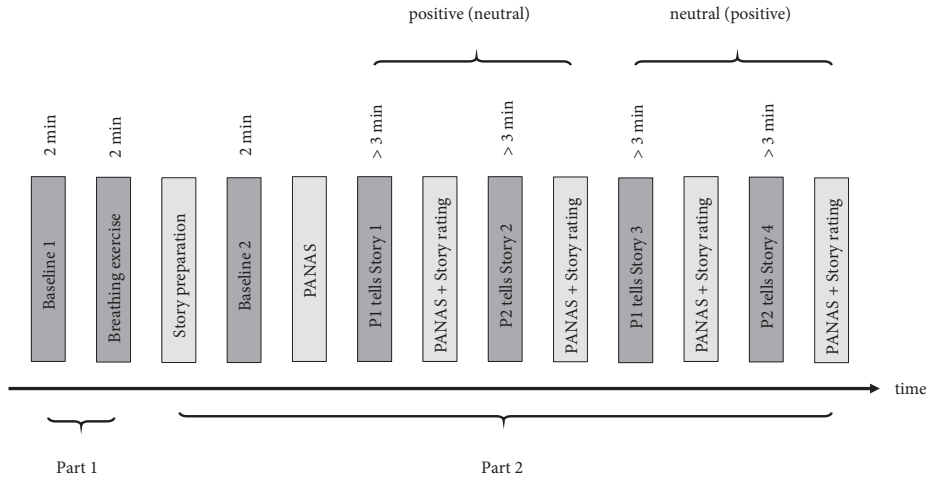

Figure 2. The time course of the study. The study was divided into two parts: breathing exercise (Part 1) and storytelling (Part 2). During the dark grey epochs, people interacted with each other; during the light grey epochs, they prepared the storytelling and filled out questionnaires; P1/P2 = Participant 1 and 2; PANAS = Positive And Negative Affect Schedule; Story 1 & 2 and Story 3 & 4 were of the same valence (positive or neutral); the order of starting with the positive or neutral story was counter balanced between dyads.

Preprocessing of the physiological measures

The physiological measures were pre-processed offline with the PhysioData Toolbox (Sjak-Shie, 2017). The heartrate data were preprocessed applying a band-filter between 1Hz and 50Hz. R-peaks were detected and transformed to inter-beat intervals (IBI) and subsequently to heartrate (bpm) values. The skin conductance signal was low-pass filtered with a cut-off of 5Hz. The EMG signal was preprocessed with a low-pass FIR filter of 28Hz and a high-pass FIR filter of 500Hz and a Notch-filter of 50Hz. The rectified signal was subsequently smoothed with a Boxcar filter of 100ms. The pupil size data were preprocessed in multiple stages according to recommended

guidelines described elsewhere (Kret & Sjak-Shie, 2018). After applying the filters, each signal was visually inspected and if necessary, manually corrected. If missing or incorrect intervals were manually detected, the signals were linearly interpolated. Finally, all signals were down-sampled to 20Hz.

Windowed Cross-Correlation analysis

Two challenges in analyzing physiological responses between two individuals include i) to statistically represent the dynamics of an interaction and ii) to quantify the associated patterns that might vary in the strength of association and the timing of the responses. Windowed Cross-Correlation (WCC) analysis offers a method that addresses both challenges. Specifically, the two time series are broken into smaller, overlapping windows before the correlation is estimated for each window. This way, the strength of association can vary between these windows accounting for the non-stationarity of the signals. The overlap between windows assures that strong synchronization that occurs at the edge of non-overlapping adjacent segments is not missed. Additionally, for each window, the two segments are lagged away from each other up to a maximum lag such that the segment of either participant 1 or participant 2 precedes the other participant's segment in time. This way the method accounts for the (varying) delay between two responses. This generates a result matrix r with correlations for the different segments and time lags defined as

$$r(Wx, Wy) = \frac{1}{T_W} \sum_{t=1}^{T_w} \frac{(Wx_t - \overline{Wx})(Wy_t - \overline{Wy})}{sd(Wx)sd(Wy)} \quad (1)$$

Where T_w is the total amount of observations (i.e., data points) in each window Wx and Wy consisting of observations Wx_t and Wy_t where $t \in \{1, \ldots, T_w\}$, Wx and Wy are the means of the observations in each window, and $sd(Wx)$ and $sd(Wy)$ the standard deviations of each window. In the result matrix, each row represents one window, while each column represents one lag. Because the first window needs to lag segments up to the maximum lag and because the window includes more than one data point, the number of rows is given by $(N - wSize - tMax) / wInc$. Dividing by *wInc* accounts for how many observations are skipped between one window and the next one. For example, if the window increment is one, then the number of rows of the result matrix will be equal to the number of observations of the time series (after accounting for the window size and maximum lag as just described). But if the increment is 10, then the steps are bigger between the windows, reducing the number of segments needed to cover the whole time series and therefore decreasing the number of rows in the result matrix. The number of columns in the result matrix is $(tMax * 2)/tInc + 1$ because the segments are shifted such that first Participant 1 and then Participant 2 precedes the other participant up to the maximum lag (i.e., twice the *tMax*). The *tInc* accounts for the size of the steps between two lags. The extra column (+1) represents the case where the lag is zero.

Peak picking. Following the WCC analysis, Boker et al. (2002) developed the so-called peak-picking algorithm where the maximum correlation across different lags is determined for each window (i.e., the maximum correlation per row of the result matrix). The maximum correlation should be preceded and succeeded by lower correlation values. For example, if Participant 1 synchronizes with Participant 2 at a lag of 1 second, then the correlation should be highest (i.e., peak) at that time lag and the correlation should be lower at both lag .5 and 1.5 seconds. This "peak" criterion is implemented to ensure that individuals indeed react to one another. If both individuals did nothing, they both would show more or less flat lines in their physiological responses and the correlation between their signals would be high for all lags. Requiring a peak in the correlation across lags prevents such events from being termed "synchrony". The peak-picking algorithm outputs a matrix with the maximum ("peak") correlation and its corresponding time lag for each window. In a last step, a summary statistic is computed by calculating the mean of the maximum correlations. This measure provides an indication of the overall level of synchrony between the two time series.

Choosing values for parameter configurations

As mentioned above, there are four parameters that need to be specified: window size, window increment, maximum lag, and lag increment. The window size (*wSize*) determines how long each window is, the window increment (*wInc*) indicates the size of the steps between two adjacent (overlapping) windows, the maximum lag (*tMax*) regulates how far the segments of the two time series are shifted away from each other, and the lag increment (*tInc*) determines the size of the steps with which the segments are shifted.

To choose the range of values we considered for the window size and maximum lag parameters, we employed a bottom-up approach by running preliminary WCC analyses on the whole time series (including all data of the study). Inspecting the result matrix plots, we examined the patterns seen in these plots. Examples of a "good" and "bad" parameter configurations are shown in Figure 3. Good parameter configurations show sharp contrasts between regions of high and low synchrony. The bad choices show a more smoothed image and thus less contrast between these regions, making differences more difficult to detect.

With regard to the maximum lag, we examined the plots inspecting whether the peak correlations fell within the range of lags or whether they fell outside the plots (not shown in Figure 3). For reasons of simplicity, the range of maximum lags was equal to the range of window sizes. In addition to the visual inspection, we ensured that the range of parameters included the parameters previously used in the literature. Finally, the minimum value for the window size was set to 3 sec to include at least 60 data points (20Hz sampling rate) per window size which is in line with previous guidelines for reliably estimating correlation coefficients (Schoeneberger, 2016). The window size and maximum lag parameters chosen for each physiological measure are listed in Table 1. For the window and lag increment parameters, we used 1/10th of the window size and the maximum lag, respectively.

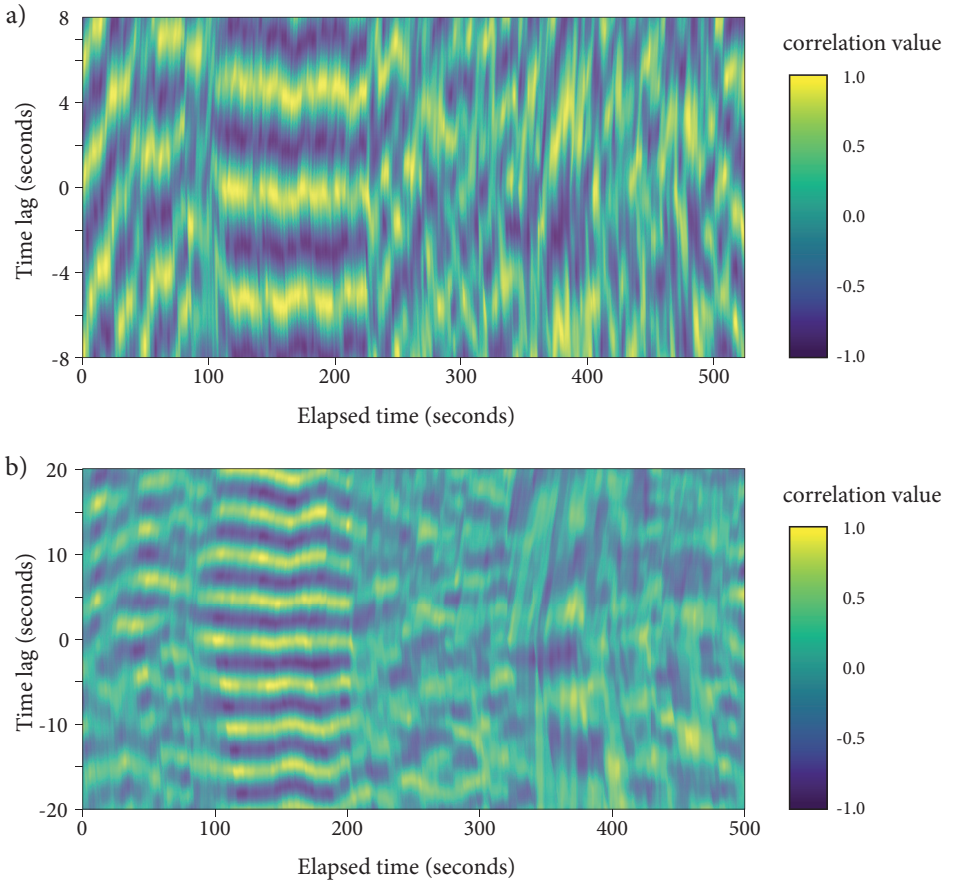

Figure 3. Examples of WCC analysis plots using heartrate data and a window size and a maximum lag of (a) 8 sec and (b) 20 sec, representing a "good" and "bad" example of parameter settings, respectively. Between around 100 and 200 seconds, people engage in a breathing exercise where they breathe synchronously which is reflected in the steadily high correlations around the time lag of zero.

Table 1

Window size and maximum lag parameters used for each physiological measure

Signal	Window size	Maximum lag
Heartrate	4 – 12 sec in steps of ½ sec	4 – 12 sec in steps of ½ sec
Skin conductance level	5 – 25 sec in steps of 1 sec	5 – 25 sec in steps of 1 sec
Pupil size	3 – 9 sec in steps of ½ sec	3 – 9 sec in steps of ½ sec
Facial expression	3 – 9 sec in steps of ½ sec	3 – 9 sec in steps of ½ sec

Note. The window and lag increments were equal to 1/10th of the window size and the maximum lag, respectively.

Choosing the best parameter settings

We conducted the WCC and peak-picking analyses for all combinations of the window size and maximum lag parameters with their corresponding increments as described in the previous section. For each parameter configuration, we calculated the mean peak correlation across window segments per dyad as the measure of synchrony. To determine the best parameter configurations for each physiological measure we used two criteria: (i) the ability to discriminate synchrony from pseudosynchrony, and (ii) the ability to detect change in synchrony. For the first criterion, we compared the original dyads consisting of the individuals who in fact interacted with each other during the experiment with the surrogate dyads consisting of all possible combinations of pairing individuals who did not interact during the experiment. If being in the specific social interaction evoked synchrony above and beyond the synchrony evoked by the fact of being in *any* actual interaction, synchrony levels are expected to be higher in the original compared to the surrogate dyads. Therefore, we calculated the mean peak correlation for both the original and the surrogate dyads and investigated whether specific parameter configurations were more sensitive to detect the difference between synchrony (original dyads) and pseudosynchrony (surrogate dyads). Sensitivity was quantified by the t-statistics of an independent t-test between the mean estimates of the two groups. A positive t-statistic indicates that the true dyads show higher levels of synchrony than the surrogate dyads. To determine the best parameter configuration, we located which configuration generated the largest t-statistic and inspected the pattern in changes of t-statistics across parameter configurations. Note that we used the t-statistic as a measure of distance between the two group means without running hypothesis testing (i.e., decide on whether the distance is significant or not). We therefore interpret the t-statistics in relative rather than absolute terms and do not draw any conclusions about whether the differences reveal significant results or not. The analysis was conducted with the data from the first baseline measure (see Figure 1). To investigate whether the results of this analysis would replicate, we additionally conducted the same analysis again with data from the second baseline measure.

For the second criterion, that is, which parameter configurations are most sensitive to detect change in synchrony, we concentrated on the original dyads and investigated which parameter configurations generated the biggest difference between two conditions of the experiment. We used the t-statistic based on a paired t-test as a measure of distance between the mean estimates of

the two conditions. A positive t-statistic indicates higher levels of synchrony during storytelling than baseline. Similar to the first criterion, we identified the largest t-statistic and inspected the pattern in changes of t-statistics across parameter configurations. We also ran the analysis twice. First, we compared story 1 and story 3 with the two baseline measures. Second, we compared story 2 and story 4 with the two baseline measures (see Appendix D1 for the reasoning behind the choice of these comparisons). To keep the length of the stories equal, we only used the first three minutes of each story. This way, both comparisons included a positive and a neutral story (a preliminary analysis yielded no differences between the positive and negative stories). The only difference was that in the first analysis, Participant 1 told the stories and in the replication analysis, Participant 2 told the stories. Being Participant 1 or 2 was based on the participant number and therefore should not have had any systematic impact on the synchrony level between the two individuals. Therefore, we could investigate whether specific parameter configurations were more sensitive than others to detect differences in synchrony levels when people just looked at each other compared to when they engaged in storytelling.

Results

Synchrony versus pseudosynchrony

Heartrate. There was a range of positive t-statistics indicating that multiple parameter configurations could differentiate between the original and the surrogate dyads (Figure 4a). The best discrimination (maximum t-statistic = 28.32) was evident for the smallest window size (4 sec) and a maximum lag of 7.5 sec (the most yellow combination in Figure 5a). When mapping the t-statistics distribution onto the parameter configuration space, a clear pattern emerged: the smaller the window size, the larger the t-statistics. This pattern was evident by the gradual changes in coloring from blue to yellow in Figure 5a when moving down the y-axis (i.e., moving from large to small window sizes). When the window size became too large, the synchrony level dropped in the original dyads such that it became lower than the synchrony level apparent in the surrogate dyads (especially, when the maximum lag was small; dark blue coloring in Figure 5a).

The maximum lag was less influential on differentiating between original and surrogate dyads than the window size, yet not trivial. The maximum t-statistic was evident for a maximum lag of 7.5 sec. The optimal maximum lag was therefore around twice the optimal window size (4 sec). Increasing or decreasing the maximum lag reduced the sensitivity to distinguish between the original and surrogate dyads as indicated by less yellow colors when moving left or right on the x-axis in Figure 5a. The replication analysis using data from the second baseline measure revealed similar results to the primary analysis and is depicted in Figure D.S1a-D.S2a. The maximum t-statistic (35.23) for a window size of 4 sec was replicated. The maximum lag differed slightly by 1.5 sec showing the highest t-statistic at 9 sec. However, the pattern was comparable with smaller window sizes and maximum lags around twice the window sizes yielding the largest difference between the original and surrogate dyads. In conclusion, if the aim of the study is to verify whether synchrony evolved as a result of interpersonal processes during a conversation above and beyond the shared environment of two participant, the range of parameters able to

detect that difference is rather wide. In general, we recommend using a small window size for heartrate synchrony. Regarding the maximum lag, the choice of parameters is less influential, however, we recommend using a maximum lag that is around twice the window size.

Skin conductance level. As with heartrate synchrony, there was a range of parameter configurations with a positive t-statistic that was sensitive to distinguish the original from the surrogate dyads (see Figure 4b). The largest t-statistic of 37.71 was observed for a window size of 6 sec and a maximum lag of 24 sec (see Figure 5b). Similar to the heartrate data, the smaller the window size, the greater the distance in estimated means between the original and surrogate dyads. Also, the outcome flipped with higher synchrony levels for the surrogate compared to the original dyads when the window size was too large paired with a small maximum lag. In contrast to heartrate, the discriminative ability steadily increased when the small window size was combined with an increasingly larger maximum lag (around four times the window size). In the replication analysis, the same pattern emerged as in the primary analysis: the greatest discrimination was seen for a small window size and a large maximum lag (see Figure D.S1b-D.S2b). The largest t-statistic (48.71) was observed for a window size of 5 sec and a maximum lag of 21 sec. Again, when the window size became too large paired with smaller maximum lags, the analysis would estimate higher synchrony levels for the surrogate compared to the original dyads. Based on these results, we recommend using a small window size and a large maximum lag that is around four times the window size.

Pupil size. The number of positive t-statistics depicted in Figure 4c indicates that there was a range of parameter configurations that could differentiate synchrony from pseudosynchrony. The maximum t-statistic of 16.12 was associated with a window size of 3 sec and a maximum lag of 9 sec. The general pattern as for the other measures was observed: the smaller the window size, the greater the difference between the original and surrogate dyads (see Figure 5c). Again, when the window size became too large, the estimates of synchrony level would become larger for the surrogate compared to the original dyads. With respect to the maximum lag, it was less influential than the window size, but showed a slight tendency to larger maximum lags. A similar pattern was observed for the replication analysis with a maximum t-statistic (18.04) evident for a window size of 3 sec and a maximum lag of 6.5 sec (see Figure D.S1c-D.S2c). In conclusion, smaller window sizes were more sensitive to distinguishing synchrony from pseudosynchrony in pupil size data. The maximum lag did not have as much of an impact, but should be set to two to three times the window size.

Facial expression. All t-statistics were positive indicating that the level of synchrony was higher for the original compared to the surrogate dyads for all parameter configurations. However, compared to the other three measures, the distribution showed less variance with t-statistics ranging from 1.68 to 5.14 (see Figure 4d). The latter was observed for a window size of 3 sec and a maximum lag of 5 sec. As shown in Figure 5d, the same pattern as for the other three measures emerged: the smaller the window size, the better the original dyads could be distinguished from the surrogate dyads. Furthermore, the maximum lag did not have a great impact on the discriminative ability, but the largest t-statistic was observed at almost twice the window size (5 sec). For the replication analysis, a similar pattern was observed (see Figure D.S1d-D.S2d) with a slightly wider range of t-statistics (maximum = 7.05; minimum = -.16). The maximum t-statistic was asso-

ciated with a window size of 3 sec and a maximum lag of 8.5 sec. Again, the smaller the window size, the greater the difference between synchrony and pseudosynchrony with limited impact of the maximum lag. In conclusion, we recommend using a small window size and a maximum lag that is two to three times the window size.

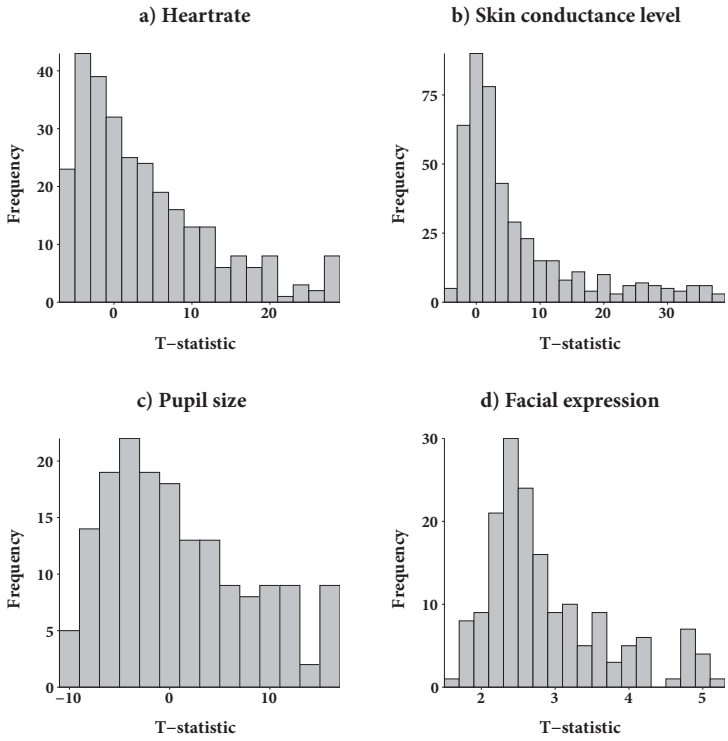

Figure 4. Distribution of t-statistics of the comparison between the original and surrogate dyads for each physiological measure. A positive value indicates higher synchrony level in the original compared to the surrogate dyads. Each data point represents one parameter configuration. For the analyses, data from the first baseline measure were used.

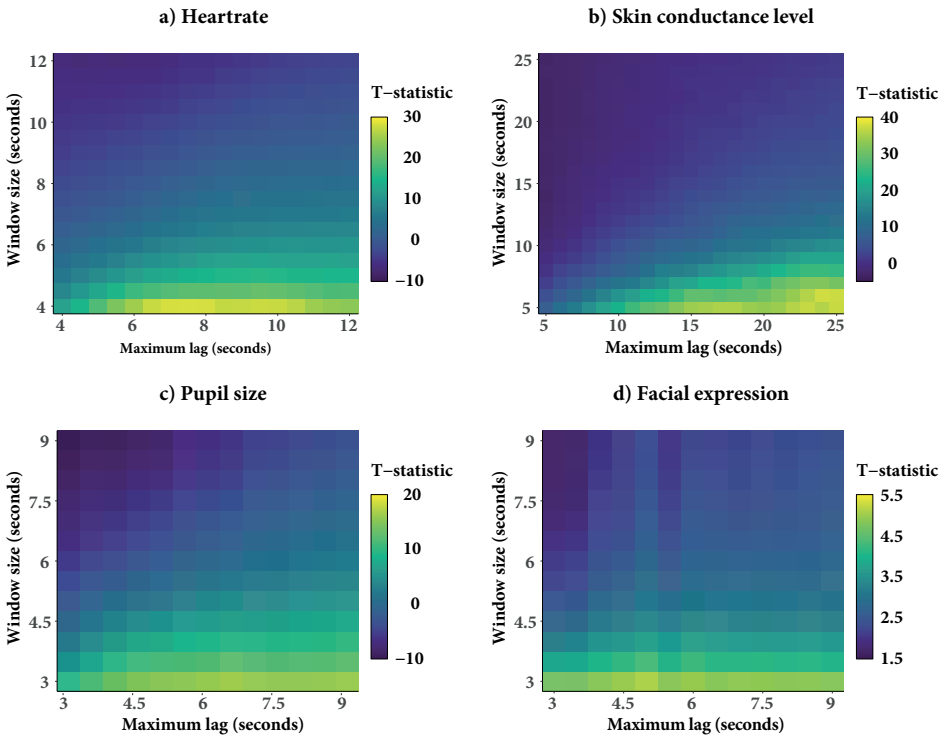

Figure 5. Distribution of the t-statistics of the comparison between the originate and surrogate dyads for all parameter configurations and each physiological measure. The color coding runs from the lowest (blue) to the highest (yellow) t-statistic. A positive t-statistic indicates that the original dyads showed higher synchrony levels than the surrogate dyads. The more yellow, the better the discrimination between the original and surrogate dyads. Data from the first baseline measure were used. Notice that the scaling of the axes and the color coding are adjusted to each physiological measure to increase comparability between parameters.

Change in synchrony

Heartrate. The largest absolute t-statistic was negative indicating that synchrony levels were higher during baseline compared to during storytelling (see Figure 6a). The highest absolute t-statistic of 4.86 was observed when the window size was set to 4 sec. Similar to the first comparison analysis, smaller window sizes could discriminate the two conditions better than large window sizes (see Figure 7a). Also, the maximum lag was less influential than the window size parameter, but the best outcome was observed for the smallest maximum lag of 4 sec. The absolute t-statistic steadily decreased with increasing maximum lags. For the replication analysis, the results were similar to the primary analysis, with smaller window sizes showing the greatest discriminative power between the conditions (see Figure D.S3a-D.S4a). Specifically, the largest absolute t-statistic was again observed for a window size of 4 sec. The maximum lag increased from 4 to 7 sec in the rep-

lication analysis with only slight changes across maximum lags. Therefore, based on both analyses the conclusion is: if the aim is to distinguish synchrony levels in heartrate responses between two (within-subject) conditions, the smaller the window size, the better. The maximum lag is less influential, but should be equal to or twice the window size.

Skin conductance level. All t-statistics were negative indicating that the level of synchrony was higher during the baseline measures compared to during storytelling (see Figure 6b). The highest absolute t-statistic of 4.37 was observed for a window size of 18 sec and a maximum lag of 25 sec. Interestingly, the previous pattern of smaller window sizes showing greater t-statistics was not evident (see Figure 7b). In fact, although there seemed to be a weak tendency for absolute t-statistics to become larger with larger window sizes and larger maximum lags, the pattern was rather weak. In addition, the difference between t-statistics was small ranging from -1.61 to -4.37. For the replication study, the range was also rather narrow from -.19 to -2.56 (see Figure D.S3b-D.S4b). The maximum absolute t-statistic was observed for a window size of 5 sec and a maximum lag of 12 sec, deviating substantially from the primary analysis. Although the general pattern (i.e., the smaller window size, the higher the t-statistic) was observed to a stronger degree compared to the primary analysis, it was still weak. In conclusion, given the lack of clear patterns in the parameter configuration space and considerable discrepancies in the results between the primary and replication analyses, we cannot draw strong conclusions about which parameter configuration is best to distinguish between two conditions when looking at skin conductance level synchrony.

Pupil size. For this measure, the parameter configurations strongly influenced whether synchrony levels were higher during baseline or storytelling (see Figure 7c). Generally, if both the window size and the maximum lag were small, synchrony levels were higher during storytelling; if the window size and maximum lag were large, synchrony levels were higher during the baseline measures. Specifically, the largest positive t-statistic of 1.72 (storytelling showed more synchrony) was observed for a window size of 3.5 sec and a maximum lag of 3 sec. However, the largest absolute t-statistic of 2.07 (baseline showed more synchrony) was associated with a window size of 8.5 sec and a maximum lag of 9 sec. A similar, but weaker pattern was evident for the replication analysis (see Figure D.S3c-D.S4c). The window sizes and maximum lags associated with the largest (absolute) t-statistic were the same as for the primary analysis. Given the ambiguity across parameters, we refrain from providing any recommendations about the best parameter configurations when the aim is to detect change in pupil size synchrony between conditions and instead caution that parameter choices can have a large effect on the outcome of this type of study.

Facial expressions. All t-statistics were positive indicating that the level of synchrony was higher during storytelling compared to baseline (see Figure 6d). The largest t-statistic of 3.99 was evident for a window size of 3 sec (see Figure 7d). Albeit weak, the general pattern emerged with larger t-statistics being associated with smaller windows sizes. The maximum lag associated

with the biggest difference between conditions was 3.5 sec, but the differences across lags were trivial. The replication analysis revealed similar results with the largest t-statistic (4.53) observed for a window size of 3 sec (see Figure D.S3d-D.S4d). The maximum lag of 9 sec deviated from the primary analysis, however, the differences across the maximum lags were again rather small. To conclude, if the aim is to detect change in synchrony between two conditions in facial expressions, then the window size should be set to a small value. The effect of the maximum lag was negligible, however, to be consistent with the other measures, we recommend a maximum lag twice the window size.

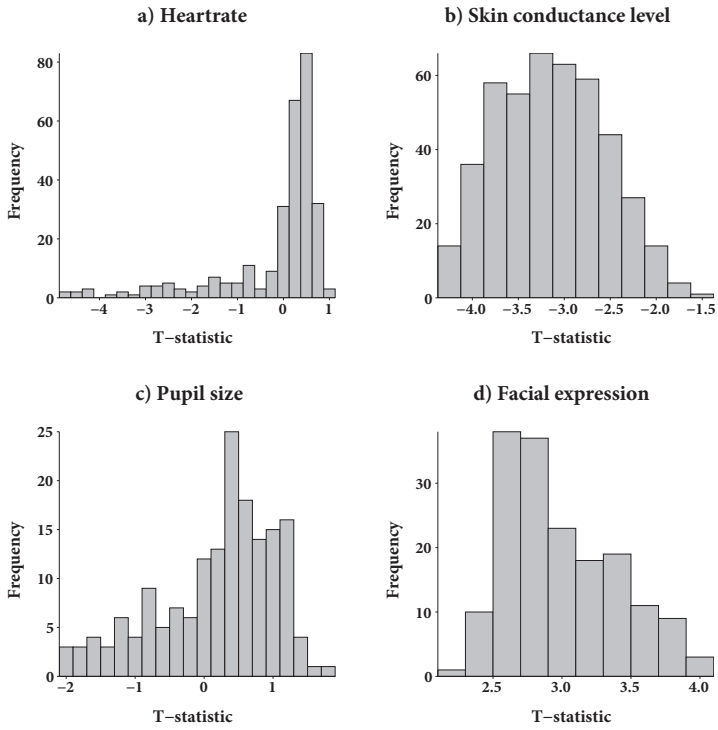

Figure 6. Distribution of t-statistics of the comparison between storytelling and baseline for each physiological measure. A positive value indicates higher synchrony levels during storytelling compared to baseline. Each data point represents one parameter configuration.

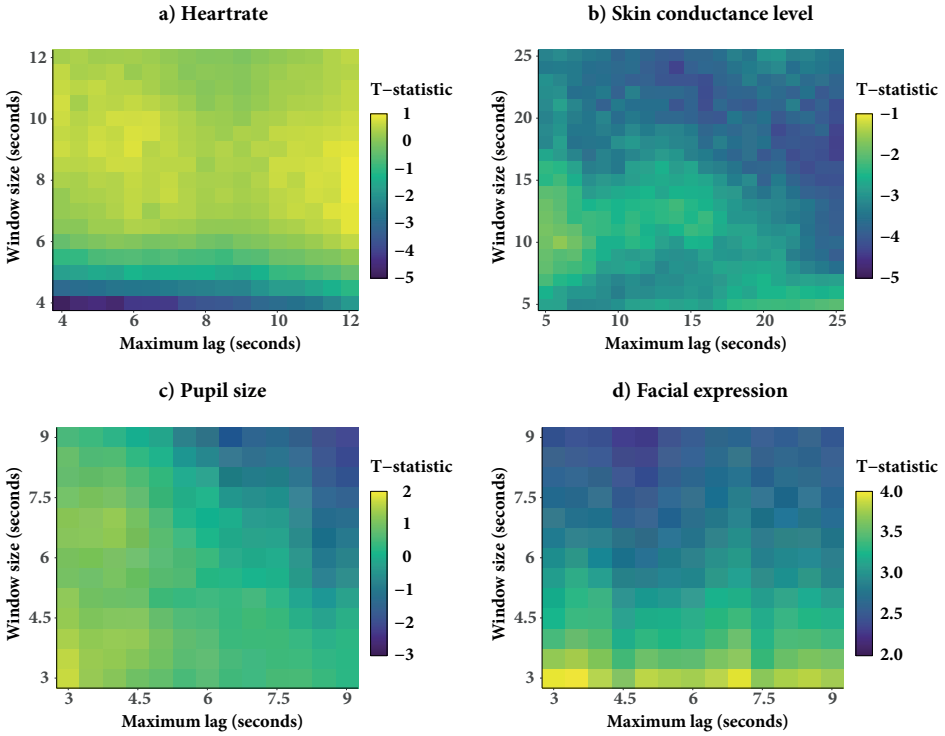

Figure 7. Distribution of t-statistics of the comparison between storytelling and baseline of all parameter configurations for each physiological measure. The color coding runs from the lowest (blue) to the highest (yellow) t-statistic. A positive t-statistic indicates that the level of synchrony was higher during storytelling than during baseline. Analysis was based on data from both baseline measures and the first and third stories. Notice that the scaling of the axes and the color coding are adjusted to each physiological measure to increase comparability between parameters. Also, the highest t-statistic was not always the highest absolute value with the latter value being discussed in the result section. However, the general idea of greater (absolute) t-statistics indicating better discrimination between the two conditions remains.

Discussion

The phenomenon that people synchronize each other's emotional expressions and physiological states has intrigued researchers in many different disciplines. Studying this phenomenon comes with the challenge of statistically capturing the dynamic nature of a social interaction. Over the years, several methods have been developed that address this dynamic to different degrees and in different ways. One such method is the Windowed Cross-Correlation analysis (Boker et al., 2002). It accounts for changes in the strength of synchrony throughout an interaction and in the different paces in which people respond. The method requires researchers to specify parameters that allow us to tailor the method to the signal of interest. Albeit increasing the method's flexibility, there is a lack of guidelines on which parameters to use for which signal, which can have an impact on the outcome of the analysis. The aim of the current study was to statistically determine the most suitable parameter settings applied to four different physiological measures: heartrate, skin conductance level, pupil size, and activity of the zygomaticus major muscle (associated with smiling). To that end, we systematically investigated the influence of a range of parameter configurations on two criteria: i) the ability to distinguish synchrony from pseudosynchrony, and ii) the sensitivity to detect change in synchrony (i.e., distinguish two within-subject conditions).

Regarding the first criterion, the results revealed that a wide range of parameter configurations could distinguish between the original dyads and dyads that participated in the study, but never engaged in an actual interaction (i.e., surrogate dyads). Additionally, a general pattern across all physiological measures emerged: the smaller the window size, the better the discriminative ability tear apart the original dyads from the surrogate dyads. In contrast, if the window size became too large, the estimated level in true dyads dropped to such an extent that it became lower than the synchrony level estimated in the surrogate dyads. With respect to the second parameter, the maximum lag was generally larger than the corresponding window size. How much larger differed between physiological measures: the optimal maximum lag was two, four, and two to three times the window size for heartrate, skin conductance level, pupil size and facial expressions, respectively.

Regarding the second criterion, that is, the sensitivity to detect change in synchrony, the results were less clear cut. Here, we compared two baseline measures where people looked at each other in silent with periods where participants engaged in storytelling. For heartrate and facial expressions, the general pattern was visible with better discriminative ability between storytelling and baseline with smaller window sizes. For facial expressions, this pattern was, however, weak at best. Interestingly, differences across measures emerged of whether synchrony levels were higher during storytelling or baseline. (Almost) all parameter configurations for the heartrate and skin conductance level measures indicated higher levels of synchrony during baseline. For pupil size, both patterns emerged with small window sizes and maximum lags showing more synchrony during storytelling, whereas large window sizes and maximum lags revealed more synchrony during baseline. For facial expressions, storytelling showed higher levels of synchrony for all parameter configurations. Other than these differences between measures, the range of t-statistics within each measure was considerably smaller than for the surrogate data analysis, suggesting less sensitivity to parameter choice. In the following, we will discuss our findings

in depth and integrate them with theoretical considerations. In Table 2, we summarize the global recommendations on determining the parameter configurations. We hope that these guidelines provide researchers with information that assist them to make well-informed decisions about the optimal parameters for their WCC analysis.

Table 2
Summary of global recommendations per parameter of the WCC analysis

Parameter	Recommendations
Window size	• Lower bound: large enough to capture meaningful information and variance within the signal of interest • Upper bound: the response duration of the signal of interest; assumption of stationarity is met
Maximum lag	• Lower bound: at least as long as the window size • Upper bound: at most twice as long as the window size
Window and lag increment	• Lower bound: 1 datapoint • Upper bound: same as the window size / maximum lag • Balance computational time and resolution: 1•5% of the window size / maximum lag

Synchrony versus pseudosynchrony–Window size

We observed that a wide range of window sizes was able to distinguish between synchrony and pseudosynchrony. However, in general smaller window sizes performed better. However, if the window size became too large, synchrony levels dropped to the extent that the levels became lower for the true dyads than the surrogate dyads. How can this general pattern across measures be explained? To understand it, let us quickly recapture the purpose of the surrogate data analysis. As introduced above, the aim is to destroy any synchrony that is the result of interpersonal processes while preserving all other statistical properties by generating new dyads that participated in the study, but never actually interacted. This way we *know* that the null hypothesis that there is no synchrony between participants is true. As the null hypothesis will be true independent of the parameter configurations, the distribution of cross-correlations stays constant across all parameters. In contrast, for the original dyads, synchrony does emerge, which we expected based on prior research. During a dynamic interaction, there are moments when dyads are in sync, but also out of sync (Boker & Rotondo, 2002). If the window size becomes too large, both moments of synchrony and "anti-synchrony" are likely to be included into one window segment, substantially reducing the strength of synchrony. This causes a drop in overall synchrony that can be lower than in the surrogate dyads with no synchrony at all (i.e., no synchrony and no "anti-synchrony"). On the other hand, decreasing the window size decreases the variance within a window causing the overall synchrony level to increase. Specifically, as seen in Equation 1, the cross-correlation is calculated by dividing the distance between each datapoint and the mean of the window segment by its standard deviation. The smaller the window size, the less change for variation to

occur within a window (i.e., the smaller the standard deviation), which causes the correlation to increase. Thus, while the distribution of correlation estimates stays constant for the surrogate dyads, the estimates for the original dyads increase with smaller window sizes. Consequently, the distance between the mean of these two groups becomes increasingly larger, causing the general pattern we see across the physiological measures. This pattern is therefore an intrinsic characteristic of the way the cross-correlation is estimated applying to all types of time series.

The question then arises whether steadily decreasing the window size will also steadily increase the ability to distinguish synchrony from pseudosynchrony. The short answer is no. Imagine the extreme case, where the window size consists of two datapoints. These two datapoints hold very little information and would only allow possible correlation values of -1 and 1. This reduces the sensitivity for measuring synchrony and therefore for distinguishing synchrony from pseudosynchrony. Consequently, somewhere between a window size containing two datapoints and the smallest window sizes we examined, there will be a turning point, where the two types of dyads will become distinguishable.

Although statistically possible, making the window size as small as possible (but above the turning point) is not advisable for two reasons: (1) a sufficient number of data points are needed to reliably estimate correlation coefficients (Schönbrodt & Perugini, 2013), and (2) the window should capture a meaningful response. As outlined earlier, in order to reliably estimate a correlation coefficient, a recent study showed that 65 to 250 datapoints are necessary depending on the strength of the correlation. With a sampling rate of 20Hz across all measures, we therefore used a window size of at least 3 seconds (60 datapoints). If researchers want to further decrease the window size, they should increase their sampling rate accordingly. With that said, a window size must include responses constricted by a meaningful upper and lower bound.

In the current study, we narrowed the possible values for the window size parameter by showing a range of parameters that qualify as potentially suitable parameters. How can researchers choose between these options? To answer this question, let us go back to the aim of cutting the time series into segments in the first place, namely, reducing the non-stationarity in the signals. A stationary signal has constant statistical properties with, among others, a constant mean and standard deviation within that signal. In a dynamic interaction, the strength of synchrony (our statistical property of interest) will vary between moments of strong and weak synchrony. The window size needs to be small enough such that the synchrony level stays constant within that window. Determining how small the window must be, depends on the nature of the signal and is contained by an upper and lower bound.

Imagine smiles of two interaction partners are coded during a conversation such that a person either smiles or not. If the two participants smile at the same time for the same duration, there will be perfect synchrony between them for the entire duration of the two smiles (given an appropriate correlation measure for categorical variables). In this case, the window size could be as large as the duration of the smile because the level of synchrony is constant during that interval. However, if the smiling response occurs in a real conversation and is measured continuously reflected in the activity of the zygomaticus major (as in the current study), there are variations in latency, magnitude and duration of the smiles within and between individuals. In this case, the level of synchrony is likely to change even within the window that would have been categorized as

a "smile" in the artificial categorical scenario just described. For example, one person might show a long, pronounced smile, while the other person might smile later and and for a shorter length of time. Then the synchrony would only occur during the short time where both people smile simultaneously. Therefore, the window size should be smaller than the duration of a "typical" smile to capture these variations. More specifically, we recommend a window size that is at most half the response duration, such that at least two estimates of the level of synchrony will be computed for that response capturing changes in synchrony that are twice the speed of the overall response.

Other than the upper bound for window size being smaller than the response duration of interest, there is a lower bound as well. In particular, the window size should be large enough to capture meaningful variations within a response. For example, if the signal of interest is skin conductance level, a window size of 1 second would contain straight lines in most windows. This produces extreme cross-correlations without capturing meaningful changes in the signal. On the other hand, applying the same window size to facial expressions might be considered a medium to large window given that a smile has been shown to last 500ms to 4 seconds (Frank, Ekman, & Friesen, 1993). Both the upper and lower bound therefore determine the potential values for the window size.

When talking about "the duration of a response" we realize that this can be difficult to define as physiological measures show great variations within and between individuals. In the section "physiological boundaries" below, we provide an overview of the "typical" temporal characteristics of each physiological measure realizing that this overview is far from being exhaustive. It is beyond the scope of the current paper to provide concrete guidelines for this matter and it is up to the researcher to decide on which responses she is interested in. As the most suitable (range of) window size(s) likely differs across situations and conditions, choosing a window size should be seen as a hypothesis that is tested, namely, that responses synchronize that are equal to or longer than the window size chosen. Although faster responses are still included in the window segments, they are likely to be averaged out and changes in the faster responses will be reduced.

If the researcher has no strong a priori hypotheses, multiple window sizes can be tested across a range of possible values taking a data-driven bottom-up approach to determine the best parameter choice. Obviously, it is not realistic that researchers perform such elaborated analyses as in the current study, however, comparing two to three potential values can shed light on the rate at which synchrony occurs in a particular context. Of course, it is unlikely that people will synchronize on one specific response duration only, so one would expect more similar results for window sizes closer together. However, referring to "skin conductance synchrony" based on one parameter setting is likely an overgeneralization and needs more detailed investigation.

To conclude, the results of the current study indicate that a range of window sizes is able to detect synchrony that occurs as a result of interpersonal processes with a preference for shorter window sizes. From a theoretical perspective, the range of potential window sizes is contained by (i) an upper bound defined by the length of the duration of the responses under investigation and (ii) a lower bound defined by sufficient variation within the window. Rather than searching for that one most suitable parameter for each physiological measure, choosing a window size should be seen as a hypothesis being tested. Importantly, researchers need to be specific about what aspect of a signal they investigate which should be clearly stated in both their hypothesis and conclusions.

Synchrony versus pseudosynchrony: maximum lag

Our results revealed that the maximum lag was less influential than the window size, yet not trivial. In contrast to the window size, the optimal maximum lag differed between the physiological measures. For heartrate, pupil size, and facial expression, the optimal maximum lag was around 5–10 seconds. Skin conductance level deviated from the other measures with the optimal parameter being around 20–25 seconds. This is consistent with the fact that skin conductance level is a considerably slower signal compared to the others. However, it contrasts the finding reported by Robinson and colleagues (1982) who showed that synchrony in skin conductance response within, but not outside the range of 7 sec was associated with the empathetic relationship between therapists and clients. Such discrepancy can be explained by the fact that while these authors concentrated on the phasic response, we have focused on the tonic, slower responses. This underscores the importance of being specific about what aspect of a signal the researcher is interested in and shows again the importance of the theoretical consideration for choosing the parameter configurations for the WCC analysis. In the following, we aim to provide the reader with a sense of how the maximum lag influences the analysis.

Essentially, the maximum lag indicates how far responses between participants can lie apart that can still be considered a response to one another. Thus, similar to choosing the window size, the maximum lag considerably depends on the interest of the researcher. Given their link, it seems reasonable to choose the maximum lag in relation to the window size. In line with our findings, we recommend using a maximum lag that is equal to or twice the size of the window. For simplicity, let us assume that stationarity is met for the length of a full response, all response cycles have the same length and the window segments start at the beginning of a new response. If the maximum lag is the same length as the window size, the window segment of Person A will be shifted away from the segment of Person B (and vice versa) until the two segments succeed one another with no overlap in time. When Person A, now later in time, shows a response, then Person B reacts *right after* the response of Person A. Thus, over the range of all considered lags, synchrony can happen between people being in sync (lag = 0) and people responding to each other in direct succession. In a similar vein, setting the maximum lag to twice the window size means that there can be up to a full response duration between the responses of the interacting individuals. For example, imagine the measure of interest is facial activity and the window size is 2 seconds. If the maximum lag is 4 seconds, then two smiles that occur simultaneously up to the situation that they are 4 seconds apart from each other are considered synchronized responses. The latter situation seems still reasonable in the context of a real conversation, yet on the upper limit. Therefore, expanding the maximum lag to 6 seconds likely increases the chance of linking two unrelated events to one another. The decision to set the maximum lag equal to or twice the window size depends on the researcher's preference of what she considers reasonable in the context of interest. In a controlled environment with straightforward, stereotypical displays of emotions, a person should react rapidly and a smaller maximum lag might be sufficient. However, in a natural interaction where ambiguous expressions and verbal conversations require more elaborated processing, a response might take longer and therefore a larger maximum lag might be appropriate. In addition, the latency of a response itself is important, especially in relation to the response duration. For exam-

ple, if a response is expected to be initiated rapidly, but last relatively long, a small maximum lag is sufficient. However, if the latency of a response is long and the duration of the response short, then a longer maximum lag is required. In sum, as a general rule of thumb we recommend a maximum lag of at least equal to and at most twice the size of the window size.

We would like to point out that the results considerably deviated for the skin conductance level. While the three other measures showed the largest discrimination between synchrony and pseudosynchrony for a maximum lag that was about twice the window size, for skin conductance level it was four times (around 20–25 seconds). As described above, this is consistent with the fact that skin conductance level is a substantially slower response compared to the other signals. One might therefore argue that the associated window size of 5 seconds might be too small capturing mostly responses with little meaningful variation. Increasing the window size might therefore be advisable, which then align with our recommendation of choosing a maximum lag that is at most twice the window size. In conclusion, our findings revealed that from a statistical point of view, the maximum lag is less influential than the window size. Nevertheless, this does not safeguard the researcher from using any parameter and tailoring it to the nature of the signal of interest is essential. Here, we have provided more information about the meaning of the maximum lag and recommended to specify the maximum lag equal to or twice the window size.

Window and lag increments

In the current study, we have adjusted the increments such that the windows and lags moved by 10% of the window size and maximum lag, respectively. This was a choice of practicality, reducing the computational time in light of the huge amount of analyses run while keeping the resolution sufficiently high. As already mentioned at the beginning of the paper, both parameters determine the resolution of the changes occurring between window segments and lags. Ideally, the increments should be as small as possible (i.e., 1 data point). However, the increments heavily influence the computational time which is why researchers might want to increase these parameters. Nevertheless, the increments should never be set higher than the window size and maximum lag themselves. In case the lag increment is equal to the maximum lag, three situations are analyzed: people responding in sync (lag = 0), Person A responds to Person B with a delay of the maximum lag, and Person B responds to Person A with a delay of the maximum lag. For the window size, two adjacent segments would not overlap. If the increment would be greater than the window size, there would be a gap between two adjacent segments. This is problematic because moments of synchrony occurring during that gap are missed. Generally, unless researchers are specifically interested in one particular time lag, we recommend keeping the increment small in relation to the window size. Using the 10%-rule of thumb was fine for the current study, however, we needed to account for an enormous amount of analyses. We believe that reducing the percentage to 1 to 5% offers a good balance between analysis sensitivity and computational time.

Change in synchrony

Besides the ability to detect synchrony, we also investigated the effect of the parameter configurations on the sensitivity to detect change in synchrony. The results were less clear-cut here. While for heartrate and facial expression synchrony, the general pattern of smaller window sizes increasing the discrimination ability was (weakly) apparent, it was not observed for skin conductance level and pupil size. Additionally, the primary and replication analyses sometimes showed large deviations. For example, for the skin conductance level, the greatest differences between conditions was apparent for a window size of 5 seconds in the primary analysis and 18 seconds in the replication analysis. On top of that, there were differences between measures and parameters in whether synchrony levels were higher during storytelling or baseline. In particular, for heartrate and skin conductance synchrony, (almost) all parameter configurations suggested higher levels of synchrony during baseline, whereas the reverse was evident for facial expressions. Such discrepancy might be explained by the function of the signal. Facial expressions are displayed for communicative purposes which is particularly important during storytelling where people react to one another more than during silent moments of eye-contact during baseline. While arousal levels also play a crucial role during conversations, during the baseline measure people could concentrate on each other nonverbally and were not "disturbed" by engaging in conversations, overall leading to higher synchrony during baseline. On top of that, the baseline condition consisted of two baselines measures with the second being preceded by the breathing exercise where participants were instructed to synchronize their breathing. This might have influenced the second baseline measure leading to higher overall synchrony levels. In general, given the lack of clear patterns and inconsistencies between the primary and replication analyses, we refrain from giving recommendations for parameter configurations based on these results.

The inconclusiveness of the results might be attributed to two potential explanations: (1) the difference between the two conditions was negligible and the sensitivity to detect such small differences was barely affected by the parameters; (2) there were differences between the two conditions, but the method was not sensitive to detect them. In support of the first explanation, in two previous studies, we have used parameters included in the current analysis with which we were able to detect differences in within-subject conditions and could link it to interpersonal outcomes (Behrens et al., 2019; Prochazkova, Sjak-Shie, Behrens, Lindh, & Kret, 2019). The method therefore has been shown to be sensitive in other contexts. However, future studies are needed to address this question using either simulated data or more extreme conditions where the difference is more pronounced and possible differences between parameters are more likely to show.

Physiological boundaries

Every physiological measure has its temporal characteristics and we will give a short overview for each of the four measures considered in the current study. First, the time course of heartrate is controlled by several physiological processes that can operate to varying degrees depending on the situation and psychological process of interest. Generally, parasympathetic nervous system activity slows the heartrate down, while sympathetic activity increases heartrate. While parasympathetic activity is associated with fast changes in heartrate and is predominantly related to breathing (changes within millisecond to second range), sympathetic activity takes more time to show and is attributed to changes in arousal levels (changes within second range) (Berntson, Cacioppo, & Quigley, 1991). The pace of the heart can change on a beat-by-beat interval and the peak of heartrate acceleration has been shown to occur within the first 4 seconds (Critchley et al., 2005). The duration of a response to an external event (e.g., stimulus presentation) usually takes around 5–8 seconds, although full recovery from stressful events can take several minutes (Berntson et al., 1991; Bradley, Codispoti, Cuthbert, & Lang, 2001; Bradley et al., 2008; McAssey et al., 2013).

Skin conductance measures are indications of arousal resulting from sympathetic nervous system activity and are divided into tonic (skin conductance level) and phasic (skin conductance response) components. The tonic activity consists of gradual changes over time that vary considerably between and within individuals. It decreases during rest and increases more quickly in response to new events (Dawson et al., 2000). The phasic activity, the high-frequency component of the skin conductance measure, is faster than the tonic response and reflects responses directly linked to an external or internal event. The latency of a response is usually between 1–3 seconds and the time to reach the peak amplitude takes between 1–4 seconds. The duration of a full response from stimulus presentation to 50% recovery of the amplitude after the response peak varies between 4 to 16 seconds (Dawson et al., 2000). This is consistent with a power spectral analysis showing that the sympathetic activity is reflected in frequencies between .045–.25 Hz, corresponding to response durations of 22 and 4 seconds, respectively (Posada-Quintero et al., 2016).

Changes in pupil size can result from both parasympathetic and sympathetic activity. Specifically, pupil constriction is mainly controlled by parasympathetic activity, whereas pupil dilation is an indication of sympathetic activity. Pupil size changes in response to light are rapid showing a constriction response 200ms after turning on the light (Mathôt, 2018). Pupil size changes in response to psychosensory processes are slower and vary with, among others, mental effort and saliency of the stimulus (for a review, see Beatty & Lucero-Wagoner, 2000). The typical response is characterized by an initial constriction in response to the stimulus and subsequently, a more pronounced dilation of the pupil with a peak after 2 to 3 seconds and a total response duration of 4 to 6 seconds (Bradley et al., 2008; Kret et al., 2015; Oliva & Anikin, 2018).

Facial expressions consist of changes in facial muscles such as the zygomaticus major, associated with smiling, and the corrugator supercilii, associated with frowning. The duration of a facial response depends on whether researchers investigate subtle, rapid changes or full-blown smiles in a natural conversation. For example, a facial response can occur as fast as 200–300ms in response to stereotypical, controlled stimuli (Achaibou, Pourtois, Schwartz, & Vuilleumier,

2008). In a more natural setting, Frank, Ekman, and Friesen (1993) showed that a Duchenne smile of enjoyment lasts between 500ms to 4 seconds. Accordingly, response windows used in previous studies greatly differ ranging from 1.4 – 5 seconds after stimulus onset showing static images (Achaibou et al., 2008; Drimalla, Landwehr, Hess, & Dziobek, 2019; Lang, Greenwald, Bradley, & Hamm, 1993), to 15 second intervals investigating facial activity in real-life interactions (Hess & Bourgeois, 2010). This section gives a brief glimpse into the temporal characteristics of the physiological measures we have focused on in this paper. However, we would like to emphasize that this overview is far from being exhaustive and researchers need more elaborated knowledge to make well-informed decisions about the signal of interest.

Limitations

There are a few limitations that we would like to point out. First, in the current study we concentrated on the window size and maximum lag parameters, while setting the window and lag increments to 10%. A systematic investigation of the effect of changing these parameters is needed. As mentioned earlier, estimations of the level of synchrony will stabilize with smaller increments such that decreasing the increments even further will add little information at the cost of extra computational time. Although we propose to set the increments to 1– 5% of the window size and maximum lag, this suggestion is not based on statistical analyses and future research is needed to determine the optimal balance between sensitivity and computational time. Second, the general guidelines we propose in Table 2 may not be generalizable to other measures of synchrony and may not be applicable to other biological time series. Researchers should therefore be careful with making any inferences about other statistical analyses and time series than used in the current study. Third, all data come from a single study and is subject to method variation. To reduce such variation, we ran all analyses twice with different data from the same study. However, this does not address method variations that are the result of the study itself and future studies should replicate our findings in a different dataset. Finally, we changed the original plan for the comparison of time intervals as outlined in Appendix D1. A more tailored study design may have observed more specific results, in particular with the regard to the sensitivity to detect change in synchrony.

Future directions and conclusions

The most important lesson the current study teaches us is that researchers need to be precise in what they (aim to) investigate as defined by the parameters specified in the analyses. In the current study, dyads synchronized on a range of response windows. However, this might not always be true, especially, if the aim is to link it to specific psychological processes that might be influenced by only particular physiological processes. Future studies are therefore needed that make more refined distinctions of which components of a particular physiological signal is involved in the process of interest and how the different components interact. This will facilitate making well-informed decisions about the response windows and shed more light on the biological underpinnings of psychological processes.

Before making well-informed decisions on the parameter configurations *within* a particular method, it is important to realize what the differences are *between* methods. WCC analysis is one of many possible methods and each method has its strengths and weaknesses. While one method might be appropriate for one, it might not be for another depending on, among others, the type of data (e.g., continuous or categorical measures) and the measure of interest (e.g., strengths versus frequency of synchrony; global versus time-sensitive measure of synchrony) (Gates & Liu, 2016; Schoenherr et al., 2018). For example, we chose to treat facial muscle activity as a continuous measure. However, researchers might also be interested in investigating concrete events of, for example, smiling and its synchrony in a conversation. Here, the analysis developed by Altmann (2011) might be appropriate where time series are first categorized into intervals of synchrony and intervals of no synchrony before measures of the strength and frequency of synchrony are computed. Despite using the same data, the outcomes can be somewhat different as demonstrated by Schoenherr and her colleagues (2018). Performing an exploratory factor analysis on seven linear time series analyses and different outcome variables (among others the WCC analysis), they reported that all these methods measure the overall phenomenon of synchrony, but could be categorized into three correlated, yet distinct facets of synchrony: the strength of synchrony of the total interaction, the strength of synchrony during synchronization intervals, and the frequency of synchrony (Schoenherr et al., 2018). The WCC analysis as performed in the current study reflects the first facet. Researchers should therefore refine which facet of synchrony they are interested in and choose the appropriate methods accordingly.

The aim of the current study was to optimize the parameters for the WCC analysis from a statistical point of view. The initial idea was to provide researchers with concrete guidelines on which specific parameters would be most appropriate for the four physiological measures. However, the results show that when the aim is to detect synchrony, the parameters follow a general pattern that is not specific to the signal of interest, but rather a result of intrinsic characteristics of how the cross-correlation is calculated. That does not mean that the parameters should not be tailored to the signal of interest. Instead, theoretical considerations should be integrated with the findings observed in the current study. Here, there is no one-fits-all solution, which might not be surprising given that we aim to capture a highly complex process. The current study narrows down the range of possible parameters and we provide guidelines on how to tailor the parameters further to the interest of the researcher. Being specific and transparent about these choices will increase the comparability across studies and add more and more pieces to the puzzle of understanding the phenomenon of synchrony.