



# Quantification of nonverbal synchrony using linear time series analysis methods: Lack of convergent validity and evidence for facets of synchrony

Désirée Schoenherr<sup>1</sup> · Jane Paulick<sup>2</sup> · Susanne Worrack<sup>1</sup> · Bernhard M. Strauss<sup>1</sup> · Julian A. Rubel<sup>2</sup> · Brian Schwartz<sup>2</sup> · Anne-Katharina Deisenhofer<sup>2</sup> · Wolfgang Lutz<sup>2</sup> · Ulrich Stangier<sup>3</sup> · Uwe Altmann<sup>1</sup>

Published online: 8 October 2018  
© Psychonomic Society, Inc. 2018

## Abstract

*Nonverbal synchrony* describes coordination of the nonverbal behavior of two interacting partners. Additionally, it seems to be important in human interactions, such as during psychotherapy. Currently, there are several options for the automated determination of synchrony based on linear time series analysis methods (TSAMs). However, investigations into whether the different methods measure the same construct have been missing. In this study,  $N = 84$  patient–therapist dyads were videotaped during psychotherapy sessions. Motion energy analysis was used to assess body movements. We applied seven different TSAMs and recorded multiple output scores (average synchrony, maximum synchrony, and frequency of synchrony; in total,  $N = 16$  scores). Convergent validity was examined using correlations of the output scores and exploratory factor analysis. Additionally, two criterion-based validations were conducted: investigations of concordant validity with a more generalized nonlinear method, and of the predictive validity of the synchrony scores for improvement in interpersonal problems at the end of therapy. We found that the synchrony measures only partially correlated with each other. The factor analysis did not support a common-factor model. A three-factor model with a second-order synchrony variable showed the best fit for eight of the selected synchrony scores. Only some synchrony scores were able to predict improvement at the end of therapy. We concluded that the considered TSAMs do not measure the same synchrony construct, but different facets of synchrony: the strength of synchrony of the total interaction, the strength of synchrony during synchronization intervals, and the frequency of synchrony.

**Keywords** Nonverbal behavior · Movement synchrony · Motion energy analysis · Time series analysis · Convergent validity

Currently, body movements can be assessed fully automatically and with high time resolution (e.g., 25 times per second) using either motion-tracking, motion capture devices, or video-based algorithms (Delaherche et al., 2012). *Motion energy analysis* (MEA) is a method that quantifies the intensity of videotaped movements frame-wise (Grammer, Honda, Juetter, & Schmitt, 1999). By determining a region of interest (ROI) for each of

two videotaped individuals (e.g., a patient and therapist during a psychotherapy session), two time series can be generated displaying the time course of the individuals' body movements. This technique has several advantages: (1) it is less time-consuming than collecting human ratings; (2) it is highly objective, reliable, and valid; and (3) in comparison to motion capture devices, no high-resolution camera equipment is necessary, and no sensors are attached to the patient's body (Altmann, 2010; Ramseyer & Tschacher, 2011). Therefore, during the past few years, the use of MEA has become enormously widespread. In behavioral and social science, MEA has been used to assess movements in mother–child interactions (Watanabe, 1983, 1987), child friendships (Altmann, 2011, 2013), and courtship behavior (Grammer et al., 1999; Grammer, Kruck, & Magnusson, 1998); for the comparison of nonverbal behavior in different types of interactions, such as tasks that elicit truthful, deceptive, argumentative, cooperative, competitive, or during fun tasks (Allsop, Vaitkus, Marie, & Miles, 2016; Altmann, 2011, 2013; Duran & Fusaroli, 2017; Paxton & Dale, 2013a;

---

✉ Désirée Schoenherr  
desiree.schoenherr@med.uni-jena.de

<sup>1</sup> Institute of Psychosocial Medicine and Psychotherapy, University Hospital Jena, Jena, Germany

<sup>2</sup> Department of Clinical Psychology and Psychotherapy, Trier University, Trier, Germany

<sup>3</sup> Department of Clinical Psychology and Psychotherapy, Goethe University Frankfurt/Main, Frankfurt/Main, Germany

Tschacher, Rees, & Ramseyer, 2014); in ecological psychology (Davis, Kay, Kondepudi, & Dixon, 2016); for the diagnosis of psychological disorders (Dean, Samson, Newberry, & Mittal, 2018; Dutschke et al., 2018; Kupper, Ramseyer, Hoffmann, Kalbermatten, & Tschacher, 2010; Kupper, Ramseyer, Hoffmann, & Tschacher, 2015); and in psychotherapy (Galbusera, Finn, & Fuchs, 2018; Paulick et al., 2018; Ramseyer, 2011, 2013; Ramseyer & Tschacher, 2011, 2016).

After generating motion energy time series, the influence of several variables (e.g., psychopathologies, attachment styles) on movement behavior can be investigated. Additionally, it has been demonstrated that the perception of a behavior shown by one interacting partner increases the probability of the other interacting partner engaging in that behavior (Chartrand & Bargh, 1999). During these sequences, an observer has the impression that the behavior of both interactors is synchronized, aligned, coordinated, co-regulated, or timed (Altmann, 2013; Bernieri & Rosenthal, 1991). Examples of such phenomena during interpersonal interactions are the imitation of facial expressions or gestures, posture mirroring, synchronous movements, or the convergence of voice parameters. Bernieri and Rosenthal (1991) pointed out that these phenomena are performed in a nonrandom manner, either following specific patterns or showing formal and temporal synchrony.

However, multiple terms and conceptualizations of these synchronization phenomena can be found, which are synonymous, partly overlapping or distinct from one another (Altmann, 2013; Feldman, 2007; Harrist & Waugh, 2002; Paxton, 2015). Paxton attempted to disambiguate different terms of synchronization phenomena (e.g., coordination, alignment, mimicry, imitation, synchrony, etc.) by natural language processing, showing that terms are rather separated by the underlying research question and study area than the phenomena under investigation. Terms are therefore rather based on different research areas with unique terminological trends (Paxton, 2015). Thus, a psychologist and a linguist may describe equal synchronization phenomena with different words. However, *synchrony* was identified as a suitable superordinate term to describe different conceptualizations such as facial imitation, movement synchrony or speech convergence. Due to the fact that the behaviors are nonverbal, we refer to these phenomena with the term nonverbal synchrony.

One possibility of grouping the different synchronization phenomena is by using a time dimension and the level of measurement as proposed by Altmann (2013). With respect to the time level, three different facets can be differentiated: (1) perfectly synchronous, simultaneous behavior or matching (*without* a time lag); (2) synchronous behavior with a time delay, echoing, alignment, imitation, or mimicry and (*with* a time lag); and (3) convergence, increasing similarity, and adaptation (*increasing similarity* over time). Referring to the level of measurement, data can be categorical, such as facial expressions (e.g., smile vs. no smile), or metric such as voice

pitch or movement intensity (Altmann, 2013). An example of the combination of both dimensions would be the examination of exactly simultaneous facial expressions (matching of categorical data). In addition to these different groups of synchrony, different algorithms exist for measuring the construct. Nevertheless, all algorithms are used to measure the construct synchrony; systematic investigations of their validity and concordance are missing. However, these investigations are needed to make judgments about the comparability of study results.

Only one study compared human ratings of nonverbal synchrony and nonverbal synchrony obtained by cross-lagged correlation (Paxton & Dale, 2013b). The authors compared the results of the cross-lagged correlation of human second-by-second coding of the movements of two persons with the results of the cross-lagged correlation of a frame-differencing technique. This study provided evidence that movements rated by humans and by the algorithms led to comparable synchrony results. However, more research will be needed to disentangle the different nonverbal synchrony constructs and to examine which time series analysis methods (TSAMs) lead to comparable results.

Since rater reliability in human synchrony ratings is rather weak, or many raters are needed to obtain high values (Bernieri, 1988), we focused on objective and reliable algorithms. As illustrated above, the construct to which synchrony obtained by algorithms is compared is crucial. To date, no true value/construct of synchrony has been defined. Therefore, in line with Cronbach and Meehl (1955), who stated that investigations of construct validity (convergent and discriminant validity) are especially relevant if no direct criterion or “universe of content” is available, we examined the *convergent* validity of different algorithms. For this aim, we first describe different linear-based algorithms that assess nonverbal synchrony, and then apply these to an identical dataset containing motion energy time series. Furthermore, we tested whether all measures load on a common factor. If not, the adequate factor solution would be determined. In addition, we conducted two further validations: First, we compared linear-based TSAMs to a more generalized nonlinear approach (cross-recurrence quantification analysis). Second, the predictive validity of the TSAM output scores was examined by inspecting the synchrony-outcome association using data from psychotherapy sessions.

However, the aim of our article was not to identify the best algorithm to assess synchrony, but rather to test whether different algorithms are equivalent operationalizations of the construct of nonverbal synchrony.

## Systematization of different linear-based algorithms assessing the nonverbal synchrony of two interacting persons

The process of co-regulation between two individuals can be understood as dynamic (Altmann, 2013; Boker, Rotondo, Xu,

& King, 2002; Fogel, 1993). In our study, nonverbal synchrony was operationalized as the degree of association between the nonverbal behaviors of two interacting persons. Approaches assuming a linear relationship between both interacting partners (correlations or regressions) are very common. Thereby, correlations or regressions are computed for the entire time series or windows of the time series. Subsequently, the obtained values are usually aggregated to an output score in order to quantify this association (also called “co-aggregation measures,” see Coco & Dale, 2014). Other approaches investigate the degree of association by using recurrence plots that show recurrence points if two systems are similar to each other with respect to their phase trajectories. Cross-recurrence quantification (CRQA) analysis can be used to analyze these points (for an overview, see Delaherche et al., 2012). This method is widely applied to analyze states of one time series and how the other revisits these states, which is especially relevant for the investigation of temporal patterns of nonverbal synchrony during an interactional process. Thereby, this method does not make the assumption of a linear relationship between two time series (Coco & Dale, 2014; Marwan & Kurths, 2002). Comparing linear methods with cross-recurrence analysis showed that a systematic association between both methods is sometimes missing (Wiltshire, 2015). However, the results from CRQA can be regarded as a generalization of cross-correlation methods (Marwan, Romano, Thiel, & Kurths, 2007). Spectral methods provide further options to analyze nonverbal synchrony (for more details, see Delaherche et al., 2012).

In the present article, we focus on methods that assume a linear relationship between time series trajectories, because (1) these methods are widely used in interactional research (Altmann, 2013; Kupper et al., 2015; Nelson, Grahe, Ramseyer, & Serier, 2014; Paulick et al., 2018; Ramseyer & Tschacher, 2011) and (2) including other approaches would result in a reduction of comparability of the different methods, since the statistical assumptions of the methods are very distinctive. As a validation, we will contrast these methods with one nonlinear method. We chose CRQA because this method is widely used in the context of interactional analyses (e.g., Shockley, 2005; Shockley, Richardson, & Dale, 2009).

Linear TSAMs can be differentiated by (a) assumptions about time lag (matching vs. echoing) and (b) the length of time series windows (global vs. local TSAMs).

**Assumptions about time lag** Two operationalizations of nonverbal synchrony can be distinguished: (a1) using no time delay (termed “*matching*”; see above) or (a2) using a specific time delay (termed “*echoing*”). Most algorithms that measure echoing also include matching. Currently, the selection of the maximum appropriate time lag is largely left up to the researcher. It characterizes the greatest interval separating the behavior of two persons, which is still considered to be

connected. Investigation of the coordination of skin conductance level showed a meaningful nonverbal synchrony with a maximum time lag of 7 s (Robinson, Herman, & Kaplan, 1982). Additionally, virtual agents that mimic persons with a time delay of exactly 4 s were rated more positively than nonmimickers (Bailenson & Yee, 2005). In accordance with that, Bilakhia, Petridis, Nijholt, and Pantic (2015) recommended a time lag of 0.04 to 4 s, based on the empirical investigation of episodes showing motor mimicry. Altmann (2011) used a maximum time lag of 2.5 s, whereas Ramseyer and Tschacher (2011) used a time lag of 5 s. Another study by Louwerse, Dale, Bard, and Jeuniaux (2012) showed that mimic expressions and head movements have a short time lag of approximately 1.5 s. However, there have also been attempts to evaluate the chosen time-lag based on the comparison with shuffled data, implying that the choice of best lag could be empirically determined.

**Length of time series window** Regarding the second dimension—the length of the assessed time series window—two types of linear TSAMs can be differentiated: (b1) global and (b2) local measures (Altmann, 2013). Two global methods are cross-lagged correlation (CLC) and cross-lagged regression (CLR; Gottman & Ringland, 1981; Paxton & Dale, 2013b). Global TSAMs calculate Pearson product-moment correlations or regressions using the two entire time series. The entire time series are incorporated with equal or different starting points. The distance between the starting points is referred to as the time lag (see above). Global methods have the advantage of being less time-consuming in terms of computational costs, as they are less complex. However, when using global methods, the assumption of global stationarity is made. Stationarity means that given the respective time series, the mean and variances stay constant over time. Additionally, by using global methods, it is assumed that person A influences person B or vice versa for the complete time course; that is, there is no changing interdependence between the interacting partners. This assumption is often violated in naturalistic contexts such as human communication (Boker et al., 2002). As a result, local TSAMs were developed.

Local TSAMs analyze the entire time series window-wise, conducting Pearson product-moment correlations (or regressions) between segments/windows of two time series. Examples are windowed cross-correlation (WCC; Tronick, Als, & Brazelton, 1977), windowed cross-lagged correlation (WCLC; Altmann, 2013; Paulick et al., 2018; Ramseyer & Tschacher, 2011), and windowed cross-lagged regression (WCLR; Altmann, 2011). By disassembling the entire time series into windows, the assumption of stationarity can be made locally, which is less restrictive. The length of the analyzed windows is referred to as the bandwidth or window width. The size of this window, in addition to the maximal time lag, is of great importance. Setting the window width too

small results in decreased reliability, whereas a large window width likely results in a violated assumption of local stationarity (Boker et al., 2002). Currently, the selection of an appropriate window width is largely up to the researcher. Theoretically, the window width must be large enough to completely capture the interrelated movements of two persons. From this perspective, Boker et al. (2002) recommended a window width of 4 s. From a methodological point of view, the window must incorporate enough values to determine a stable correlation. Schönbrodt and Perugini (2013) recommended a sample size of at least 65 values to stably detect high correlations. For weaker correlations, 250 values are needed to obtain stable correlations. Tronick et al. (1977) used a window size of 10 s (= 10 values). Cappella (1996) postulated that a window width of at minimum 50–70 values should be used and that this window width should be about 4–5 times larger than the maximal time lag used. However, the optimal trade-off, which comprises a high reliability of correlations, a stationary time series window and a theoretically plausible episode of interrelated nonverbal behaviors, has yet to be empirically determined.

Another important aspect is whether these windows are applied as moving/rolling windows that overlap or whether the time series is split into nonoverlapping windows. Splitting the time series into windows that do not overlap may result in synchrony events at the splitting point not being detected. Therefore, rolling windows with overlap are usually preferred. Regarding forecasting, an application of rolling origins is also preferable to fixed origins in processing time series, as they yield higher efficiency and reliability in out-of-sample tests (Tashman, 2000).

Human interaction is characterized by the interdependence of both interacting partners, meaning that there are periods during which the behavior of person A influences the behavior of person B (A = drive, B = driven), and vice versa (B = drive, A = driven). The drive is also known as the *zeitgeber*, or the person who sets the pace as well as the person who leads (Boker et al., 2002; Kupper et al., 2015; McGarva & Warner, 2003). When using global methods, the reciprocity of an interaction may not be captured, because by shifting one time series with a time lag and calculating the correlation with the second time series, the *zeitgeber* of the interaction is fixed. That is, person A always influences person B, or vice versa. If windows are used, the *zeitgeber* may vary for each window (Ramseyer & Tschacher, 2011). This change of leading may be an important aspect of nonverbal communication (Boker et al., 2002). Using a local method makes examining dynamic interaction changes possible (Boker et al., 2002).

**Output scores** By using different linear TSAMs, different values result as output scores. Paxton and Dale (2013b) reported that a variety of output scores, including the average synchrony score and the highest value, can be used to quantify

synchrony. Most measures assess the strength of synchrony. Altmann (2011, 2013) proposed the frequency of synchrony—that is, the ratio of synchronous time to total time—as an output score. The selection of an output score depends largely on the research question, the researcher (Paxton & Dale, 2013b), and the algorithm, since not every algorithm provides every output score (e.g., the frequency of synchrony can only be calculated using Altmann’s, 2011, 2013, algorithms). With respect to the research question, it is possible to use the average score as an indicator of the overall interrelatedness of an interaction (e.g., Ramseyer & Tschacher, 2011). Additionally, the maximum score of a window can be used to investigate microprocesses in psychotherapy. The frequency of nonverbal synchrony may be used to evaluate interaction on a temporal level (i.e., what percent of the interaction was synchronous). This means that many differences exist with respect to linear TSAMs: Various algorithms have been used, and different output scores may be calculated. Some recent approaches are listed in Table 1 and introduced in the following sections.

**Cross-lagged correlation (CLC)** In comparison to a simple Pearson product-moment correlation, CLC additionally considers the time lag between two time series, so that the correlation between the time series is calculated for each time lag until a maximum value of the time lag is reached (Kato et al., 1983). With respect to its interpretation, CLC refers to echoing. The output of the CLC is a function of the correlations displaying the strength of the association between the time series with respect to every permitted time lag. The two most common outcome scores are the average of all CLC values and the maximum CLC value. By averaging the correlations of different lags, averaged degrees of global matching and echoing are obtained. Using the highest correlation yields the highest global matching or echoing in the sequence.

**Cross-lagged regression (CLR)** An important issue with regard to the analysis of time series data is autocorrelation. Referring to body movement time series, an autocorrelation means that the individual’s current movement is influenced by their previous movements. Not considering this autocorrelation can result in spurious correlations (Altmann, 2011, 2013; Gottman & Ringland, 1981). In recent years, the global measure CLR was used to overcome this shortcoming (Cappella, 1996). When using the CLR, a maximum time lag must also be specified. One difference to CLC is that CLR uses a regression. In this regression, two predictors are incorporated to predict a person’s later movement: (1) the previous movements of the person (autocorrelation) and (2) the previous movement of the interacting partner (cross-correlation). If the model including the autocorrelation and the cross-correlation cannot explain significantly more variance than the model including the autocorrelation only, the association

**Table 1** Global and local TSAMs and their applications

TSAM	Time lag	Windowed	Control for auto-corr	Outcome	Output score			For example, applied by
					av	max	freq	
CLC	X			Correlation function	X	X		Kato et al. (1983), Paxton and Dale (2013b)
CLR	X		X	$R^2$ function	X	X		Gottman and Ringland (1981)
WCC		X		Correlation function	X	X		Tronick et al. (1977)
WCLC <sub>S1</sub>	X	X		Correlation function	X	X		Paulick et al. (2018), Ramseyer and Tschacher (2011)
WCLC <sub>S2</sub>	X	X		Correlation function	X	X		Kleinbub and Ramseyer (2018)
WCLC <sub>F</sub>	X	X		$R^2$ function	X	X	X	Altmann (2011, 2013)
WCLR	X	X	X	$R^2$ function	X	X	X	Altmann (2011, 2013)

C = correlation; CLC = cross-lagged correlation; CLR = cross-lagged regression; WCC = windowed cross-correlation; WCLC<sub>S1</sub> = windowed cross-lagged correlation (e.g., Paulick et al.); WCLC<sub>S2</sub> = windowed cross-lagged correlation R (e.g., Kleinbub & Ramseyer); WCLC<sub>F</sub> = windowed cross-lagged correlation (e.g., Altmann); WCLR = windowed cross-lagged regression; av = average output score; max = maximum output score; freq = frequency output score.

is categorized as nonsynchronous. Models can be compared using  $R^2$  difference testing. Similar to the CLC, the CLR results in a function of  $R^2$  values referring to every permitted time lag. The interpretation of possible output scores (average, maximum) is similar to the CLC.

Note that the distinction between correlational and regressive approaches refers to the way in which the association between both persons' time series is operationalized (Altmann, 2011, 2013). However, both approaches are similar, because a correlation is a one-predictor regression with standardized variables.

**Windowed cross-correlation (WCC)** The WCC represents the simplest local measure. Similar to the simple Pearson product-moment correlation, the correlation is calculated without considering a time lag. The only difference is that the time series is disassembled into windows, resulting in the correlation being computed window-wise (Tronick et al., 1977). The WCC can be applied with overlapping or nonoverlapping windows. Before computing the WCC, the bandwidth must be specified. As a result of the WCC, a function displaying the strength of the association between the two time series is obtained for each window. With respect to the WCC, a commonly used output score is the mean of all obtained cross-correlations (Bilakhia et al., 2015; Nagaoka & Komori, 2008). Accordingly, the resulting output score of the WCC can be interpreted as the averaged degree of local matching. The maximum score can also be determined as the maximum correlation of one window, operationalizing the maximum synchrony in the sequence.

**Windowed cross-lagged correlation (WCLC)** To our knowledge, the WCLC was first applied by Watanabe (1983) and became more popular after its introduction by Boker et al. (2002), who used automated motion capture time series and combined WCLC with a peak-picking algorithm. In WCLC, the association between the time series windows of two

individuals is calculated using either overlapping or nonoverlapping windows. The correlation is calculated window-wise—that is, as a first step the correlation between the first window of the time series of person A and the first window of the time series of person B is conducted (time lag = 0). Afterward, one time series window is shifted by one time lag and the correlation between the first window of the time series of person A and the second window of the time series of person B is conducted (time lag = 1). This procedure is repeated until the maximum time lag is reached and all windows of both time series have been used. After applying the WCLC an  $m \times n$  matrix of values is obtained, where  $m$  characterizes the number of different time lags and  $n$  denotes the time point (in frames). Note that a positive time lag refers to predictions in which person A is the drive and person B is the driven, whereas a negative time lag means that person B is the drive and person A the driven.

Various versions of the WCLC can be distinguished. Versions differ mainly in their preprocessing and processing of the time series or how they use the matrix of correlations/ $R^2$  to generate an output score quantifying synchrony. In the following study, three versions of the WCLC are presented. The first version, called WCLC<sub>S1</sub> here and developed by Paulick et al. (2018), is based on correlations and computes the strength of the association between the two time series as an outcome measure (subscript S stands for strength, and 1 indicates that it is the first WCLC strength measure). More precisely, the authors re-created the algorithm by Ramseyer and Tschacher (2011), which is the most popular method within the psychotherapeutic context. The corrected motion energy time series ( $z$ -transformed, with the threshold for minimal movement; see Grammer et al., 1999) are cross-correlated in nonoverlapping windows of a 1-min duration (for each window, cross-correlations were computed for positive and negative time lags of up to 5 s, in steps of 0.04 s). Subsequently, the cross-correlations of the matrix are

Fisher's- $Z$  transformed and their absolute values are aggregated to an output score for the nonverbal synchrony of each video sequence.

The second version is  $WCLC_{S2}$ , developed by Kleinbub and Ramseyer (2018), which was recently published as the R package *rMEA* (again the subscript  $S$  is for strength, and 2 for the second WCLC strength measure). The function incorporates the following input variables: *mea* (*MEA* time series), *lagSec* (maximum time lag, in seconds), *winSec* (bandwidth, in seconds), *incSec* (step size, in seconds), *rZ* (application of Fisher's  $z$ -transformation), and *ABS* (transformation to absolute values). It results in a cross-correlation matrix that is aggregated to an output score measuring the strength of the association between two time series. A difference between the two presented  $WCLC_S$  versions is the applied minimal movement threshold in  $WCLC_{S1}$ . Since Paulick et al.'s (2018)  $WCLC_{S1}$  is based on Ramseyer and Tschacher's (2011) procedure, Grammer et al.'s (1999) threshold for minimal movements was applied before calculating the cross-correlations. However, in  $WCLC_{S2}$  this threshold is not present anymore. Apart from this difference,  $WCLC_{S1}$  and  $WCLC_{S2}$  are comparable.

The third version is  $WCLC_F$  (where the subscript  $F$  stands for *frequency*) by Altmann (2011, 2013). Altmann (2013) used the determination coefficient (squared correlation) as a result of the WCLC because it contains only positive values, and higher correlations are weighted higher than very low correlations. An example of a time series including a synchronization interval, surface and contour plots of a possible matrix are displayed in Fig. 1. The distinction between the algorithms measuring the strength ( $WCLC_{S1}$ ,  $WCLC_{S2}$ ) and this algorithm lies in the peak-picking algorithm, that is used to identify synchronization intervals. The peak-picking algorithm is described in the Peak-Picking Algorithm paragraph below. In this version of the WCLC, not the synchrony of the entire interaction is quantified, but the local synchronization intervals. These are characterized by the highest values/ $R^2$ 's of synchrony in the local environment (Boker et al., 2002), which are arranged on a horizontal line (see Fig. 2). To compute an output score of synchrony referring to the  $WCLC_F$ , the duration of all synchronization intervals of a particular time series pair was summed up. Afterward the ratio of the time, synchronized to the total duration of the sequence, was calculated (Altmann, 2011, 2013). Thus, the output score indicates the percentage of synchronous interactions. Referring to the  $WCLC_{S1}$  and  $WCLC_{S2}$ , correlations of the different time lags and time points were averaged to obtain an output score. Therefore, another difference between  $WCLC_F$  and both versions of  $WCLC_S$  is the output score, because this version assesses instead the frequency of the associations between the two time series as the primary output score. However, the algorithm also computes the strength of the association in the synchronization intervals. Therefore, the average and maximum output scores are also calculated. Note that in all

versions, only positive values are used. In  $WCLC_{S1}$  and  $WCLC_{S2}$ , absolute values of the correlations are used, whereas in  $WCLC_F$ ,  $R^2$  is used. The procedures differ in that calculating the  $R^2$  values results in weighting higher correlations more than lower correlations. WCLC conceptualizes both matching and echoing locally.

**Windowed cross-lagged regression (WCLR)** Altmann (2011, 2013) developed WCLR, which is a combination of CLR and WCLC, in order to benefit from the advantages of both methods. WCLR is similarly to WCLC. The main difference is that a regression is conducted that includes the previous behavior of an individual as a predictor (similar to the CLR). Interrelatedness is computed window-wise, whereby the position of the reference window and time-lag to the other window are varied. Altmann (2011, 2013) proposed the application of two models on one pair of windows:

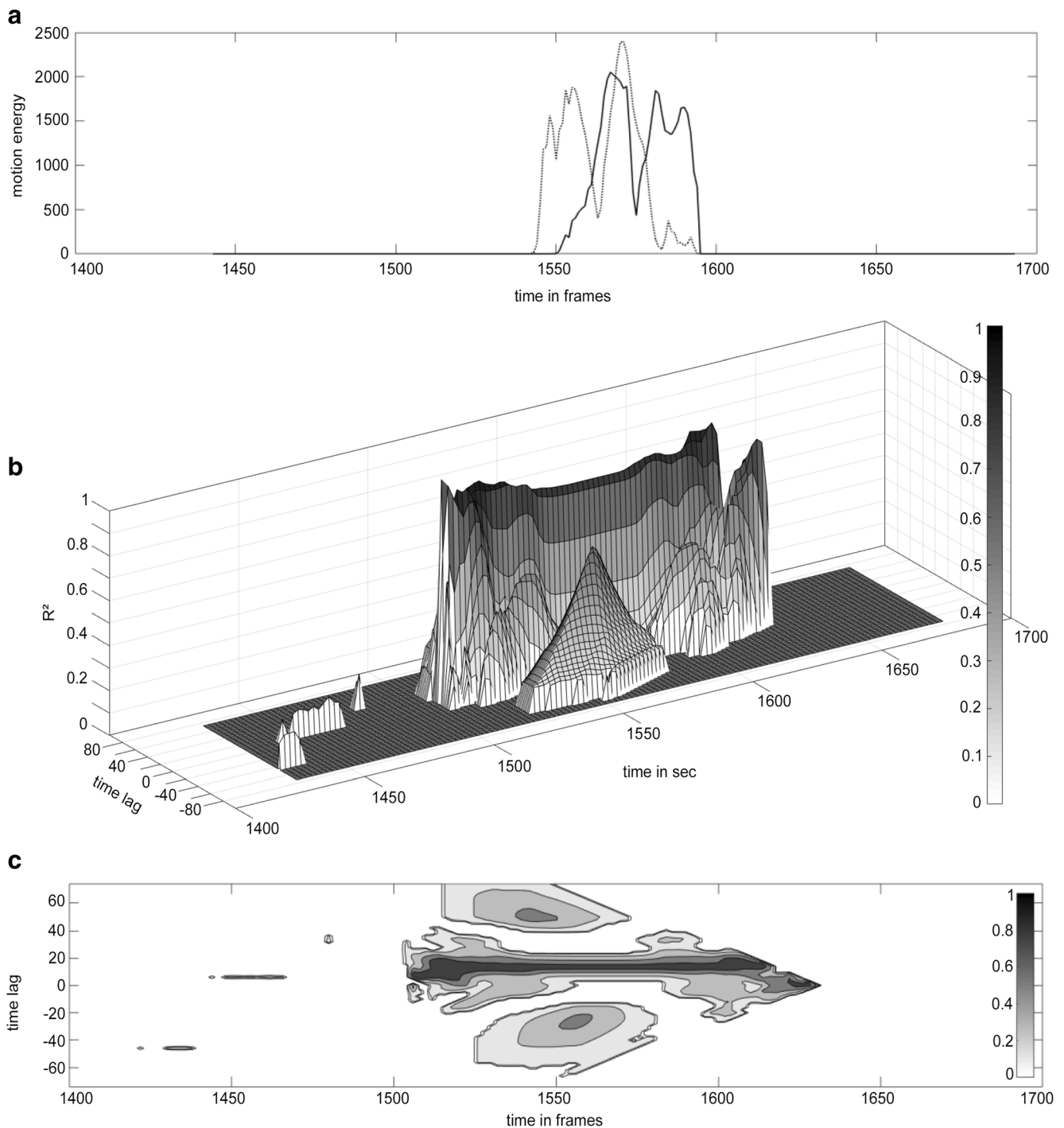
$$\text{Model 1 : } X_{\text{self } t} = \alpha_0 + \alpha_1 X_{\text{self } s} + \varepsilon_{M1 t} \quad (1)$$

$$\text{Model 2 : } X_{\text{self } t} = \beta_0 + \beta_1 X_{\text{self } s} + \beta_2 X_{\text{partner } s} + \varepsilon_{M2 t} \quad (2)$$

Equation 1 shows Model 1, incorporating the previous behavior of an individual in order to predict the current behavior—that is, the regression accounts for possible autocorrelation.  $X_{\text{self } t}$  refers to the individual's own behavior at time point  $t$ , and  $X_{\text{self } s}$  refers to the individual's previous behavior at time point  $s$  (i.e.,  $s$  is prior to  $t$ ).  $\alpha_0$  denotes the intercept,  $\alpha_1$  the slope of the linear regression, and  $\varepsilon_{M1 t}$  refers to the residual variance of Model 1 at a time point  $t$ . These terms are used analogously in Eq. 2. Additionally, a second predictor,  $X_{\text{partner } s}$ , is included, which denotes the previous behavior of the interacting partner at time point  $s$ .

Accordingly, if Model 2 fits significantly better than Model 1, synchrony between the two individuals is detected. Whether Model 2 fits significantly better than Model 1 can be investigated with an  $R^2$  difference test (Altmann, 2011). Computing the WCLR results in an  $m \times n$  matrix of  $R^2$  difference values, where  $m$  characterizes the number of different time lags and  $n$  denotes the time point (in frames). To determine an output score, a peak-picking algorithm is applied to the matrix. The output of the WCLR (and the WCLC) is a list of intervals that were classified by the algorithm as synchronous (time-lagged) time series sequences. On the basis of this list, the ratio of synchronous time to the total duration of the sequence is calculated. Altmann (2013) showed that the WCLR is preferable to the WCLC by using simulated (cyclic) time series. The WCLR also conceptualizes local matching and echoing.

**Peak-picking algorithm** The output of the  $WCLC_F$  and WCLR (implemented by Altmann, 2011, 2013) is an  $R^2$  matrix, which stores the strength of window-wise associations between both persons' time series. To extract the synchronization intervals

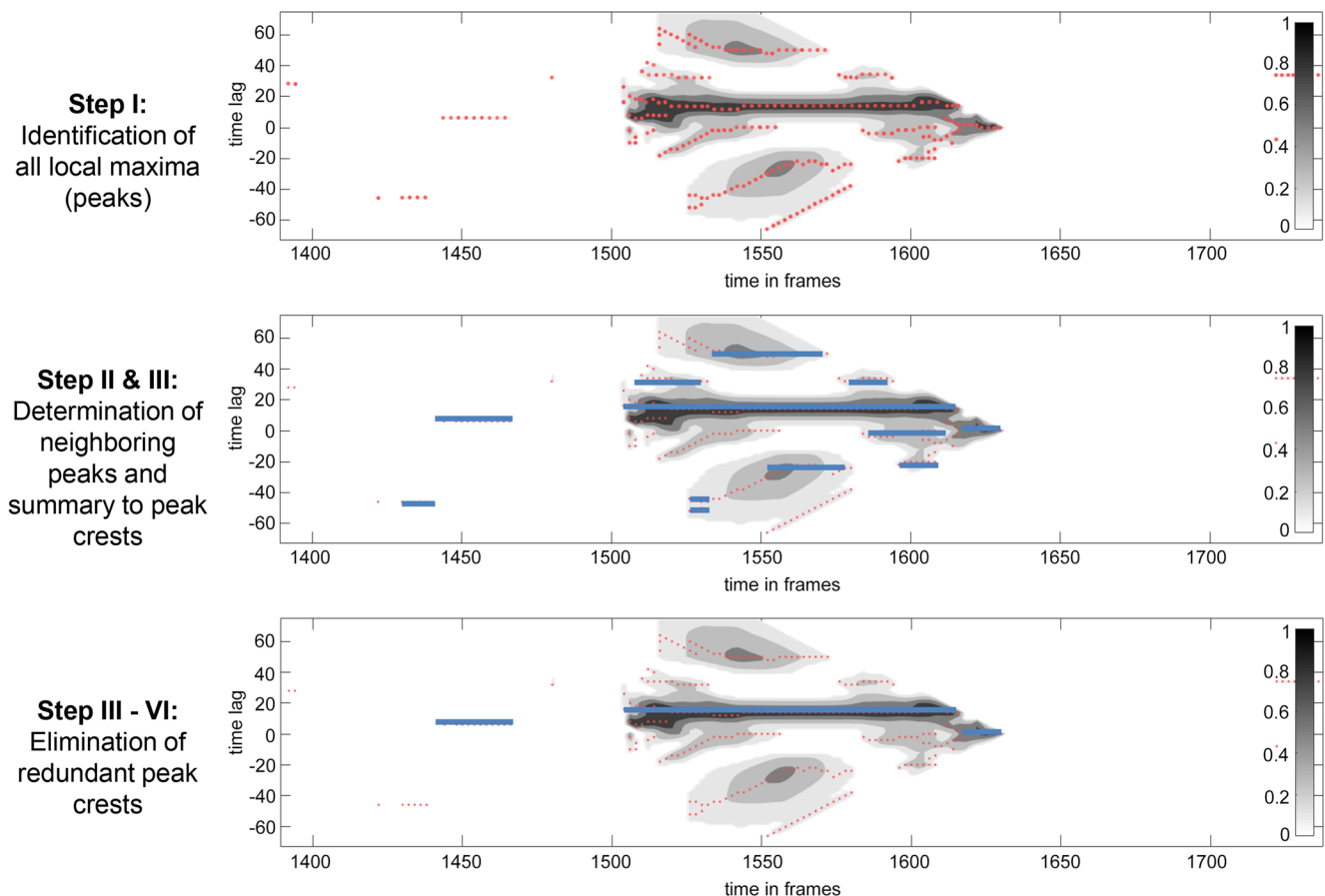


**Fig. 1** Time series and two plots of the resulting  $R^2$  matrix after applying  $WCLC_F$ . Panel A shows the two time series (dotted and solid) with an interval displaying synchrony (interval duration approximately 2 s,

applied bandwidth 3 s = 75 frames displayed in gray). Panel B shows a surface plot of the  $R^2$  matrix, and panel C a contour plot of the  $R^2$  matrix. The time lag is displayed in frames

from the matrix, peaks must be found. Boker et al. (2002) defined the maximum of the cross-correlation as a peak, where values decrease on either side of the particular peak in a specific local region. In their algorithm, the peak that has the closest distance to time lag = 0 is chosen as the best peak to identify the synchronization interval. Altmann (2011, 2013) created an alternative peak-picking algorithm (for an

illustration, see Fig. 2). First, all local maxima in the  $R^2$  matrix (peaks) are identified. Next, neighboring peaks with an equal time lag are determined. The third step is a summary of neighboring peaks to peak crests (with a small time lag tolerance). Fourth, all intervals that last less than 0.4 s are removed, because this amount of time is considered too short to contain meaningful synchrony. Subsequently, an adjustment of the



**Fig. 2** Contour plot of the  $R^2$  matrix, which is a result of the analysis of two time series with  $WCLC_F$ . Red dots indicate the  $R^2$  peaks selected by the Boker et al. (2002) algorithm, whereas blue lines are the  $R^2$  “mountain crests” selected by the Altmann algorithm

onset of an interval takes place using the time lag. Finally, overlapping peak crests are identified and peak crests with the lower  $R^2$  average are removed (see Fig. 2).

From a theoretical point of view, the peak-picking algorithms by Boker et al. (2002) and Altmann (2013) have fundamentally different implications for the concept of nonverbal synchrony. The former implies that synchrony occurs over the entire duration of the interaction, with only the strength of synchrony varying over time. Therefore, the primary aim is not the detection of synchronization intervals but the quantification of synchrony. This implies that investigations of microprocesses within therapy cannot be conducted, because no synchronization intervals are determinable. The latter peak-picking algorithm assumes the existence of so-called *synchronization intervals* and that the time series are interrelated only during these intervals. On the basis of such an on–off pattern, the frequency of synchrony can be quantified. However, within a synchronization interval, Altmann’s (2013) peak-picking algorithm also computes the strength of synchrony in terms of the average  $R^2$  values of all identified synchronization intervals. This implies the assumption that an interaction contains in-synchrony intervals and also intervals during which no synchrony can be found. It allows for an examination of microprocesses within psychotherapy, in

which the content of therapy may be related to different synchronization intervals and in which specific moments of change (e.g., repairs of alliance ruptures) may be investigated further.

## Research questions

The aim of the present study was twofold: to test construct validity and investigate criterion validity with two different criteria. Since authors have used different parameter settings (e.g., bandwidth or overlapping vs. nonoverlapping windows), we conducted all analyses on a set of parameters recommended by the authors and on a completely equal set of parameters, for comparability both between studies and within the present study.

**Convergent validity** First, we examined the convergent validity of linear TSAMs used in behavioral and social science. We applied these TSAMs to motion energy time series, resulting in movement synchrony output scores. Thereby, we used various output scores (average, highest value, and the ratio of synchronous time). We hypothesized that all correlations would be moderate to high because all TSAMs conceptualize nonverbal synchrony and are widely applied to motion energy time series



in order to assess movement synchrony. Additionally, we tested whether a common factor model fits the data. Should this not be the case, an exploratory factor analysis would be conducted to systemize the output scores of different TSAMs.

**Criterion validity** (1) We described CRQA as a widely applied nonlinear approach to investigating synchrony that is more generalized than the linear TSAMs presented. Thus, we also conducted correlations with this measure and included it in the factor analysis in order to validate the results of the linear TSAMs. (2) Our sample comprised patients suffering from social anxiety disorder (SAD), who are considered to show a high level of impairment in interpersonal contexts (Wenzel, Graff-Dolezal, Macho, & Brendle, 2005). Furthermore, with respect to therapy outcome in a disorder-heterogeneous sample, research has shown that psychotherapy enhances interpersonal abilities and reduces interpersonal problems, especially when movement synchrony during the early stage of therapy is high (Ramseyer & Tschacher, 2011). Therefore, we hypothesized that output synchrony scores would be negatively related to interpersonal problems at the end of therapy. We conducted partial correlations (controlling for initial impairment) to examine the predictive validity of the presented TSAMs.

## Method

### Background

Our data were gleaned from video recordings from the multi-center randomized controlled SOPHONET treatment study, conducted from 2007 to 2009 by outpatient clinics at universities in Bochum, Dresden, Jena, Mainz, and Goettingen, Germany (Leichsenring et al., 2013, 2014).

### Setting and material

The therapies included five preparatory sessions that are compulsory in the German health care system (for more details, see Leichsenring et al., 2013, 2014), as well as 25 individual 50-min treatment sessions. Due to the fact that the videos were recorded in different centers, camera positions varied slightly (lateral view  $\sim 90^\circ$ , angular view  $\sim 45^\circ$ ). The psychotherapist and patient were always recorded with one video camera. Furthermore, only video files with a constant camera position, camera settings, and light conditions were considered. To be included in the study, videos had to show both persons (therapist and patient) for at least 15 min during the first half of the therapy session. The starting point of the considered sequence was set after welcoming, administrative questions, and the filling out of questionnaires. The latest possible start time was set to 10 min. Videos that showed the filling in of questionnaires for more than 10 min were excluded. Further exclusion criteria

included: the presence of another person apart from the therapist and patient, one person leaving his/her chair during the 15 min, and one person changing his/her position so that he/she was no longer visible in the video. To ensure the comparability of video files, they were converted to equivalent formats using a size of  $640 \times 480$ , a frame rate of 25 fps, and a bit rate of 2,000 Kbit per second (Any Video Converter 3.0, AVC, 2009).

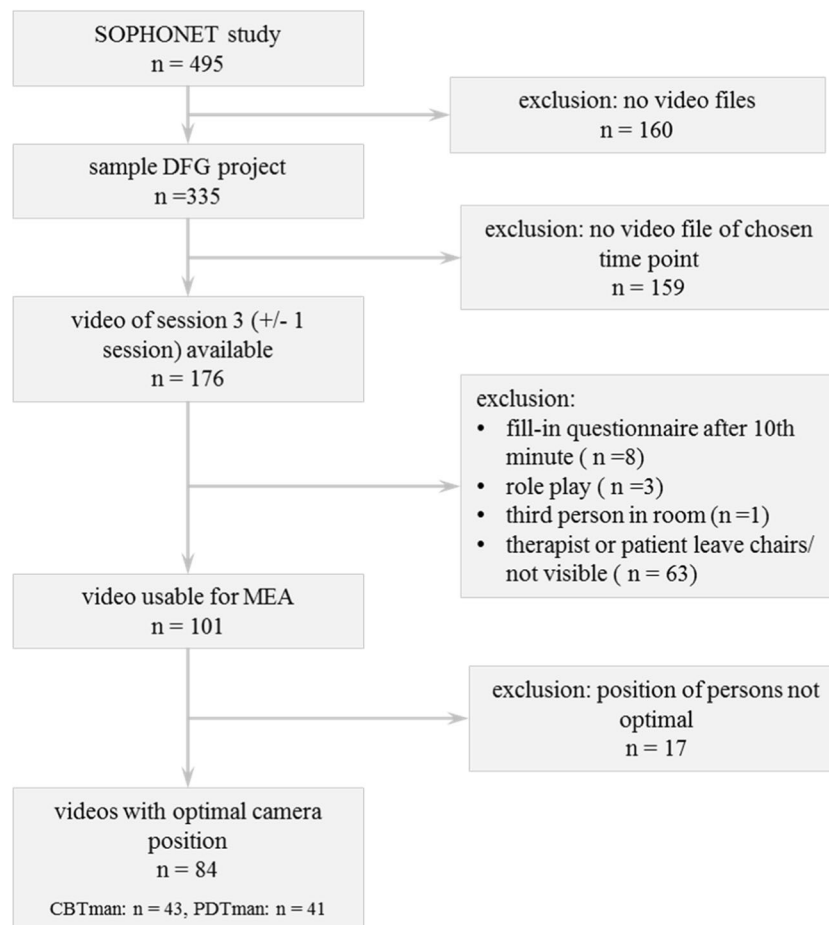
### Study subjects

On the basis of the inclusion criteria of the SOPHONET study, the patients had to be 18–70 years old, have a diagnosis of SAD according to the German version of the Structured Clinical Interview for DSM-IV (Wittchen, Wunderlich, Gruschwitz, & Zaudig, 1997), a score  $> 30$  on the Liebowitz Social Anxiety Scale (LSAS; Mennin et al., 2002), and a primary diagnosis of SAD with respect to the Anxiety Disorders Interview Schedule (Brown, Barlow, & Di Nardo, 1994). Comorbid disorders less severe than SAD were allowed, with the exception of psychotic disorders, acute substance-related disorders, and cluster A and B personality disorders. Additional exclusion criteria were a risk of self-harm, organic mental disorders, severe medical conditions, and concurrent psychotherapeutic or psychopharmacological treatments (see Leichsenring et al., 2013, 2014).

In the present study, we examined a subsample of the SOPHONET sample. The selection was driven by the following inclusion criteria for video quality: (1) patients had to be visible for the first 15 min of the third therapy session, (2) video quality was sufficient without severe video errors, and (3) no third person was present. After applying the inclusion criteria to the video files, 84 of 495 study subjects remained for the present investigation (study flow is presented in Fig. 3). Regarding the present investigation, 53 patients were female, 52 had a high school diploma or higher, and 37 were living in a current relationship. The mean therapy duration per patient was 23.87 sessions ( $SD = 6.93$ ). The mean pretreatment severities were 73.44 ( $SD = 21.79$ ) on the LSAS, 10.41 ( $SD = 6.51$ ) on the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), and 1.74 ( $SD = 0.47$ ) on the Inventory of Interpersonal Problems (IIP; Horowitz, Strauss, Thomas, & Kordy, 2016).

### Assessment of movement

We used the first 15 min of the third therapy session ( $\pm$  one session) to conduct motion energy time series analysis, because it has been demonstrated that movement synchrony of the first 15 min is representative of the entire 50 min of a therapy session (Paulick et al., 2018; Ramseyer & Tschacher, 2011). To assess the time course of individual body movements, we used motion energy analysis (MEA), implemented in MATLAB 2016 (The MathWorks, Inc, 2016; Altmann, 2013).



**Fig. 3** Study flow. DFG = German Research Foundation, CBT = cognitive–behavioral therapy, PDT = psychodynamic-oriented therapy, man = manual-guided

The ROI covered the upper body from the chair’s seating-base upward.

We determined the cutoff value for meaningful pixel changes empirically (according to Altmann, 2013). For this purpose, we used  $n = 20$  video files. We specified 30 different pixels in the background as control pixels, which should not show any grayscale intensity changes as a result of movements. Time series of these 30 pixels were calculated, showing the intensity change from one frame to another. Afterward, we determined the 99% quantile of all values, which equaled 12—that is, 99% of the intensity changes were lower than 12. Thus, the cutoff value was set to 12. Furthermore, we applied a 2-D median filter in order to reduce video noise (i.e., irregularities in the video sequences), as described in Altmann, Schoenherr, Paulick, Knitter, et al. (submitted). In addition, we transformed all motion energy time series to a value range of 0–100 by dividing each value by the number of pixels in the ROI and multiplying it by 100. The value of 100 means that 100% of the ROI was activated; the value of 1 means that 1% of the ROI was activated. Therefore, the values of different persons were comparable, even when they did not have the same ROI size. Moreover, we filtered video errors by

applying the MATLAB script described in Altmann, Schoenherr, Paulick, Knitter, et al. (submitted), available at <https://github.com/10101-00001/MEA>.

### Parameter settings for the present analysis

When choosing parameter settings, we wanted to have maximum comparability to the literature and maximum comparability within the article. Therefore, we applied the algorithms first as described in the literature (heterogeneous parameter settings), and second with entirely equal parameter settings and output metric (homogeneous parameter settings). If no standard was available in the literature, we chose parameter settings that facilitated comparability of the algorithms. Since there is no common standard with respect to data transformation and smoothing procedure, we neither smoothed nor transformed the time series, to avoid information loss. Referring to the maximal time lag considered, we used a homogeneous maximal time lag of 5 s (in accordance to Ramseyer, 2011).

**Heterogeneous parameter settings** For the CLC or CLR, we applied a simple cross-correlation or cross-regression with

step size 0.04 s (= 1 frame). The CLC results in an absolute global correlation, the CLR in an  $R^2$  value. For the WCC, we used a bandwidth of 10 s (= 250 frames) with a step size of 1 s (= 25 frames; Tronick et al., 1977). As a global score, an averaged/maximum global absolute correlation was used. With respect to  $WCLC_{S1}$  and  $WCLC_{S2}$ , we applied a 1-min window as bandwidth (= 1,500 frames) with nonoverlapping windows (step size = 1 min = 1,500 frames; Paulick et al., 2018; Ramseyer & Tschacher, 2011). Since a moving median had already been applied to the raw data, we omitted the moving average of the Ramseyer/Paulick procedure in order to apply the algorithms to the same raw data. Before calculating the correlations, a  $z$ -transformation was applied. In addition,  $WCLC_{S1}$  applies a minimal threshold for movement (see Grammer et al., 1999; Ramseyer & Tschacher, 2011). Both  $WCLC_{S1}$  and  $WCLC_{S2}$  result in an absolute averaged/maximum correlation as the output score. For  $WCLC_F$  and WCLR, we used a bandwidth of 5 s (= 125 frames) with a step size of 0.04 s (= 1 frame), as well as overlapping windows (Altmann, Schoenherr, Paulick, Deisenhofer, et al., 2018). In the present study, Altmann's (2013) peak-picking algorithm was used for the  $WCLC_F$  and WCLR with a time lag tolerance of one frame (which equates to 0.04 s). All synchronization intervals that lasted less than 0.4 s were removed because they were too short to display meaningful synchrony (for details, see Altmann, 2013). All algorithms had the same time series as input. All time-lagged methods had a maximum time lag of 5 s.

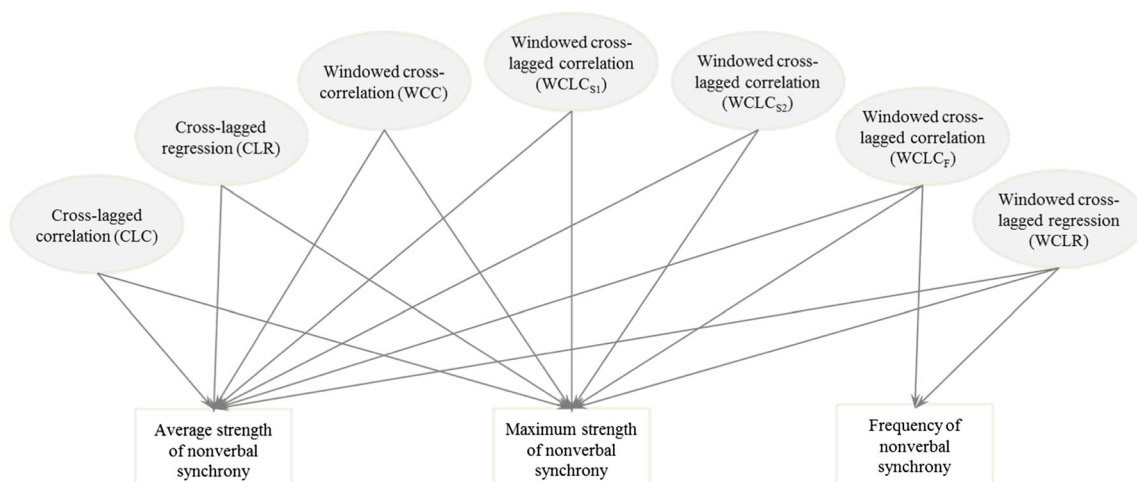
**Homogeneous parameter setting** To enhance comparability, all algorithms were used with equal settings: same time series, maximum time lag of 5 s (= 125 frames), bandwidth of 5 s (= 125 frames), and overlapping windows with a step size of 0.04 s (= 1 frame). To allow comparison of the correlations and  $R^2$  values of the different methods and obtain a homogeneous metric, we squared the simple correlations.

We calculated various output scores for each method, when possible (average absolute correlation = average strength of synchrony, highest absolute correlation = maximum strength of synchrony, ratio of synchronous time to total duration = frequency of synchrony). The global output scores used with respect to the different TSAMs are displayed in Fig. 4.

For a criterion-based validation with a nonlinear method, we also computed synchrony scores using CRQA, implemented in R by Coco and Dale (2014). We preprocessed our time series via  $z$ -transformation and determined the optimal radius and embedded dimensions by using the `optimizeParam` function. The mean radius of all time series was 8.85, and the optimal value of the embedded dimensions was 3. We calculated recurrence rate (RR) profiles by applying the `runCrqa` function for continuous data with a window size of 125 frames (= 5 s). Afterward, we averaged the RR profiles to obtain an output score (average) representing nonlinear movement synchrony. Additionally, we calculated the mean of the maximal RR (maxrec) to obtain an output maximum score.

## Statistical analysis

After video analysis with MEA and computation of synchrony indices with multiple TSAMs, we examined the association between different synchrony output scores. Descriptive statistics (mean, standard deviation) were investigated. Then we calculated bivariate Pearson product-moment correlations between the synchrony output scores in order to investigate convergent validity. Correlations lower than .35 were considered low, those between .35 and .67 moderate, and those between .68 and 1.0 high (Taylor, 1990).  $P$  values of the correlations were corrected for multiple testing by Bonferroni correction. Correlations with the CRQA score were also conducted with respect to all three output scores (average, max, freq). Then we tested the theoretical assumption that a common factor model holds. Thereby, we first tested the average and ratio/frequency



**Fig. 4** Output measures for the different linear TSAMs

scores, and second we examined the common factor structure of the maximum scores, because including the correlated average and maximum scores in one analysis might have biased the factor analysis. The root mean square error of approximation (RMSEA) should be  $< .05$ , the comparative-fit index (CFI)  $> .9$ , and the Tucker–Lewis index (TLI)  $> .9$  to suggest a satisfactory fit. Because the results showed that a common factor model was not appropriate, we conducted an exploratory factor analysis (maximum likelihood estimator, rotation oblimin), first for the average and ratio scores and second for the maximum scores, in *Mplus 7.4*. Thereby, we specified that the factor solution lay between one and six factors and tested different factor solutions (e.g., two-factor, three-factor, etc.) for data fit. In our exploratory factor analysis, we allowed for correlating factors.

Finally, we examined whether the output scores were equally related to a construct: interpersonal problems at the end of therapy. Interpersonal problems were measured with the IIP (Horowitz et al., 2016), which is composed of eight scales, including dominance, self-sacrifice, and social inhibition, that measure interpersonal problems. Each scale consists of four items based on a 5-point Likert scale. We conducted partial correlations while controlling for interpersonal problems at the beginning of therapy.

It should be noted that we performed each analysis on two datasets. The first dataset ( $N = 84$  time series pairs) were synchrony measures based on the parameter settings suggested by the authors of the algorithms. The second dataset were synchrony measures computed when the same parameter settings were applied to each algorithm. This was done to maximize both comparability with other studies (heterogeneous settings) and comparability within our own study (homogeneous settings).

## Results

### Descriptive statistics

The means and standard deviations of the examined synchrony output scores and time lags are displayed in Table 2.

### Correlation analysis

The correlations of different synchrony output scores are displayed in Table 3 with respect to the average strength and frequency of synchrony. The table captures correlations in which each algorithm was applied both as recommended in

**Table 2** Means and standard deviations of synchrony indices and their time lag variables

	Synchrony indices			Time lag		
	Mean (SD)			Mean (SD)		
TSAM	Average score	Max score	Ratio score	Average score	Max score	Ratio score
Heterogeneous parameter settings						
CLC	.036 (.023)	.075 (.039)	–	0	– 8.05 (76.85)	–
CLR	.005 (.014)	.013 (.039)	–	0	108.52 (23.97)	–
WCC	.115 (.029)	.704 (.121)	–	0	0	–
WCLC <sub>S1</sub>	.047 (.015)	.111 (.041)	–	0	0	–
WCLC <sub>S2</sub>	.084 (.015)	.750 (.245)	–	0	0	–
WCLC <sub>F</sub>	.269 (.046)	.871 (.059)	.442 (.108)	64.18 (2.97)	– 9.57 (73.70)	64.18 (2.97)
WCLR	.252 (.041)	.854 (.055)	.444 (.108)	65.71 (3.01)	– 2.29 (72.35)	65.71 (3.01)
CRQA	.168 (.095)	.177 (.095)	–	0	0	–
Homogeneous parameter settings						
CLC	.002 (.003)	.007 (.009)	–	0	– 8.05 (76.85)	–
CLR	.005 (.014)	.013 (.039)	–	0	108.52 (23.97)	–
WCC	.018 (.008)	.738 (.148)	–	0	0	–
WCLC <sub>S1</sub>	.008 (.002)	.042 (.032)	–	0	0	–
WCLC <sub>S2</sub>	.038 (.008)	.997 (.006)	–	0	0	–
WCLC <sub>F</sub>	.269 (.046)	.871 (.059)	.442 (.108)	64.18 (2.97)	– 9.57 (73.70)	64.18 (2.97)
WCLR	.252 (.041)	.854 (.055)	.444 (.108)	65.71 (3.01)	– 2.29 (72.35)	65.71 (3.01)
CRQA	.168 (.095)	.177 (.095)	–	0	0	–

Heterogeneous parameter settings according to the literature. TSAM = time series analysis method; SD = standard deviation; CLC = cross-lagged correlation; CLR = cross-lagged regression; WCC = windowed cross-correlation; WCLC = windowed cross-lagged correlation (WCLC<sub>S</sub>, with respect to strength of synchrony; WCLC<sub>F</sub>, with respect to frequency of synchrony); WCLR = windowed cross-lagged regression.  $N = 84$ , all values in frames

**Table 3** Correlations between *average* output scores of nonverbal synchrony assessed with various TSAMs

	CLC	CLR	WCC	WCLC <sub>S1</sub>	WCLC <sub>S2</sub>	WCLC <sub>F</sub>	WCLR	CRQA	WCLC <sub>F</sub>	WCLR
									freq	freq
CLC	.960*	.001	.250	.194	.256	.132	.110	.272	.193	.236
CLR	.004	1	-.117	.067	-.026	-.033	-.011	-.110	-.155	-.155
WCC	.268	-.107	.954*	.271	.866*	.770*	.755*	.768*	.874*	.881*
WCLC <sub>S1</sub>	.349*	.037	.607*	.713*	.292	.217	.190	.172	.137	.146
WCLC <sub>S2</sub>	.528*	.085	.734*	.692*	.781*	.791*	.772*	.681*	.733*	.749*
WCLC <sub>F</sub>	.147	-.033	.827*	.492*	.588*	1	.984*	.604*	.753*	.749*
WCLR	.123	-.011	.803*	.460*	.565*	.984*	1	.601*	.767*	.757*
CRQA	.269	-.110	.738*	.454*	.555*	.604*	.601*	1	.769*	.777*
WCLC <sub>F</sub>	.204	-.155	.880*	.443*	.550*	.753*	.767*	.769*	1	.987*
freq										
WCLR	.239	-.155	.892*	.464*	.579*	.749*	.757*	.777*	.987*	1
freq										

$R^2$  upper triangle (gray shaded): correlations of different synchrony indices ( $R^2$ ) based on homogeneous parameter settings; lower triangle: correlations of different synchrony indices (corelations/ $R^2$ ) based on heterogeneous parameter settings from literature. CLC = cross-lagged correlation; CLR = cross-lagged regression; WCC = windowed cross-correlation; WCLC = windowed cross-lagged correlation (WCLC<sub>S</sub>, with respect to strength of synchrony; WCLC<sub>F</sub>, with respect to frequency of synchrony); WCLR = windowed cross-lagged regression; CRQA = Cross recurrence quantification analysis. \* Significant correlations, Bonferroni-adjusted alpha level = .001, freq = frequency of synchrony/ratio

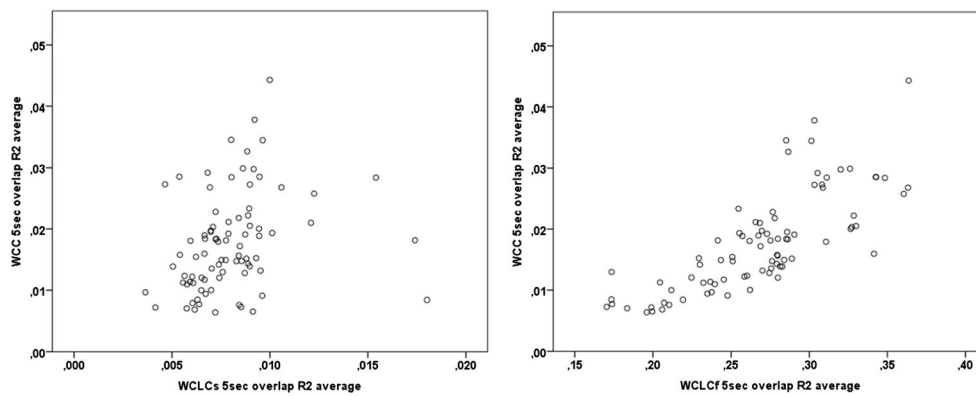
the literature (lower triangle) and with entirely equal parameter settings (upper triangle, gray-shaded). For the sake of clarity, we only present results referring to the average scores and ratio score, here. The other correlations (maximum output scores) are displayed in the [Appendix](#). High correlations between the scores of the more local methods indicate a substantial association between parameter settings, with one exception: The correlations with WCLC<sub>S1</sub> using equal parameter settings are nonsignificantly different from zero. The global methods are neither correlated with each other nor highly correlated with the scores of the local methods. Two associations are illustrated by scatterplots (Fig. 5).

With respect to the maximum scores, most of the output scores were nonsignificantly correlated. Significant associations were only shown referring to the frequency measures (WCLC<sub>F</sub>, WCLR) and between the WCC and frequency measures and CRQA. Additionally, WCLC<sub>S1</sub> (max) and CLC (max) were

associated (see the Appendix, Table 6). The maximum and average scores for each algorithm were associated with each other mostly, but ranged from  $r = -.210$  to  $r = .999^*$  (see the Appendix, Tables 7 and 8; here and below, asterisks indicate results significant at the alpha level given in particular tables).

### Exploratory factor analysis

With respect to our two parameter settings, we conducted two factor analyses (A: output scores resulting with heterogeneous parameter settings [according to literature], B: output scores resulting with homogeneous parameter settings). We included average scores and ratio/frequency scores only because the highly correlated average and maximum scores might have biased the factor analysis. For neither of the settings, a common factor model fit the data (A:  $\chi^2 = 380.25$ ,  $df = 35$ ,  $p < .01$ ,



**Fig. 5** Scatterplots showing the correlations between WCC (average) and  $WCLC_{S1}$  (average), on the left, and  $WCLC_F$  (average), on the right

RMSEA = 0.34, CFI = 0.67, TLI = 0.58; B:  $\chi^2 = 331.87$ ,  $df = 35$ ,  $p < .01$ , RMSEA = 0.32, CFI = 0.71, TLI = 0.63).

For the exploratory analyses, the results showed that a three-factor model fit the data for both parameter settings (A:  $\chi^2 = 19.24$ ,  $df = 18$ ,  $p = .38$ , RMSEA = 0.029, CFI = 0.999, TLI = 0.997, B:  $\chi^2 = 9.80$ ,  $df = 18$ ,  $p = .94$ , RMSEA < 0.00, CFI = 1.00, TLI = 1.02). The factor loadings of the three-factor solution are displayed in Table 4. The correlations were, between Factors 1 and 2,  $r = .742^*$  for the dataset generated with heterogeneous parameter settings for the algorithms (respectively,  $r = .726^*$  for homogeneous parameter settings), between Factors 1 and 3,  $r = .576^*$  (respectively,  $r = .737^*$ ), and between

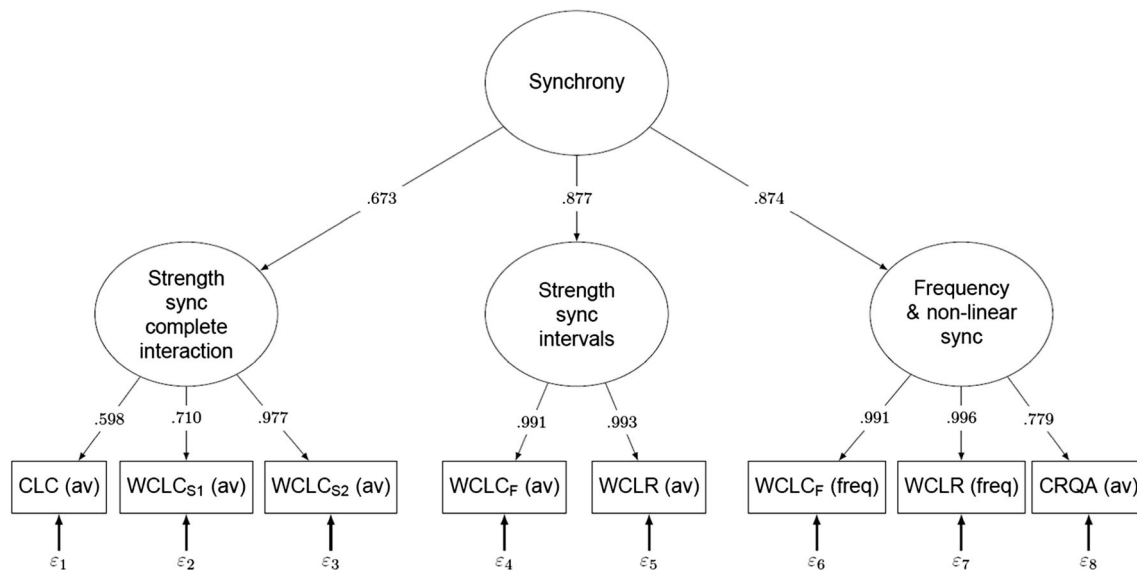
Factors 2 and 3,  $r = .564^*$  (respectively,  $r = .702^*$ ). For maximum scores, no adequate factor solution was found.

**Post-hoc analysis** On the basis of the three-factor solution of the exploratory factor analysis (EFA) and our main hypothesis (common factor for synchrony), we also conducted a confirmatory factor analysis to test whether the three factors loaded on a single factor. Therefore, we specified three latent endogenous variables based on the factor loadings and one superior latent synchrony variable. Referring to the heterogeneous parameter settings, we found an adequate model by excluding the WCC and CLR (which had double loadings in EFA). The

**Table 4** Factor loadings (geomin-rotated loadings) of the exploratory factor analysis

	Heterogeneous parameter settings			Homogeneous parameter settings		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
CLC (av)	-.041	-.128	.630*	.034	-.228	.448*
CLR (av)	-.296*	-.010	.263*	-.364*	.251	.028
WCC (av)	.561*	.232*	.283*	.422*	.017	.592*
$WCLC_{S1}$ (av)	.026	.097	.638*	-.266	.002	.538*
$WCLC_{S2}$ (av)	.019	.006	.967*	-.003	.224	.767*
$WCLC_F$ (av)	-.013	1.038*	.006	.003	.882*	.126*
$WCLC_F$ (freq)	1.018*	.000	-.042	.969*	.041	.000
WCLR (av)	.140	.852*	-.002	.037	.990*	-.009
WCLR (freq)	.996*	-.009	.007	.920*	.011	.080
CRQA (av)	.689*	-.005	.163	.542*	-.062	.370*

Heterogeneous parameter settings according to the literature. CLC = cross-lagged correlation; CLR = cross-lagged regression; WCC = windowed cross-correlation;  $WCLC$  = windowed cross-lagged correlation ( $WCLC_S$ , with respect to strength of synchrony;  $WCLC_F$ , with respect to frequency of synchrony); WCLR = windowed cross-lagged regression. The highest factor loadings are shaded in gray. \* Significant at a 5% alpha level, freq = frequency



**Fig. 6** Common factor model synchrony (heterogeneous parameter settings)

model showed an excellent fit ( $\chi^2 = 7.99$ ,  $df = 18$ ,  $p = .98$ ,  $RMSEA < 0.001$ ,  $CFI = 1.000$ ,  $TLI = 1.018$ ). The model with significant (standardized) path coefficients is displayed in Fig. 6. Regarding the homogeneous parameter settings, no adequate and converging model was found.

### Criterion-based validation: Correlations with interpersonal problems

Correlations of the output scores with interpersonal problems at the end of therapy (while controlling for initial interpersonal problems) are displayed in Table 5. The hypothesized negative association between nonverbal synchrony and interpersonal problems posttherapy was found with respect to WCC (average), WCLCF

(ratio), and WCLR (ratio), which showed significant correlations. Additionally, marginally significant associations were found with regard to WCLCF (average) and WCLR (average).

### Discussion

The aim of the present study was to examine the convergent validity of linear TSAMs for the assessment of nonverbal synchrony. Besides the diversity of TSAMs, most of these algorithms provide more than one output score—for instance, an average and a maximum score. Therefore, various output scores per TSAM were calculated when possible. In the literature, these different output scores and TSAMs are all used to

**Table 5** Partial correlations between output scores and interpersonal problems at the end of therapy, with  $p$  values

Output score	Heterogeneous parameter settings		Homogeneous parameter settings	
	$r$ IIP <sub>Post</sub>	$p$ Value	$r$ IIP <sub>Post</sub>	$p$ Value
CLC (average)	-.029	.799	-.090	.419
CLR (average)	.076	.497	.076	.497
WCC (average)	-.268*	.015	-.305*	.005
WCLCS <sub>1</sub> (average)	-.179	.108	-.086	.444
WCLCS <sub>2</sub> (average)	-.123	.272	-.147	.187
WCLCF (average)	-.204	.066	-.204	.066
WCLCF (ratio)	-.262*	.018	-.262*	.018
WCLR (average)	-.205	.065	-.205	.065
WCLR (ratio)	-.282*	.010	-.282*	.010
CRQA (average)	-.138	.215	-.138	.215

Heterogeneous parameter settings according to the literature.  $r$  = correlation; CLC = cross-lagged correlation; CLR = cross-lagged regression; WCC = windowed cross-correlation; WCLC = windowed cross-lagged correlation (WCLCS; with respect to strength of synchrony; WCLCF; with respect to frequency of synchrony); WCLR = windowed cross-lagged regression; CRQA = Cross recurrence quantification analysis. \* Significant correlations at the 5% alpha level

assess nonverbal synchrony. Since it is not clear if all output scores measure the same construct, all TSAMs were applied to an identical dataset of time series pairs. We conducted all analyses with a set of heterogeneous parameter settings (according to literature) and homogeneous parameter settings (highest comparability between algorithms within this study).

The present study was able to demonstrate that not all output scores that are used to calculate synchrony are correlated. Especially, global and local TSAMs measure different facets of synchrony. Whereas global TSAMs assume the interrelatedness of both interacting partners to be stable (i.e., person A always influences person B, or vice versa), local TSAMs operationalize interrelatedness dynamically with varying leading and pacing. Therefore, it is plausible for global and local TSAMs not to be associated with each other. The correlations largely support this inference. Furthermore, there is no evidence that TSAMs that have the same methodical approach (e.g., correlational vs. regressive methods) would inevitably assess the same construct. Additionally, we found no common construct underlying all TSAMs with either of the two parameter settings. The results of the exploratory factor analyses suggest a three-factor solution. However, we were able to show a common factor as a latent second-order variable of these three factors by excluding WCC and CLR for the heterogeneous parameter settings.

### Multifactor structure of synchrony measures

The examined three factors of the EFA incorporated with the results of the correlations are described in the following paragraphs.

**Synchrony Factor 1** The first factor is formed by the  $WCLC_F$  (ratio),  $WCLR$  (ratio), and  $CRQA$  (average) measures. In line with that, all three output scores are highly correlated (ranging from  $r = .77$  to  $r = .99$ ). Both linear measures are frequency measures that capture the ratio of time that was synchronized to the total duration of the sequence. Correlations are very high indicating nearly equal scores. The facet of synchrony that is measured differs enormously in comparison to the other (strength) measures. Apparently, the frequency measure is highly associated with the nonlinear output score ( $r = .77$ ). The result of the  $CRQA$  is the recurrence rate indicating the frequency of revisiting states of both phase trajectories. The construct can be described as the *frequency of synchrony*.

**Synchrony Factor 2** This factor incorporates the average output scores of the  $WCLC_F$  and  $WCLR$ . With respect to the  $WCLC_F$  and  $WCLR$ , a peak-picking algorithm is used, which identifies the start and end points of synchronization intervals as well as synchrony strength within the identified intervals. Their output values are based on the identified synchronization intervals. Intervals that do not show synchrony are

neglected. The construct can best be described as the *strength of synchrony within the identified synchronization intervals*. The strength of synchrony in intervals (Factor 2) and the frequency of synchrony (Factor 1) are both measured by  $WCLC_F$  and  $WCLR$  and are therefore highly associated; however, they are not equivalent. If synchrony is highly frequent, it is not, by default, always strong. The strength of the association is partially determined by the peak height of both time series. If both are high and similar, high strength is identified. In the context of psychotherapy, this would mean that both persons gesture in a large or a very space-consuming fashion. If the magnitude of the persons' movements is very different within their space, strength will be lower, although the interval will also be identified as showing synchrony.

The results showed that the  $WCLC_F$  and  $WCLR$  measures correlated very highly with each other ( $r = .98$ ). Since  $WCLR$  was developed on the basis of cyclical data (Altmann, 2011, 2013), it is probably advantageous for this type of data as compared to  $WCLC_F$ . With respect to the noncyclical data in the present study, the scores were nearly equivalent. Therefore,  $WCLC_F$  seems to be preferable, because the computational expense of  $WCLC_F$  is much lower than that of  $WCLR$ .

**Synchrony Factor 3** Factor 3 of the EFAs incorporates the  $CLC$  output score and the  $WCLC_{S1}$  and  $WCLC_{S2}$  output scores. In comparison to the associations within the other factors, the correlations between  $CLC$  and the other scores of this factor are only moderate. This may be explained because  $CLC$  is a global measure and both forms of  $WCLC_S$  are local measures. However, their belonging to this factor may be explained by the large bandwidth of 1 min used in both forms of  $WCLC_S$ . Since the total duration of the sequence was 15 min, the results of  $WCLC_S$  are similar to those of the  $CLC$  and are assigned to the same factor. Both  $WCLC_S$  measures are highly associated ( $r > .7$ ), which is plausible because both measures differ only in the used the Grammer et al. (1999) minimal movement threshold (see the [Method](#) section). The scores of Factor 3 quantify the strength of synchrony and aggregate the value with respect to the total interaction. Thus, the  $r$  or  $R^2$  values of nonsignificant sequences (sequences without movement synchrony) and movement synchrony intervals are included in the aggregation. The construct is best described as the *strength of synchrony of the total interaction*. The main difference between this factor and the  $WCLC_F$  and  $WCLR$  measures is the peak-picking algorithm. The clear advantage of  $WCLC_F$  and  $WCLR$  lies in their determination of synchronization intervals, which also clearly distinguishes Factors 1 and 2 from Factor 3.

**Special cases CLR and WCC** Our results showed that the  $CLR$  is not related to any other measure, suggesting that  $CLR$  maps a completely different construct than do the other TSAMs. This is emphasized by the finding that no converging and adequate



model was found by incorporating CLR. We do not recommend using CLR as a synchrony measure.

The measure WCC is related to the output scores of all local TSAMs. This corresponds to the results of the factor analysis showing significant factor loadings for more than one factor. Apparently, WCC is related to each facet of synchrony—especially to the frequency of synchrony and the strength of synchrony of the total interaction. This might be plausible because sequences may have similar proportions of matching and echoing. Therefore, the assessed matching of the WCC correlates highly with the sum of echoing and matching of the other algorithms. This does not necessarily imply that it is beneficial not to include a time lag; rather, it presents the opportunity to estimate the general level of synchrony of an interaction using WCC.

The existence of a diversity of procedures and low concordance of scales is also found in other areas of psychological research, such as attachment research. The investigation of the attachment–outcome relation is complicated by the fact that measures of attachment are quite heterogeneous (Bouthillier, Julien, Dubé, Bélanger, & Hamelin, 2002; Kirchmann, Fenner, & Strauß, 2007; Manes et al., 2016; Roisman et al., 2007) and show weak, if any, convergence. Thus, the comparability of studies is reduced. Evidence related to attachment as a predictor of outcome in specific psychological treatments remains unclear (Manes et al., 2016). Recently, it has been assumed that (i) the methods measure different aspects of the construct and (ii) attachment is therefore not unidimensional but, rather, multidimensional. Referring to the results of the present study, the same can be said of nonverbal synchrony: Nonverbal synchrony does not seem to be a unidimensional but, rather, a multidimensional construct. However, the different facets can be related to a superior construct of nonverbal synchrony.

### Criterion-based validations

In the current literature, it is often discussed that nonlinear methods better reflect interpersonal interactions. Therefore, we also used the most commonly applied nonlinear method: cross-recurrence quantification analysis. Different linear output scores are differently associated with the output scores of CRQA: The average score of the CRQA shows high correlations with the local TSAMs. Only the  $WCLC_{S1}$  score with homogeneous parameter settings seems to be an exception. Therefore, the  $WCLC_{S1}$  should not be applied with small bandwidth and overlapping windows. To summarize, in comparison to some linear models, nonlinear models do not necessarily result in completely different synchrony indices. This result can be considered a validation of the local TSAMs with the parameter settings that are recommended in the literature.

When examining empirical research questions with TSAMs, which are related to different synchrony constructs,

inconsistent findings may be found. Tronick et al. (1977), for instance, showed significant nonverbal synchrony between mother–child dyads using WCC, whereby Gottman and Ringland (1981) did not find this association when reanalyzing the very same dataset using CLR. We also conducted partial correlations between ten nonverbal synchrony output scores and therapy outcome (IIP). The assumed significant negative relationship between synchrony and the outcome was only observable when using three of the ten output scores. This would mean that a high level of synchrony at the beginning of therapy is associated with fewer interpersonal problems at the end of the therapy. The scores showing this relationship can be assigned to the same synchrony facet (frequency of synchrony), supporting the idea of different synchrony facets. The association between the IIP results and synchrony was also present marginally significantly with respect to synchrony in the intervals. However, total synchrony was only descriptively and nonsignificantly associated with interpersonal problems. This is inconsistent with the findings of Ramseyer and Tschacher (2011), who reported a significant association. However, Paulick et al. (2018) did not find a linear association between IIP and total synchrony. Assuming an association between synchrony and interpersonal problems, the frequency measures of the  $WCLC_F$  and  $WCLR$  and the WCC (average) are the only valid output scores.

### Practical implications

Since the output scores do not measure the same nonverbal synchrony construct, the question arises, which of the TSAMs and their output scores best measure nonverbal synchrony. The present study is unable to answer this question completely, because we have no “true value” of nonverbal synchrony to which we can compare the output scores of the applied TSAMs. In the future it may be beneficial to create a database that includes sequences with and without nonverbal synchrony. To date, only two brief studies exist that have compared output scores with simulated (Altmann, 2013) or human-rated (Paxton & Dale, 2013b) nonverbal synchrony. Therefore, extensive studies will be needed. Additionally, the criterion to which the results of an algorithms are compared is important: If the aim is to measure an aspect of synchrony that is associated with interpersonal problems, the results of algorithms that correlate highly with interpersonal problems are best. If the aim is to measure a synchrony facet that is similar to human-rated synchrony, human-rated sequences will be needed as a criterion against which to compare the algorithms. In addition, there have also been attempts to define synchrony empirically in comparison to pseudosynchrony. Therefore, the choice of the best-suited algorithms is inherently connected to the criterion.

In this context, however, it should be emphasized that the results of the TSAMs presented are dependent on the parameter settings. These include, for example, the degree of

smoothing, transformation, and bandwidth. In the present study, standard settings from the literature and equal parameter settings were used. The results were more consistent using the parameter settings from the literature. Therefore, future studies should investigate which parameter settings really are best suited to optimally capture nonverbal synchrony. Since, for example, bandwidth has a high impact on results, which is why we do not recommend the direct comparison of results that were conducted using different bandwidths, even when identical TSAMs were used. Nevertheless, in line with Delaherche and Chetouani (2010), we recommend the application of local methods, because they are more likely to cope with statistical challenges such as nonstationarity, and also take zeitgeber changes or a changing time lag into account.

With respect to the present study, it can be said that  $WCLC_F$  or  $WCLR$  can be used to assess the strength of synchrony in predefined intervals or the frequency of synchrony in an interaction.  $CLC$ ,  $WCLC_{S1}$ , and  $WCLC_{S2}$  can be applied to measure the strength of synchrony of a total interaction.  $CLR$  measures are not recommended, because they are not comparable to the other TSAMs assessing nonverbal synchrony. If one is solely interested in the amount of synchrony (not in pacing, leading, or time-lag-related variables), the  $WCC$  (average) is a valid output score with which to estimate the amount of interrelatedness between two time series. Nevertheless, that a common construct of synchrony may be found for parameter settings from the literature, future studies should explicitly define and characterize the facet of synchrony they intend to measure.

### Strengths and limitations

A limitation is that the TSAM parameters were not systematically varied (e.g., in terms of time lag or bandwidth). This indicates that our results cannot be generalized to algorithms using other parameter settings. Since parameter settings are an important source of variance between the results of different TSAMs, we can only draw conclusions about the presented algorithms *with* their parameter settings. In addition, no correction of spurious correlations, apart from the correction for autocorrelation, was applied. Prewhitening can be performed to reduce bias due to autocorrelation, as was shown by R. T. Dean and Dunsmuir (2016). However, autocorrelation should have influenced all methods to a comparable extent (except for the methods that include autocorrelation within their model specification, such as  $CLR$  and  $WCLR$ ). Furthermore, we did not control for any further spurious correlations that might have been caused by randomly occurring synchrony. Different methods exist by which spurious correlations can be controlled for. One possibility is the use of surrogate/virtual pairs—that is, time series pairs that are split and randomly recombined (Louwerse et al., 2012; Moulder, Boker, Ramseyer, & Tschacher, 2018). Synchrony determined within

these surrogate pairs can be used as a baseline to evaluate the meaningfulness of genuine synchrony. Another opportunity to build a baseline is to randomly shuffle the data points (Louwerse et al., 2012) or windows (Ramseyer & Tschacher, 2010) of one time series. Altmann (2013) and Gottman and Ringland (1981) proposed a parametric test to solve this issue. Another possibility is to increase the cutoff for distinguishing between randomly occurring and meaningful synchrony (R. T. Dean & Dunsmuir, 2016). However, the lack of control for spurious correlations should also have a comparable effects on all TSAMs. Additionally, our study only compared algorithms assuming a linear relationship between the interacting persons. In further analyses, it may also be interesting to investigate other dependencies, by using spectral analysis, for example. Additionally, the inconsistent results with respect to the synchrony–outcome association can also be attributed to the limited sample size or to varying parameter settings. Valid parameter settings for each TSAM have to be examined in order to optimally map the corresponding synchrony construct. Overall, it can be assumed that a method effect exists, which explains the contrary results. For future studies, it will be important to specify exactly which method (including which parameter settings—e.g., for bandwidth) was used and which facet of synchrony is being addressed. Regarding the confirmatory factor analysis, it should be noted that typically exploratory and confirmatory analyses are run on different samples. However, due to our limited sample size, this was not possible, which should also be addressed in future studies.

One of the strengths of this study is that different algorithms were used to measure the linear dependency of two time series using an identical dataset. On the basis of the different algorithms and output scores, we systemized the approaches to different synchrony facets. Thereby, we used (1) the parameter settings recommended in the literature and (2) equal parameter settings across all TSAMs. In addition, the study draws attention to the fact that the choice of the algorithm and parameters is essential to the subsequent analyses (e.g., association analysis of nonverbal synchrony and therapeutic success). We applied the methods to a sample from a psychotherapeutic setting—that is, to real-world data—to ensure comparability of the results and their statistical challenges to previous studies (Galbusera et al., 2018; Paulick et al., 2018; Ramseyer & Tschacher, 2011). Thus, we are confident of the generalizability of our results to this field of research. However, a replication of our findings based on another dataset is advisable.

### Conclusion and future directions

The present study shows that confidence in the convergent validity of TSAMs based on the apparent fit of the construct and method is critical. In the literature, different methods have been presented to determine nonverbal synchrony. Thus, the facet of synchrony that is being measured depends on the

algorithm applied and the output score used. Therefore, comparing study results calculated with TSAMs that measure different facets is critical. However, with the parameter settings used in the literature, we found a superior latent factor synchrony for most of the TSAMs presented.

In the future, the construct to be measured should be explicitly defined, and the chosen method should also be checked to determine whether the defined construct is actually being measured. Further validation studies should be conducted to measure content validity. Human ratings of synchrony should be compared with the results obtained by these

methods. Furthermore, which method or parameter combinations lead to valid results will need to be investigated.

**Author note** This study was funded by the German Research Foundation (DFG project “Timing of Nonverbal Patient-Therapist-Interaction and Therapeutic Success of Social Phobic Patients (TIMPATHIN),” funding number GZ: STR 306/28-1 and LU 660/8-1) and used video recordings from the SOPHONET treatment study, funded by the German Federal Ministry of Education and Research (BMBF, FKZ 01GV0607). We cordially thank U. Willutzki, S. Herpertz, J. Hoyer, P. Joraschky, W. Hiller, E. Leibing, M. Beutel, and F. Leichenring for providing video recordings for the project. We thank K. Boyle, a native speaker of English, for proof reading.

## Appendix: Results—Maximum output score

The correlations of the maximum output scores are displayed in Table 6. Tables 7 and 8 show the correlations between maximum and average scores, respectively ratio scores (Table 7 with heterogeneous parameters, Table 8 with homogeneous parameter settings).

**Table 6** Correlations between *maximum* output scores of nonverbal synchrony assessed with various TSAMs

	CLC	CLR	WCC	WCLC <sub>S1</sub>	WCLC <sub>S2</sub>	WCLC <sub>F</sub>	WCLR	CRQA
CLC	.967*	.012	.088	.123	-.095	.065	.043	.155
CLR	.018	1	.100	-.067	.046	-.176	-.161	-.107
WCC	.073	.157	.595*	.126	-.076	.406*	.309	.411*
WCLC <sub>S1</sub>	.590*	-.081	.248	.255*	.014	.008	-.027	.193
WCLC <sub>S2</sub>	.170	.101	-.014	.165	.081	-.011	-.068	-.558*
WCLC <sub>F</sub>	.085	-.176	.293	.125	-.154	1	.779*	.306
WCLR	.094	-.161	.155	.142	-.170	.779*	1	.261
CRQA	.216	-.107	.355*	.215	-.190	.306	.261	1

$R^2$  values in the upper triangle (gray-shaded) are based on algorithms with exactly equal parameter settings, and correlations in the lower triangle are based on algorithms with the settings recommended from the literature. CLC = cross-lagged correlation, CLR = cross-lagged regression, WCC = windowed cross-correlation, WCLC = windowed cross-lagged correlation (WCLC<sub>S</sub>, with respect to the strength of synchrony; WCLC<sub>F</sub>, with respect to the frequency of synchrony), WCLR = windowed cross-lagged regression, CRQA = cross-recurrence quantification analysis. \* Significant correlations, Bonferroni-adjusted alpha level = .001

**Table 7** Correlations between the *average* output scores and the *maximum* output scores of nonverbal synchrony assessed with various TSAMs—Heterogeneous parameter settings

	CLC (av)	CLR (av)	WCC (av)	WCLC <sub>S1</sub> (av)	WCLC <sub>S2</sub> (av)	WCLC <sub>F</sub> (av)	WCLR (av)	CRQA (av)	WCLC <sub>F</sub> (ratio)	WCLR (ratio)
CLC (max)	.817*	.010	.169	.368	.419*	.109	.078	.216	.144	.161
CLR (max)	.013	.999*	-.101	.038	.095	-.029	-.007	-.105	-.152	-.150
WCC (max)	.023	.154	.515*	.450*	.389*	.433*	.366	.347	.364	.380*
WCLC <sub>S1</sub> (max)	.410*	-.080	.252	.596*	.502*	.162	.123	.219	.199	.186
WCLC <sub>S2</sub> (max)	.009	.104	-.437*	-.040	-.210	-.420*	-.433*	-.175	-.442*	-.448*
WCLC <sub>F</sub> (max)	.085	-.181	.506	.325	.369	.631	.626	.298	.437	.417
WCLR (max)	.203	-.167	.418*	.318	.378*	.573*	.569*	.254	.362	.355
CRQA (max)	.271	-.113	.752*	.457*	.567*	.616*	.613*	.999*	.780*	.789*
WCLC <sub>F</sub> (ratio)	.204	-.155	.880*	.443*	.550*	.753*	.767*	.769*	1	.987*
WCLR (ratio)	.239	-.155	.892*	.464*	.579*	.749*	.757*	.777*	.987*	1

CLC = cross-lagged correlation, CLR = cross-lagged regression, WCC = windowed cross-correlation, WCLC = windowed cross-lagged correlation (WCLC<sub>S</sub>, with respect to the strength of synchrony; WCLC<sub>F</sub>, with respect to the frequency of synchrony), WCLR = windowed cross-lagged regression, CRQA = cross-recurrence quantification analysis. \* Significant correlations, Bonferroni-adjusted alpha level = .001

**Table 8** Correlations between the *average* output scores and the *maximum* output scores of nonverbal synchrony assessed with various TSAMs—Homogeneous parameter settings: bandwidth 5 s, overlapping windows

	CLC (av)	CLR (av)	WCC (av)	WCLC <sub>S1</sub> (av)	WCLC <sub>S2</sub> (av)	WCLC <sub>F</sub> (av)	WCLR (av)	CRQA (av)	WCLC <sub>F</sub> (ratio)	WCLR (ratio)
CLC (max)	.687*	.004	.071	.183	.111	.056	.025	.157	.101	.115
CLR (max)	.010	.999*	-.113	.072	-.020	-.029	-.007	-.105	-.152	-.150
WCC (max)	.037	.095	.431*	.266	.400*	.503*	.479*	.405*	.448*	.448*
WCLC <sub>S1</sub> (max)	.059	-.067	.207	.283	.211	.141	.113	.199	.132	.143
WCLC <sub>S2</sub> (max)	-.274	.050	-.386*	.047	-.216	-.126	-.147	-.566*	-.374*	-.368
WCLC <sub>F</sub> (max)	.037	-.181	.410*	.289	.463*	.631*	.626*	.298	.437*	.417*
WCLR (max)	.181	-.167	.359	.272	.428*	.573*	.569*	.254	.362	.355
CRQA (max)	.273	-.113	.778*	.167	.694*	.616*	.613*	.999*	.780*	.789*
WCLC <sub>F</sub> (ratio)	.193	-.155	.874*	.137	.733*	.753*	.767*	.769*	1	.987*
WCLR (ratio)	.236	-.155	.881*	.146	.749*	.749*	.757*	.777*	.987*	1

CLC = cross-lagged correlation, CLR = cross-lagged regression, WCC = windowed cross-correlation, WCLC = windowed cross-lagged correlation (WCLC<sub>S</sub>, with respect to the strength of synchrony; WCLC<sub>F</sub>, with respect to the frequency of synchrony), WCLR = windowed cross-lagged regression, CRQA = cross-recurrence quantification analysis. \* Significant correlations, Bonferroni-adjusted alpha level = .001

## References

- Allsop, J. S., Vaitkus, T., Marie, D., & Miles, L. K. (2016). Coordination and collective performance: Cooperative goals boost interpersonal synchrony and task outcomes. *Frontiers in Psychology*, 7, 1462. <https://doi.org/10.3389/fpsyg.2016.01462>
- Altmann, U. (2010). Interrater-Reliabilität = 1 in Videostudien? Automatisierte Erhebung von Nonverbalität in einem Experiment zur Kooperation von Schülern [Automated coding of nonverbal behavior in an experiment on the cooperation of students]. *Beiträge zur Erziehungswissenschaftliche Forschung–Nachhaltige Bildung*, 5, 261–267.
- Altmann, U. (2011). Investigation of movement synchrony using windowed cross-lagged regression. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, & A. Nijholt (Eds.), *Analysis of verbal and nonverbal communication and enactment: The processing issues* (pp. 335–345). Berlin, Germany: Springer. [https://doi.org/10.1007/978-3-642-25775-9\\_31](https://doi.org/10.1007/978-3-642-25775-9_31)
- Altmann, U. (2013). *Synchronisation nonverbalen Verhaltens: Weiterentwicklung und Anwendung zeitreihenanalytischer Identifikationsverfahren*. Berlin, Germany: Springer.
- Altmann, U., Schoenherr, D., Paulick, J., Deisenhofer, A.-K., Schwartz, B., Rubel, J., ... Strauss, B. M. (2018). *Timing of nonverbal patient-therapist interaction and therapeutic success of social phobic patients*. Manuscript in preparation.
- Altmann, U., Schoenherr, D., Paulick, J., Knitter, L., Worrack, S., Schiefele, A.-K., ... Strauss, B. M. (submitted). *Introduction, practical guide, and validation study for measuring body movements using motion energy analysis*. Manuscript submitted for publication.
- Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, 16, 814–819. <https://doi.org/10.1111/j.1467-9280.2005.01619.x>
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4, 561–571. <https://doi.org/10.1001/archpsyc.1961.01710120031004>
- Bernieri, F. J. (1988). Coordinated movement and rapport in teacher–student interactions. *Journal of Nonverbal Behavior*, 12, 120–138. <https://doi.org/10.1007/BF00986930>
- Bernieri, F. J., & Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 401–432). New York: Cambridge University Press.
- Bilakhia, S., Petridis, S., Nijholt, A., & Pantic, M. (2015). The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern Recognition Letters*, 66, 52–61. <https://doi.org/10.1016/j.patrec.2015.03.005>
- Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7, 338–355. <https://doi.org/10.1037/1082-989X.7.3.338>
- Bouthillier, D., Julien, D., Dubé, M., Bélanger, I., & Hamelin, M. (2002). Predictive validity of adult attachment measures in relation to emotion regulation behaviors in marital interactions. *Journal of Adult Development*, 9, 291–305. <https://doi.org/10.1023/A:1020291011587>
- Brown, T. A., Barlow, D. H., & Di Nardo, P. A. (1994). *Anxiety disorders interview schedule for DSM-IV (ADIS-IV): Client interview schedule*. Albany, NY: Graywind.
- Cappella, J. N. (1996). Dynamic coordination of vocal and kinesic behavior in dyadic interaction: Methods, problems, and interpersonal outcomes. In J. H. Watt & C. A. VanLear (Eds.), *Dynamic patterns in communication processes* (pp. 353–386). Thousand Oaks, CA, US: Sage.
- Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, 76, 893–910. <https://doi.org/10.1037/0022-3514.76.6.893>
- Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: An R package. *Frontiers in Psychology*, 5, 510. <https://doi.org/10.3389/fpsyg.2014.00510>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302. <https://doi.org/10.1037/h0040957>
- Davis, T. J., Kay, B. A., Kondepudi, D., & Dixon, J. A. (2016). Spontaneous interentity coordination in a dissipative structure. *Ecological Psychology*, 28, 23–36. <https://doi.org/10.1080/10407413.2016.1121737>
- Dean, D. J., Samson, A. T., Newberry, R., & Mittal, V. A. (2018). Motion energy analysis reveals altered body movement in youth at risk for psychosis. *Schizophrenia Research*, 200, 35–41. <https://doi.org/10.1016/j.schres.2017.05.035>
- Dean, R. T., & Dunsmuir, W. T. M. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, 48, 783–802. <https://doi.org/10.3758/s13428-015-0611-2>
- Delaherche, E., & Chetouani, M. (2010, October). *Multimodal coordination: Exploring relevant features and measures*. Paper presented at the 2nd international Workshop on Social Signal Processing, Florence, Italy.
- Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3, 349–365. <https://doi.org/10.1109/T-AFFC.2012.12>
- Duran, N. D., & Fusaroli, R. (2017). Conversing with a devil’s advocate: Interpersonal coordination in deception and disagreement. *PLoS ONE*, 12, e0178140. <https://doi.org/10.1371/journal.pone.0178140>
- Dutschke, L. L., Stegmayer, K., Ramseyer, F., Bohlhalter, S., Vanbellinghen, T., Strik, W., & Walther, S. (2018). Gesture impairments in schizophrenia are linked to increased movement and prolonged motor planning and execution. *Schizophrenia Research*, 200, 42–49. <https://doi.org/10.1016/j.schres.2017.07.012>
- Feldman, R. (2007). Parent–infant synchrony and the construction of shared timing: Physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry*, 48, 329–354. <https://doi.org/10.1111/j.1469-7610.2006.01701.x>
- Fogel, A. (1993). Two principles of communication: Co-regulation and framing. In J. Nadel & L. Camaioni (Eds.), *New perspectives in early communicative development* (pp. 9–22). London, UK: Routledge.
- Galbusera, L., Finn, M. T., & Fuchs, T. (2018). Interactional synchrony and negative symptoms: An outcome study of body-oriented psychotherapy for schizophrenia. *Psychotherapy Research*, 28, 457–469. <https://doi.org/10.1080/10503307.2016.1216624>
- Gottman, J. M., & Ringland, J. T. (1981). The analysis of dominance and bidirectionality in social development. *Child Development*, 52, 393–412. <https://doi.org/10.2307/1129157>
- Grammer, K., Honda, M., Juette, A., & Schmitt, A. (1999). Fuzziness of nonverbal courtship communication unblurred by motion energy detection. *Journal of Personality and Social Psychology*, 77, 487–508. <https://doi.org/10.1037/0022-3514.77.3.487>
- Grammer, K., Kruck, K. B., & Magnusson, M. S. (1998). The courtship dance: Patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Nonverbal Behavior*, 22, 3–29. <https://doi.org/10.1023/A:1022986608835>
- Harrist, A. W., & Waugh, R. M. (2002). Dyadic synchrony: Its structure and function in children’s development. *Developmental Review*, 22, 555–592. [https://doi.org/10.1016/S0273-2297\(02\)00500-2](https://doi.org/10.1016/S0273-2297(02)00500-2)

- Horowitz, L. M., Strauss, B., Thomas, A., & Kordy, H. (2016). *IIP-D Inventar zur Erfassung Interpersonaler Probleme—Deutsche Version*. [Inventory for the assessment of interpersonal problems—German version] (3rd ed.). Göttingen, Germany: Hogrefe.
- Kato, T., Takahashi, E., Sawada, K., Kobayashi, N., Watanabe, T., & Ishh, T. (1983). A computer analysis of infant movements synchronized with adult speech. *Pediatric Research*, *17*, 625–628. <https://doi.org/10.1203/00006450-198308000-00004>
- Kirchmann, H., Fenner, A., & Strauß, B. (2007). Konvergenz des Erwachsenen-Bindungsprototypen-Ratings (EBPR) mit verschiedenen Selbstbeschreibungsinstrumenten zur Erfassung von Bindungsmerkmalen. *PPmP—Psychotherapie: Psychosomatik-Medizinische Psychologie*, *57*, 334–342. <https://doi.org/10.1055/s-2007-970860>
- Kleinbub, K. R., & Ramseyer, F. (2018). rMEA: Synchrony in motion energy analysis (MEA) time-series (R package version 1.0.0.9012). <https://cran.r-project.org/web/packages/rMEA/index.html>
- Kupper, Z., Ramseyer, F., Hoffmann, H., Kalbermatten, S., & Tschacher, W. (2010). Video-based quantification of body movement during social interaction indicates the severity of negative symptoms in patients with schizophrenia. *Schizophrenia Research*, *121*, 90–100. <https://doi.org/10.1016/j.schres.2010.03.032>
- Kupper, Z., Ramseyer, F., Hoffmann, H., & Tschacher, W. (2015). Nonverbal synchrony in social interactions of patients with schizophrenia indicates socio-communicative deficits. *PLoS ONE*, *10*, e0145882. <https://doi.org/10.1371/journal.pone.0145882>
- Leichsenring, F., Salzer, S., Beutel, M. E., Herpertz, S., Hiller, W., Hoyer, J., ... Poehlmann, E. (2013). Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: A multicenter randomized controlled trial. *American Journal of Psychiatry*, *170*, 759–767. <https://doi.org/10.1176/appi.ajp.2013.12081125>
- Leichsenring, F., Salzer, S., Beutel, M. E., Herpertz, S., Hiller, W., Hoyer, J., ... Poehlmann, E. (2014). Long-term outcome of psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder. *American Journal of Psychiatry*, *171*, 1074–1082. <https://doi.org/10.1176/appi.ajp.2014.13111514>
- Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science*, *36*, 1404–1426. <https://doi.org/10.1111/j.1551-6709.2012.01269.x>
- Manes, S., Nodop, S., Altmann, U., Gawlytta, R., Dinger, U., Dymel, W., ... Petrowski, B. (2016). Social anxiety as a potential mediator of the association between attachment and depression. *Journal of Affective Disorders*, *205*, 264–268. <https://doi.org/10.1016/j.jad.2016.06.060>
- Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*, 299–307. [https://doi.org/10.1016/S0375-9601\(02\)01170-2](https://doi.org/10.1016/S0375-9601(02)01170-2)
- Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*, 237–329. <https://doi.org/10.1016/j.physrep.2006.11.001>
- Matlab (2016). The MathWorks, Inc., Natick, Massachusetts, United States.
- McGarva, A. R., & Warner, R. M. (2003). Attraction and social coordination: Mutual entrainment of vocal activity rhythms. *Journal of Psycholinguistic Research*, *32*, 335–354. <https://doi.org/10.1023/a:1023547703110>
- Mennin, D. S., Fresco, D. M., Heimberg, R. G., Schneier, F. R., Davies, S. O., & Liebowitz, M. R. (2002). Screening for social anxiety disorder in the clinical setting: Using the Liebowitz Social Anxiety Scale. *Journal of Anxiety Disorders*, *16*, 661–673. [https://doi.org/10.1016/S0887-6185\(02\)00134-2](https://doi.org/10.1016/S0887-6185(02)00134-2)
- Moulder, R. G., Boker, S. M., Ramseyer, F., & Tschacher, W. (2018). Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000172>
- Nagaoka, C., & Komori, M. (2008). Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. *IEICE Transactions on Information and Systems*, *91*, 1634–1640. <https://doi.org/10.1093/ietisy/e91-d.6.1634>
- Nelson, A., Grahe, J., Ramseyer, F., & Serier, K. (2014). Psychological data from an exploration of the rapport/synchrony interplay using motion energy analysis. *Journal of Open Psychology Data*, *2*, e5. <https://doi.org/10.5334/jopd.ae>
- Paulick, J., Deisenhofer, A.-K., Ramseyer, F., Tschacher, W., Boyle, K., Rubel, J., & Lutz, W. (2018). Nonverbal synchrony: A new approach to better understand psychotherapeutic processes and drop-out. *Journal of Psychotherapy Integration*, *28*, 367–384. <https://doi.org/10.1037/int0000099>
- Paxton, A. (2015). *Coordination: Theoretical, methodological, and experimental perspectives* (Doctoral dissertation). University of California, Merced.
- Paxton, A., & Dale, R. (2013a). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, *66*, 2092–2102. <https://doi.org/10.1080/17470218.2013.853089>
- Paxton, A., & Dale, R. (2013b). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, *45*, 329–343. <https://doi.org/10.3758/s13428-012-0249-2>
- Ramseyer, F. (2011). Nonverbal synchrony in psychotherapy: Embodiment at the level of the dyad. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 193–207). Exeter, UK: Andrews.
- Ramseyer, F. (2013, June). *Synchronized movement in social interaction*. Paper presented at the 2013 Inputs–Outputs Conference: An Interdisciplinary Conference on Engagement in HCI and Performance, Brighton, UK.
- Ramseyer, F., & Tschacher, W. (2010). Nonverbal synchrony or random coincidence? How to tell the difference. In A. Esposito (Ed.), *Development of multimodal interfaces: Active listening and synchrony* (pp. 182–196). Berlin, Germany: Springer.
- Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology*, *79*, 284–295. <https://doi.org/10.1037/a0023419>
- Ramseyer, F., & Tschacher, W. (2016). Movement coordination in psychotherapy: Synchrony of hand movements is associated with session outcome. A single-case study. *Nonlinear Dynamics, Psychology, and Life Sciences*, *20*, 145–166.
- Robinson, J. W., Herman, A., & Kaplan, B. J. (1982). Autonomic responses correlate with counselor–client empathy. *Journal of Counseling Psychology*, *29*, 195–198. <https://doi.org/10.1037/0022-0167.29.2.195>
- Roisman, G. I., Holland, A., Fortuna, K., Fraley, R. C., Clausell, E., & Clarke, A. (2007). The Adult Attachment Interview and self-reports of attachment style: An empirical rapprochement. *Journal of Personality and Social Psychology*, *92*, 678–697. <https://doi.org/10.1037/0022-3514.92.4.678>
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. <https://doi.org/10.1016/j.jrp.2013.05.009>
- Shockley, K. (2005). Cross recurrence quantification of interpersonal postural activity. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 142–177). Arlington, VA: National Science Foundation.
- Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*, 305–319. <https://doi.org/10.1111/j.1756-8765.2009.01021.x>
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*, 437–450. [https://doi.org/10.1016/S0169-2070\(00\)00065-0](https://doi.org/10.1016/S0169-2070(00)00065-0)

- Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, 6, 35–39. <https://doi.org/10.1177/875647939000600106>
- Tronick, E. D., Als, H., & Brazelton, T. B. (1977). Mutuality in mother–infant interaction. *Journal of Communication*, 27, 74–79. <https://doi.org/10.1111/j.1460-2466.1977.tb01829.x>
- Tschacher, W., Rees, G. M., & Ramseyer, F. (2014). Nonverbal synchrony and affect in dyadic interactions. *Frontiers in Psychology*, 5, 1323. <https://doi.org/10.3389/fpsyg.2014.01323>
- Watanabe, T. (1983). A study of motion–voice synchronization. *Bulletin of JSME*, 26, 2244–2250. <https://doi.org/10.1299/jsme1958.26.2244>
- Watanabe, T. (1987). A quantitative analyzing method for the synchrony between a speaker's voice and a listener's nodding in interpersonal communication. *JSME International Journal*, 30, 2040–2043. <https://doi.org/10.1299/jsme1987.30.2040>
- Wenzel, A., Graff-Dolezal, J., Macho, M., & Brendle, J. R. (2005). Communication and social skills in socially anxious and nonanxious individuals in the context of romantic relationships. *Behaviour Research and Therapy*, 43, 505–519. <https://doi.org/10.1016/j.brat.2004.03.010>
- Wiltshire, T. (2015). Team interaction dynamics during collaborative problem solving (Doctoral dissertation). University of Central Florida, Orlando, FL.
- Wittchen, H.-U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). *SKID-I: Strukturiertes klinisches Interview für DSM-IV*. Göttingen, Germany: Hogrefe.