Check for updates

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021), 74, 86–109
© 2020 The British Psychological Society
www.wileyonlinelibrary.com

Editor's Choice

Accounting for auto-dependency in binary dyadic time series data: A comparison of model- and permutation-based approaches for testing pairwise associations

Nadja Bodner[1]*, Francis Tuerlinckx[1], Guy Bosmans[2] and Eva Ceulemans[1]

[1]Quantitative Psychology and Individual Differences Research Group, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium
[2]Clinical Psychology Research Group, Faculty of Psychology and Educational Sciences, KU Leuven, Belgium

Many theories have been put forward on how people become synchronized or co-regulate each other in daily interactions. These theories are often tested by observing a dyad and coding the presence of multiple target behaviours in small time intervals. The sequencing and co-occurrence of the partners' behaviours across time are then quantified by means of association measures (e.g., kappa coefficient, Jaccard similarity index, proportion of agreement). We demonstrate that the association values obtained are not easy to interpret, because they depend on the marginal frequencies and the amount of auto-dependency in the data. Moreover, often no inferential framework is available to test the significance of the association. Even if a significance test exists (e.g., kappa coefficient) auto-dependencies are not taken into account, which, as we will show, can seriously inflate the Type I error rate. We compare the effectiveness of a model- and a permutation-based framework for significance testing. Results of two simulation studies show that within both frameworks test variants exist that successfully account for auto-dependency, as the Type I error rate is under control, while power is good.

1. Introduction

Human emotions, thoughts and behaviours often arise in the context of interpersonal relationships (Tiedens & Leach, 2004). Many theories have been put forward on how people influence each other in daily interactions, advancing notions such as synchronization, co-regulation and contagion (Butler, 2011; Butler & Randall, 2013; Sels, Ceulemans, & Kuppens, 2018a), and on how these interaction characteristics affect psychological well-being. For instance, dyadic parent–child interactions are considered one of the basic building blocks that drive development and adaptation. In these interactions, children learn how to express, interpret and regulate their emotions

*Correspondence should be addressed to Nadja Bodner, Quantitative Psychology and Individual Differences Research Group, Faculty of Psychology and Educational Studies, KU Leuven, Tiensestraat 102 – Box 3713, 3000 Leuven, Belgium (email: nadja.bodner@kuleuven.be).

DOI:10.1111/bmsp.12222

(Cassidy, 1994; Hunter, Hessler, & Fainsilber Katz, 2008; Schwartz, Sheeber, Dudgeon, & Allen, 2012). Not surprisingly, the quality of these interactions has been related to developmental psychopathology (e.g., Allen & Sheeber, 2008; Cummings, Davies, & Campbell, 2000; Hollenstein, Allen, & Sheeber, 2016; Kobak & Bosmans, 2019; Lichtwarck-Aschoff, Kunnen, & van Geert, 2009). As another example, interpersonal emotion dynamics in intimate relationships have been related to both individual well-being and relationship quality (e.g., Randall & Schoebi, 2015; Sels, Ceulemans, Bulteel, & Kuppens, 2016).

To get a clear picture of the mechanisms at work, many researchers observe dyads while they are performing a lab task, such as playing together or discussing a conflictual topic (e.g., Bodner, Kuppens, Allen, Sheeber, & Ceulemans, 2018; Sels, Ceulemans, & Kuppens, 2018b; Van keer, 2017). The partners' behaviours and emotions are then tracked across time. To this end, researchers very often code the presence or absence of behavioural expressions in small time intervals, resulting in a binary time series for each behaviour. Next to relative frequencies and duration of the separate behaviours (Hollenstein et al., 2016), such data allow one to investigate synchronicity (e.g., 'Do behaviours co-occur?') or temporal sequencing (e.g., 'Do specific partner behaviours follow upon each other in that behaviour X in time interval $t - 1$ elicits behaviour Y in time interval t?') between partners (Bodner et al., 2018; Van keer, 2019).

To assess synchronicity and temporal sequencing, pairwise association measures or contingency indices are computed between the time series (Bakeman & Quera, 2011). A wide range of association measures for binary data have been proposed and used in dyadic research: the proportion of agreement and its complement the Hamming distance, odds ratio, log odds ratio, Yule's Q (Bakeman, McArthur, & Quera, 1996; Bakeman & Quera, 2011) and the risk difference (Lloyd, Yoder, Tapp, & Staubitz, 2016), the recurrence rate of cross-recurrence quantification analysis (Main, Paxton, & Dale, 2016), the Jaccard index (Bodner et al., 2018) and normalized forms thereof (Van keer et al., 2019), Cohen's kappa and Wampold's transformed kappa (Bakeman et al., 1996; Holloway, Wampold, & Nelson, 1990), and the phi coefficient (Bakeman et al., 1996).

As we will demonstrate, interpreting the association values that are obtained for a specific measure is not straightforward, for two reasons. First, many association measures partly reflect the relative frequency of the behaviours under study. This implies that comparing the association values obtained across different time series pairs is complex, because higher values may result from differences in relative frequencies or from differences in association strength. Second, the values are also affected by the auto-dependency within each time series. This is a conceptual issue that is well known in time series analysis of continuous data (e.g., Lütkepohl, 2005) but seems hitherto to have been overlooked in this context. A natural solution to the interpretation question is to conduct a significance test that indicates whether a value is clearly higher than would be expected by chance, given the relative frequency and the auto-dependency within the data. However, for many of the measures, either no procedure has been proposed to test its significance, or the proposed test is only applicable under stringent assumptions (e.g., independence of the binary scores across time intervals).

The aim of this paper is to investigate the pros and cons of two significance testing approaches that can account for relative frequency and auto-dependency. Specifically, we propose a model-based significance test that applies the principles of simulation-based testing and parametric bootstrap methods (Efron & Tibshirani, 1993) and compare it to a permutation-based approach that was recently proposed by Moulder, Boker, Ramseyer, and Tschacher (2018) and involves shuffling the original data. We also aim to showcase

the importance of accounting for auto-dependency, in that ignoring it can lead to inflated Type I probabilities. To this end, we derive two variants for each testing approach, where one variant incorporates auto-dependency and the other ignores it. Specifically, for the model-based approach, we distinguish the Model-AD test (model-based test accounting for auto-dependency) and Model-No test (model-based test ignoring auto-dependency). Similarly, for the permutation-based approach we discern the Permut-AD and Permut-No tests. All four tests are very flexible and can be applied to whichever association measure one prefers. To demonstrate this, we study three association measures: Cohen's kappa, because of its popularity; the proportion agreement, because it is a basic building block of many agreement measures (e.g., Cohen's kappa); and the Jaccard index, because this measure has been applied in recent papers on dyadic interactions (Bodner et al., 2018; Van keer et al., 2019). This paper squarely focuses on the aims listed above, which target the performance of the compared tests within each association measure. Therefore, fleshing out the interpretational differences between the three considered association measures or giving advice on which association measure to use in which situation falls beyond the scope of this paper.

The remainder of this paper is organized as follows. First, by means of a real data example, we introduce the notation that we will use in this paper. Second, we recapitulate how to compute the three association measures and demonstrate that they are impacted by the relative frequency and auto-dependency. Next, we introduce the four significance tests. We then report the results of two simulation studies, one investigating the Type I error rate of the tests and one focusing on power. Finally, we discuss our findings and some future directions.

2. Method

2.1. Data and notation, with a real data example

To illustrate why interpreting pairwise association measures for binary data is challenging when applied to time series, we will consider two dyads from a data set on the presence and absence of affective expressions of three family members (mother, father, adolescent) during a family interaction (Allen, Kuppens, & Sheeber, 2012; Sheeber et al., 2012). The second-to-second binary anger expression data of two mother–adolescent dyads are visualized in Figure 1 for 100 seconds of the interaction. The scores on the anger expression variables of mother and adolescent, which we will generally denote by X and Y, are represented by a vertical tick, where low-level ticks indicate absence and high-level ones presence.

2.1.1. Relative frequency

A first important characteristic of each time series is the relative frequency with which X and Y are present, denoted by p_1^X and p_1^Y.[1] For the first dyad, p_1^{Mom} equals .05 and $p_1^{\text{Ado}} = .16$, while in the second dyad p_1^{Mom} amounts to .29 and $p_1^{\text{Ado}} = .40$.

2.1.2. Auto-dependency

A second important characteristic is the auto-dependency of each variable (i.e., the dependency of the variable score at time point t on the variable score at time point $t - 1$). It

[1] p_1^X is short for $P(X(t) = 1)$, where $X(t)$ denotes the score in interval t.

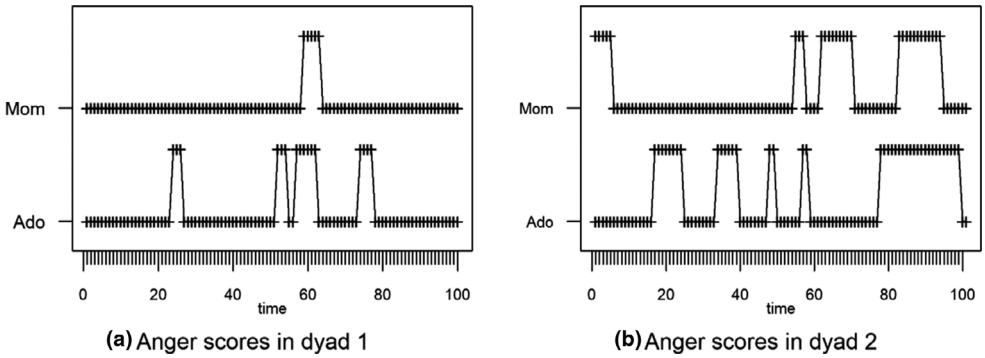


Figure 1. Visualization of the presence/absence of anger in a mother–adolescent interaction.Anger of the mother (Mom) and anger of the adolescent (Ado) across 100 one-second intervals is visualized, with the level of the line indicating absence (low) or presence (high) of the emotion.

is obvious from the plot that if individuals express anger in a certain time interval they are more likely also to show anger in the next interval, while they are less likely to be angry in a certain interval if they did not show anger in the previous one. Assuming time-homogeneity,[2] meaning that auto-dependency does not change over time, the auto-dependency of variable X can thus be assessed by computing two conditional probabilities $p^X_{1|1}$ and $p^X_{1|0}$,[3] where

$$p^X_{1|1} = \frac{\sum_t (X(t) = 1 \wedge X(t-1) = 1)}{\sum_t (X(t-1) = 1)},$$

$$p^X_{1|0} = \frac{\sum_t (X(t) = 1 \wedge \mathrm{X}(t-1) = 0)}{\sum_t (X(t-1) = 0)}.$$

Note that $\wedge$ is the logical AND operator. In dyad 1, $p^{\text{Mom}}_{1|1} = .80$ and $p^{\text{Mom}}_{1|0} = .01$, and $p^{\text{Ado}}_{1|1} = .75$ and $p^{\text{Ado}}_{1|0} = .05$; the large difference for both mother and adolescent shows that there is high auto-dependency. Similarly, in dyad 2, $p^{\text{Mom}}_{1|1} = .87$ and $p^{\text{Mom}}_{1|0} = .04$, and $p^{\text{Ado}}_{1|1} = .88$ and $p^{\text{Ado}}_{1|0} = .08$.

2.2. *Measuring associations*

As mentioned in the introduction, pairwise association measures are computed to quantify either the co-occurrence or the temporal sequencing of the behaviours X and Y across time. We will focus on co-occurrence here, because assessing sequencing simply boils down to computing co-occurrence on the lagged variables (e.g., Bodner et al., 2018). To quantify co-occurrence of X and Y, we will focus on the following three indices in this paper: the proportion of agreement, the Jaccard coefficient, and Cohen's kappa.

[2] The assumption of time homogeneity and stationarity is very frequently made in time series analysis. This assumption can be checked by applying nonparametric change point detection methods that scan the time series for changes in pre-specified statistics (Cabrieto, Tuerlinckx, Kuppens, Hunyadi, & Ceulemans, 2018).

[3] $p^X_{1|1}$ is short for $P(X(t) = 1|X(t–1) = 1)$ and $p^X_{1|0}$ for $P(X(t) = 1|X(t–1) = 0)$.

2.2.1. Three association measures

Proportion of agreement. The proportion of agreement is the simplest measure. It boils down to the proportion of time intervals in which behaviours X and Y are either both shown or not shown:

$$P_A(X,Y) = \frac{\sum_{t=1}^{N}(X(t)=1 \wedge Y(t)=1) + \sum_{t=1}^{N}(X(t)=0 \wedge Y(t)=0)}{N},$$

where N denotes the total number of time intervals. The proportion of agreement lies between 0 and 1. When neither behaviour X nor behaviour Y is shown across time, $P_A(X,Y)=1$. For our two dyads $P_A(\text{Mom},\text{Ado})$ amounts to .87 and .57, respectively.

Jaccard index. The Jaccard index (Jaccard, 1901, 1912) expresses the proportion of co-occurrence of two behaviours X and Y over the intervals in which at least one of X and Y is shown:

$$J(X,Y) = \frac{\sum_{t=1}^{N}(X(t)=1 \wedge Y(t)=1)}{N - \sum_{t=1}^{N}(X(t)=0 \wedge Y(t)=0)}.$$

The Jaccard index takes values between 0 and 1, but cannot be computed when neither X nor Y is shown in any time interval. The $J(\text{Mom},\text{Ado})$ values for our two dyads amount to .24 and .23.

Cohen's kappa. Cohen's kappa (Cohen, 1960) refines the proportion of agreement by correcting for the proportion of agreement P_E that is expected on the basis of the relative frequencies of X and Y alone, with $P_E = p_1^X p_1^Y + p_0^X p_0^Y$:

$$\kappa(X,Y) = \frac{P_A - P_E}{1 - P_E}.$$

Kappa takes values between –1 and 1, but cannot be computed when both X and Y always equal 0 (or 1) in all time intervals. For our two dyads, $\kappa(\text{Mom},\text{Ado})$ amounts to .33 and .07, respectively.

2.2.2. Impact of relative frequency and auto-dependency

This simple example already showcases that it is hard to interpret the values of a specific measure. Specifically, it is somewhat puzzling that the Jaccard values for the two dyads are very similar, whereas this does not hold for the other two measures. Indeed, if equal Jaccard values indicated similar association strength, we would expect the two kappa values and the two proportions of agreement to be almost equal as well (Note that we could have taken other dyads and then it might have happened that the kappa scores or the proportions of agreement were very similar, whereas the other measures yield different scores). We argue that this phenomenon is at least partly due to the impact of relative frequency and auto-dependency on the association values.

To better illustrate the impact of these characteristics, we simulated data and show (Figure 2 and Figure 3) that both characteristics affect the sampling distribution of the association measures. To generate Figure 2, we simulated 1,000 independent (X, Y) pairs, consisting of 1,000 time points, for a range of relative frequency values and calculated the association measures under study for each pair. Both variables have the same relative frequency ($p_1^X = p_1^Y$)and no auto-dependency ($p_1^X = p_{1|1}^X = p_{1|0}^X$ and $p_1^Y = p_{1|1}^Y = p_{1|0}^Y$). Although the variables X and Y are always generated independently and one might thus naively expect similar association values, this only holds for the kappa scores. Indeed, the expected proportions of agreement P_E(see kappa) increases when the relative frequency of X and Y becomes more extreme (e.g., if $p_1^X = p_1^Y$ is less than .1 or greater than .9, the observed proportion of agreement will be greater than .8). Moreover, higher relative frequencies lead to higher Jaccard values. Regarding the kappa values, the literature on inter-rater agreement (Bakeman, McArthur, Quera, & Robinson, 1997) points out that kappa scores are also related to relative frequencies if the variables are not independent.

The plots in Figure 3 show that introducing auto-dependency while generating X and Y independently further impacts the sampling distribution, in that it affects its shape and variance. Specifically, stronger auto-dependency leads to wider distributions, implying that an observed association value will point towards a diverse range of population values. This makes sense as high auto-dependencies lead to behaviour scores that linger longer (i.e., longer sequences of 1s or 0s). Therefore, occasional co-occurrence or co-absence of two behaviours will also linger longer, leading to more extreme association values. In addition to becoming wider, the sampling distributions are more asymmetric. For the Jaccard index and the proportion of agreement, this asymmetry is due to ceiling or floor effects when the expected value is close to the high (proportion of agreement) or low end of the scale (Jaccard index). Since the proportion of agreement is a building block of kappa, the associated sampling distribution also becomes skewed.

This impact of relative frequency and auto-dependency hampers the interpretation of the association values obtained, in that higher or lower scores on the same association measure may simply come into existence due to differences in these two characteristics.

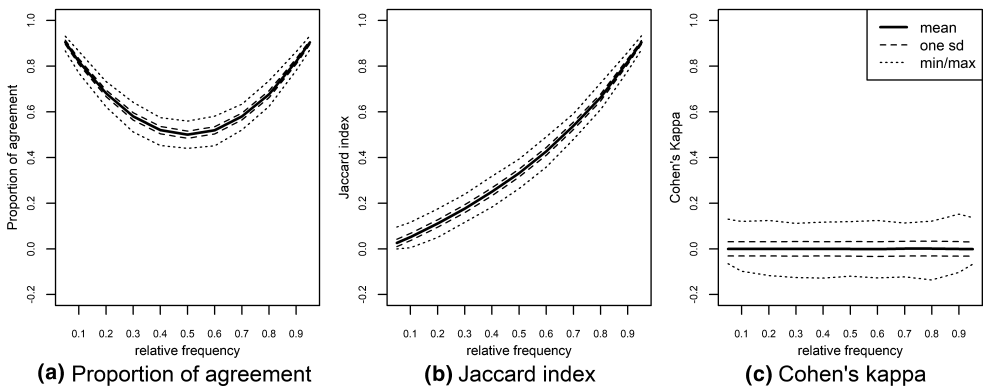


Figure 2. Values of (a) proportion of agreement, (b) Jaccard index, and (c) Cohen's kappa as a function of relative frequencies, with the different lines representing the mean (solid line), ±1 standard deviation (dashed lines) and minimum/maximum (dotted lines). The values are calculated based on simulated, independent variables X and Y ($p_1^X = p_1^Y$ range between .05 and .95; each pair contains 1,000 time points) without auto-dependency ($p_1^X = p_{1|1}^X = p_{1|0}^X = p_1^Y = p_{1|1}^Y = p_{1|0}^Y$).

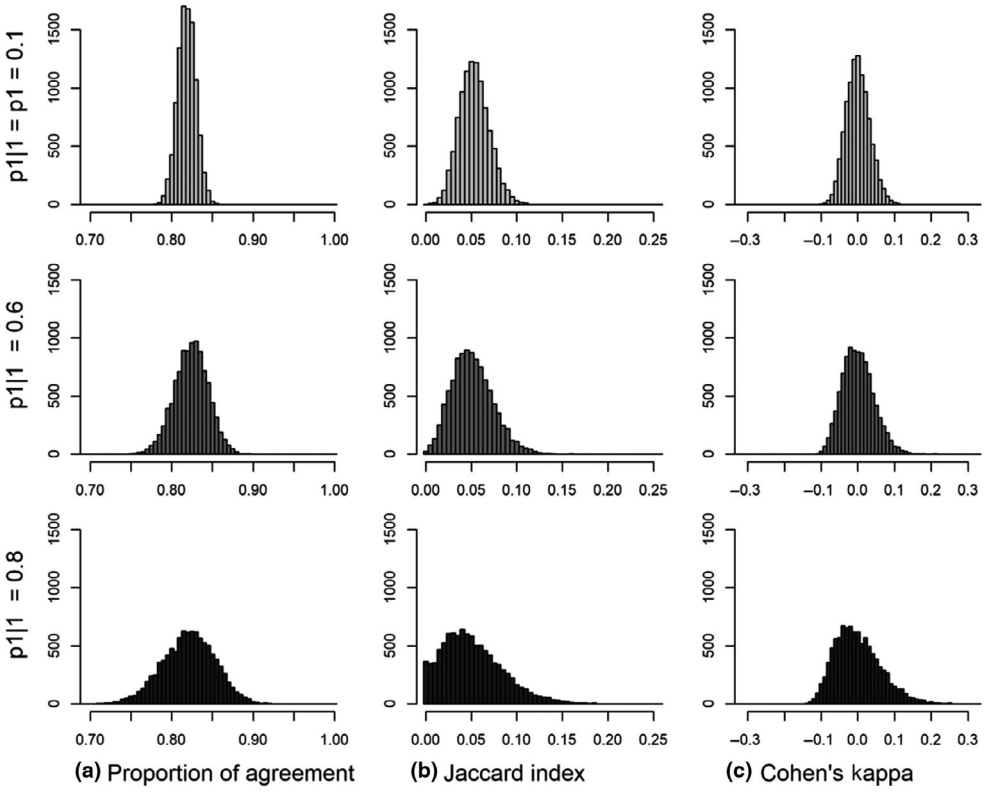


Figure 3. Distributions of (a) proportion of agreement, (b) Jaccard index and (c) Cohen's kappa as a function of auto-dependency. While the relative frequency of X and Y always equals .1, their auto-dependency varies across the rows, in that in the top row $p_{1|1} = p_{1|0} = .1$ (light grey), in the middle row, $p_{1|1} = .6$ and $p_{1|0} = .044$ (medium grey), and in the bottom row $p_{1|1} = .8$ and $p_{1|0} = .022$ (black). We generated 100 replicates of the pairwise independent variables X and Y (each consisting of 1,000 time points) and calculated the association values for all 10,000 possible pairs.

To investigate whether a certain obtained association value is higher than could be expected by chance, we could rely on significance tests. Yet, up to now, significance tests have only been proposed for some of the measures under consideration, and are subject to too strict assumptions. For instance, the significance test for Cohen's kappa that is provided in some statistical software packages assumes independence of observations and thus cannot handle auto-dependency of X and Y. Since we demonstrated that auto-dependency increases the variance of the sampling distribution, we expect this 'independent observations' kappa test to have an inflated Type I error rate.[4]

[4] Indeed, if we compare the results of the traditional kappa test with the results of the two-sided version of the model-based approaches on data that pertain to anger expressions of 598 mother–adolescent interactions, the traditional kappa test indicated 186 pairs as significant, the two-sided version of the Model-AD test revealed 55 significant associations only (54 of them were flagged as significant by both tests), and the Model-No test indicates 184 pairs as significant kappa (with 174 indicated as significant by both the Model-No and the traditional test). This suggests that the Type I error inflation would lead to almost 70% false positives.

2.3. Testing significance

To properly take into account the relative frequencies and auto-dependencies, we introduce and compare two approaches for testing pairwise association: a model-based procedure and a permutation-based one. For both approaches, we discuss a variant that accounts for auto-dependency and one that does not, which leaves us with four tests to compare. As these four tests have a lot in common, we first explain the general idea behind them. Then we introduce the four tests and point out the differences between them using the data of our first example dyad.

2.3.1. General idea of the four tests

The different steps of all four tests are visualized in Figure 4. First, the data (two time series X and Y) are collected (step 1) and the corresponding association value for this dyad (step 2) is calculated. In step 3, the original data (or certain properties thereof) are used to generate new data. Importantly, the replicates for X are generated independently of those of Y. Exactly how this step is performed differs across the four tests. In step 4, the association values of all possible pairs of the simulated X and Y series are computed. These values constitute the sampling distribution of the association measure under the assumption of independence. Finally, in step 5, we compare the observed association value and the 95th percentile of the sampling distribution. If the value exceeds the percentile, we consider the observed association to be significant at the 5% significance level.[5]

The four tests differ only in how the new data are generated in step 3. The model-based approach builds on the observed relative frequency and auto-dependency, in that it simulates new time series by drawing from a Bernoulli distribution, accounting for relative frequency (Model-No) or for both relative frequency and auto-dependency (Model-AD). The permutation-based approach shuffles segments of the originally observed time series (Moulder et al., 2018) where the segments can consist of single scores, thus removing auto-dependency (Permut-No; called 'data shuffling' by Moulder *et al.*), or of a number of consecutive scores, which partly retains auto-dependency (Permut-AD or segment shuffling).[6]

2.3.2. Model-AD

The data generation mechanism of this test builds on a Markov model in which the probability of showing a behaviour at time point t depends on whether this behaviour has been shown at time point $t - 1$. To estimate the parameters of this model, we quantify the relative frequency and the auto-dependency of the two behaviours (mother's anger $p_1^{\text{Mom}} = .05$, $p_{1|1}^{\text{Mom}} = .80$, $p_{1|0}^{\text{Mom}} = .01$; adolescent's anger $p_1^{\text{Ado}} = .16$, $p_{1|1}^{\text{Ado}} = .75$, $p_{1|0}^{\text{Ado}} = .05$). The scores on the replicates of X and Y are drawn time point by time point from different Bernoulli distributions until the simulated time series has the same length as the observed data. Specifically, $X(1)$ is sampled from a Bernoulli distribution with parameter p_1^X. Next,

[5] Note that we perform a one-sided test, as this corresponds with the research question underlying dyadic interaction studies: do the behaviours of both partners co-occur more than expected by chance?

[6] We also investigated the data sliding approach (Moulder et al., 2018) where one time series is cut at a random time point and both resulting segments are concatenated in a different order. However, this approach performed worse (especially with the short time series of 100 and 200 observations) than the data shuffling and segmented shuffling approaches and was therefore excluded.

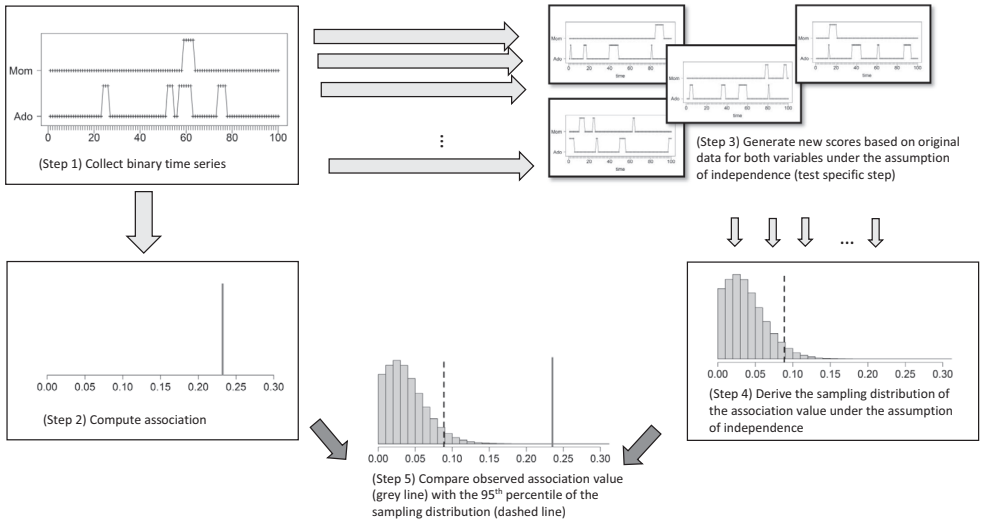


Figure 4. Schematic representation of the five steps of the test procedures.

all other scores $X(t)$ are drawn from a Bernoulli distribution with parameter $p^X_{1|0}$ if $X(t-1) = 0$ or from a Bernoulli distribution with parameter $p^X_{1|1}$ if $X(t-1) = 1$. To generate the replicates of Y the parameters p^Y_1, $p^Y_{1|0}$ and $p^Y_{1|1}$ are used. Note that the relative frequencies and auto-dependencies obtained are expected to vary somewhat across the time series generated.

The Model-AD test shows a significant result for the Jaccard index ($p = .0404$) and for Cohen's kappa ($p = .0248$), but not for proportion of agreement ($p = .2814$). The sampling distributions of the Model-AD test (Figure 5) are not symmetric and show that the Jaccard values and kappa values obtained very often equal 0. The latter is probably due to the low relative frequencies p^X_1 and p^Y_1 as well as to the high probabilities $p^X_{0|0}$ and $p^Y_{0|0}$, leading to long sequences of 0s, implying that the co-occurrence of 1s will be very low in many of the simulated pairs.

2.3.3. Model-No

This second model-based test is a simplified version of the Model-AD test. Auto-dependency is simply ignored in that $p^X_1 = p^X_{1|1} = p^X_{1|0}$ and $p^Y_1 = p^Y_{1|1} = p^Y_{1|0}$. This means that when generating data in step 3 all scores are drawn from one Bernoulli distribution with the parameter p^X_1 or p^Y_1.

For the example dyad, the p-values are well below those of the Model-AD test and equal $p = .0012$ for the Jaccard index, $p = .0009$ for Cohen's kappa, and $p = .0668$ for proportion of agreement. This finding naturally follows from the sampling distributions resulting from the Model-No test being narrower. We also see that the Jaccard and kappa scores equal 0 less frequently, because the high probabilities $p^X_{0|0}$ and $p^Y_{0|0}$ are not taken into account.

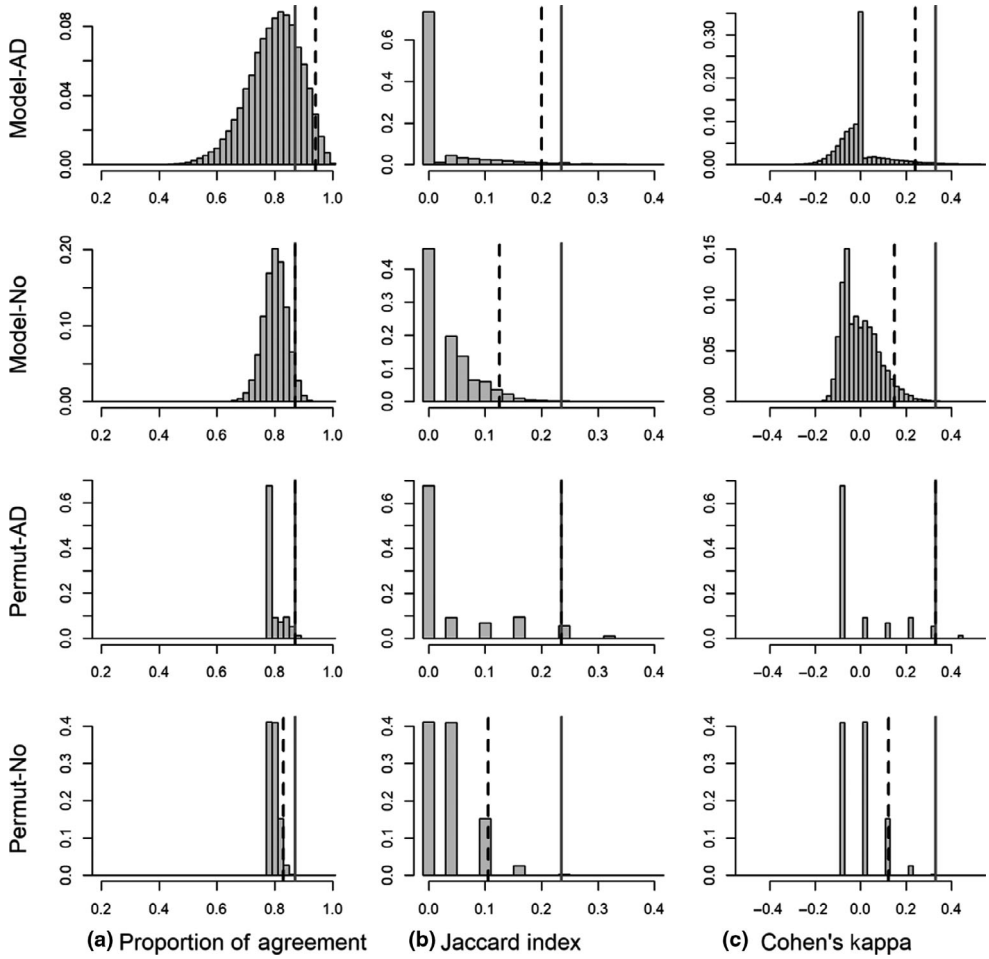


Figure 5. Significance test on the first dyad for (a) proportion of agreement, (b) Jaccard index, (c) Cohen's kappa, executed by the Model-AD test (first row), the Model-No test (second row), the Permut-AD test (third row) and the Permut-No test (last row). All sampling distributions are generated by calculating the association measures on all possible pairs of the 1,000 replicates for each variable. The *y*-axis shows proportions. The dashed line indicates the 95th percentile and the solid grey line the observed value. For the Model-AD and the Model-No tests the Jaccard index and Cohen's kappa are significant, for Permut-No all association measures are significant, and for the Permut-AD none is.

2.3.4. Permut-AD

When using the Permut-AD test, the two observed time series are each cut into 10 (almost) equally sized segments[7] of adjacent time points (for a similar idea in a cross-validation

[7] In empirical data the length of the time series is often not a multiple of 10, making it impossible to have equally sized segments. Which segments are longer is randomly chosen in each permutation and can thus differ for different permutations. Next, even if the length of the time series is a multiple of 10, we include some additional variation in the segment length by shortening two segments by 2, two by 1 and lengthening four of the other intervals accordingly. Which segments become longer and which shorter is again randomly chosen and may differ for each permutation.

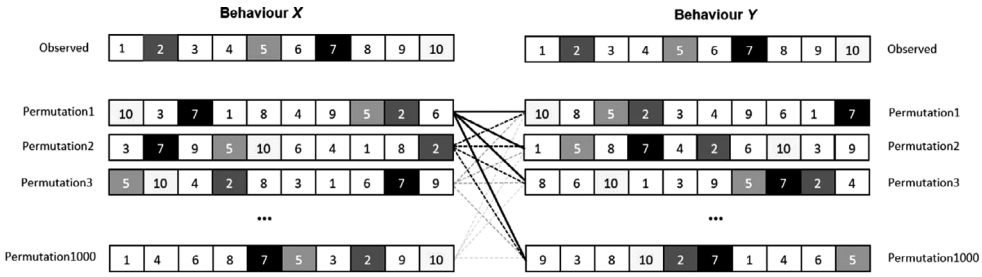


Figure 6. Schematic representation of the data generation step of the permutation approach. The two observed time series *X* and *Y* are first cut into 10 (almost) equally sized segments; we show the 2nd, 5th, 7th and 10th segment in different colours. These segments are randomly shuffled and rearranged, yielding 1,000 permuted versions of *X* and 1,000 permuted versions of *Y*. For each possible permuted *X* and permuted *Y* pair, the association value is calculated.

approach, see Bulteel, Mestdagh, Tuerlinckx, & Ceulemans, 2018). The order of these 10 segments is then randomly permuted for *X* and *Y* separately (see Figure 6). As a result, the relative frequencies are held constant (all 0s and 1s will be retained, but in a different order) and the auto-dependency in the original data is largely retained, because score sequences are partly moved together. Unlike Moulder *et al.*, we permute both *X* and *Y*, because this enables us to combine all permutations of *X* with all permutations of *Y*, as we did for the model-based test.

According to the Permut-AD test, none of the observed association values for the first dyad are significant ($p = .0641$ for proportion of agreement, $p = .0665$ for the Jaccard index and $p = .0668$ for Cohen's kappa). In Figure 5, we see that the sampling distribution includes six possible values only. This can be explained by noting that the mother expresses anger in five intervals only. Since the relative frequency is kept constant, these five 1s can co-occur with 0, 1, . . ., 5 expressions of the adolescent, yielding six possible values.

2.3.5. ***Permut****-No*

The Permut-No test is obtained by setting the segment size equal to 1. This shuffling procedure destroys most auto-dependency, while the relative frequency is kept constant. Therefore, we note in Figure 5 that the sampling distributions again pertain to six values only. However, the frequencies of these six values differ from those for Permut-AD, because the five 1s of the mother are not kept together. All *p*-values obtained are very small.

3. Simulation study I: Type I error

In the first simulation study, we investigated the Type I error rate of the four proposed tests. Ideally, this error rate should equal the significance level used. Based on Figure 3, we expected that Model-No and Permut-No would result in inflated Type I error rates, especially when the auto-dependency deviates strongly from the relative frequency.

3.1. Design

The relative frequencies and auto-dependencies of X and Y were set to the same values. We used a design that consisted of 60 cells, obtained by fully crossing the following three factors:

- Relative frequency of X (p_1^X) and Y (p_1^Y), .1, .2, .3;
- Auto-dependency of X ($p_{1|1}^X$) and Y ($p_{1|1}^Y$), .4, .5, .6, .7, .8 – note that, given $p_{1|1}$ and the relative frequencies p_1, the associated probability $p_{1|0}$ can be approximated as

$$p_{1|0}^X \approx \frac{p_1^X - p_1^X * p_{1|1}^X}{1 - p_1^X}$$

  (see the Appendix for a derivation);
- Length of time series: 100, 200, 500 or 1,000 time points.

Within each of the 60 cells, we generated 1,000 'observed', independent (X, Y) pairs, using the generation procedure described in step 3 of the Model-AD test. For each pair, we computed the value of the three association measures under study and tested the significance at the 5% significance level, using the four discussed tests (Model-AD, Model-No, Permut-AD, Permut-No).

3.2. Results

Figure 7 clearly shows that the two tests that ignore the auto-dependency (Model-No and Permut-No), have inflated Type I errors, with the amount of inflation related to the amount of auto-dependency. Furthermore, the Type I error rates of the tests that account for auto-dependency are not affected by the amount of auto-dependency, but differ somewhat across the three association measures. In particular, the Model-AD test has too low Type I error rates for the proportion of agreement. Note that we also inspected the effects of

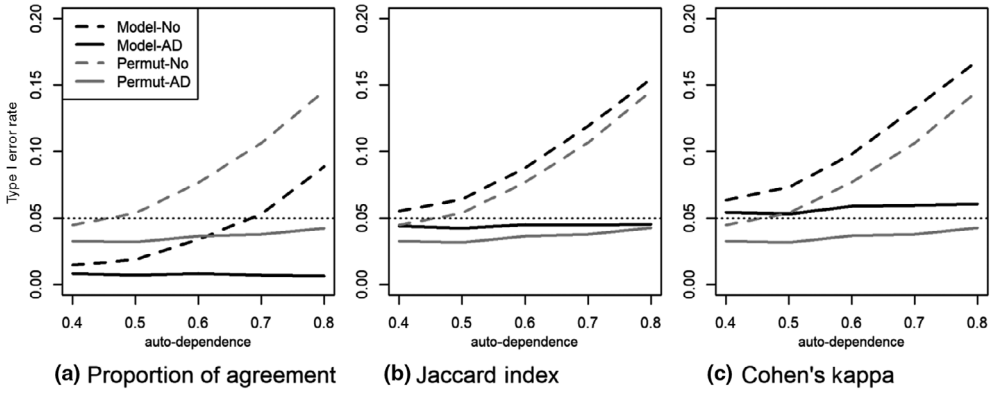


Figure 7. Type I error rate for time series with equal relative frequencies and auto-dependencies. Results are shown for (a) proportion of agreement, (b) Jaccard index and (c) Cohen's kappa as a function of auto-dependency and test (solid line = AD tests, dashed line = No tests, dotted line = nominal alpha level, black = model-based tests; grey = permutation-based tests). The results are averaged over number of time points and relative frequency.

relative frequency and time series length, but these are negligible (see the supplementary material).

Critical readers might note that we only investigated two time series that have the same relative frequency and auto-dependency and that remain stable throughout the whole interaction. These considerations leave us with two follow-up questions:

1. How do the four tests behave when the two time series have different relative frequencies or auto-dependencies?
2. How do the tests cope with stationarity violations (e.g., relative frequency or auto-dependency changes abruptly at a certain moment)?

To investigate these questions we ran follow-up experiments.

3.2.1. How do tests behave when the two time series have different relative frequency or auto-dependency?

We examined some settings where X and Y, both consisting of 500 time points, had unequal characteristics. The relative frequencies were set to $p_1^X = .05$.and $p_1^Y = .3$. Moreover, we combined five auto-dependency levels for X ($p^X_{1|1} = .4, .5, .6, .7$ and $.8$) with three auto-dependency levels for Y($p^Y_{1|1} = .5, .7$ and $.9$). Within each of the 15 design cells, 1,000 replications were generated and analysed with the four tests.

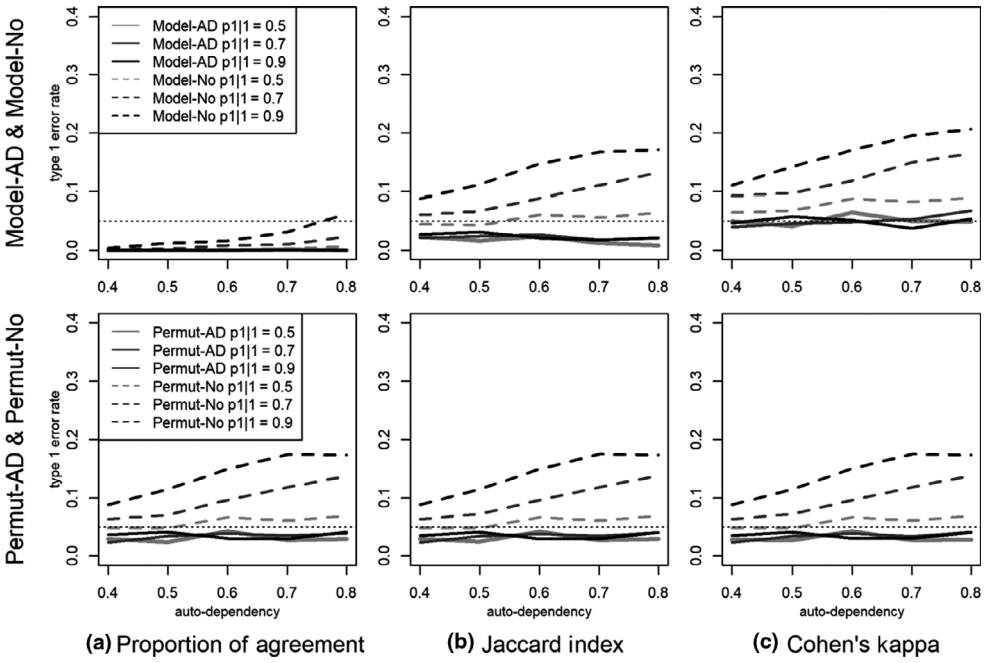


Figure 8. Type I error rate for two time series with different relative frequencies (p_1^X = .05; p_1^Y = .3) for (a) proportion of agreement, (b) Jaccard index, (c) Cohen's kappa for the model-based tests (first row) and the permutation-based tests second row). The solid lines refer to the AD variants and the dashed lines to the No variants. The colour of the lines represents the auto-dependency of Y (light grey,$p^Y_{1|1}$ = .5; middle grey, .7; black, .9). The auto-dependency of X is represented on the x-axis.

The results, shown in Figure 8, are in line with the results for the time series with equal properties. The Type I error rates of the No tests are too high due to the incorrect independence assumption. The Type I error rates of the AD test are close to the nominal alpha level, except for the proportion of agreement, where the Type I error rate is too low for the model-based test.

3.2.2. How do the tests cope with stationarity violations?

To evaluate how the tests cope with violations of the stationarity assumption, we executed two simulation studies in which we concatenate two pairs of time series that have either different relative frequency or different auto-dependency. The properties of X and Y thus change simultaneously. In the first study, we let the relative frequency switch from $p_1^X = p_1^Y = .1$ to $p_1^X = p_1^Y = .3$, while the auto-dependency $p_{1|1}^X = p_{1|1}^Y$ takes the values .4, .5, .6, .7 or .8 across the whole time series. We also varied the length of the time series that were concatenated (i.e., the number of time points before and after the change), considering three scenarios: [500,500], [100,900], and [900,100]. In the second study, the auto-dependency $p_{1|1}^X = p_{1|1}^Y$ changes from .4 to .8, while the relative frequency $p_1^X = p_1^Y$ equals .1, .2 or .3 throughout. Here, we focused on the cases with [500,500] and [900,100] time points.

Figure 9 shows that changing the relative frequencies has a clear impact. Both the Model-AD and the Permut-AD tests have too high Type I error rates, especially in the [500,500] settings and in settings with low auto-dependency. Except for the proportion of agreement, the results of both tests are quite comparable, in that both tests have lower Type I error rates when the higher relative frequency is shown for 900 time points. As expected, the Model-No and Permut-No tests perform worse in all conditions.

Setting aside the proportion of agreement, which again leads to somewhat deviant results, Figure 10 shows that both AD tests cope relatively well with the change in auto-dependency. The Model-No and Permut-No tests perform worse in all conditions.

4. Simulation study 2: Power

In this second simulation study, we investigated and compared the power of the model-based and the permutation-based approaches. Given the results of the first simulation study, we only considered the two tests that take into account auto-dependency (Model-AD and Permut-AD). Furthermore, we excluded the proportion of agreement, given that it showed too low Type I error rates for the Model-AD test. To investigate power, we calculated the Jaccard and kappa values of (X, Y) variable pairs that were generated to have a specific association strength (operationalized as Jaccard score). Power can then be estimated by computing how often the Model-AD and Permut-AD tests yield a significant result.

4.1. Design

We considered 400 design cells, which were obtained by manipulating the following four factors:

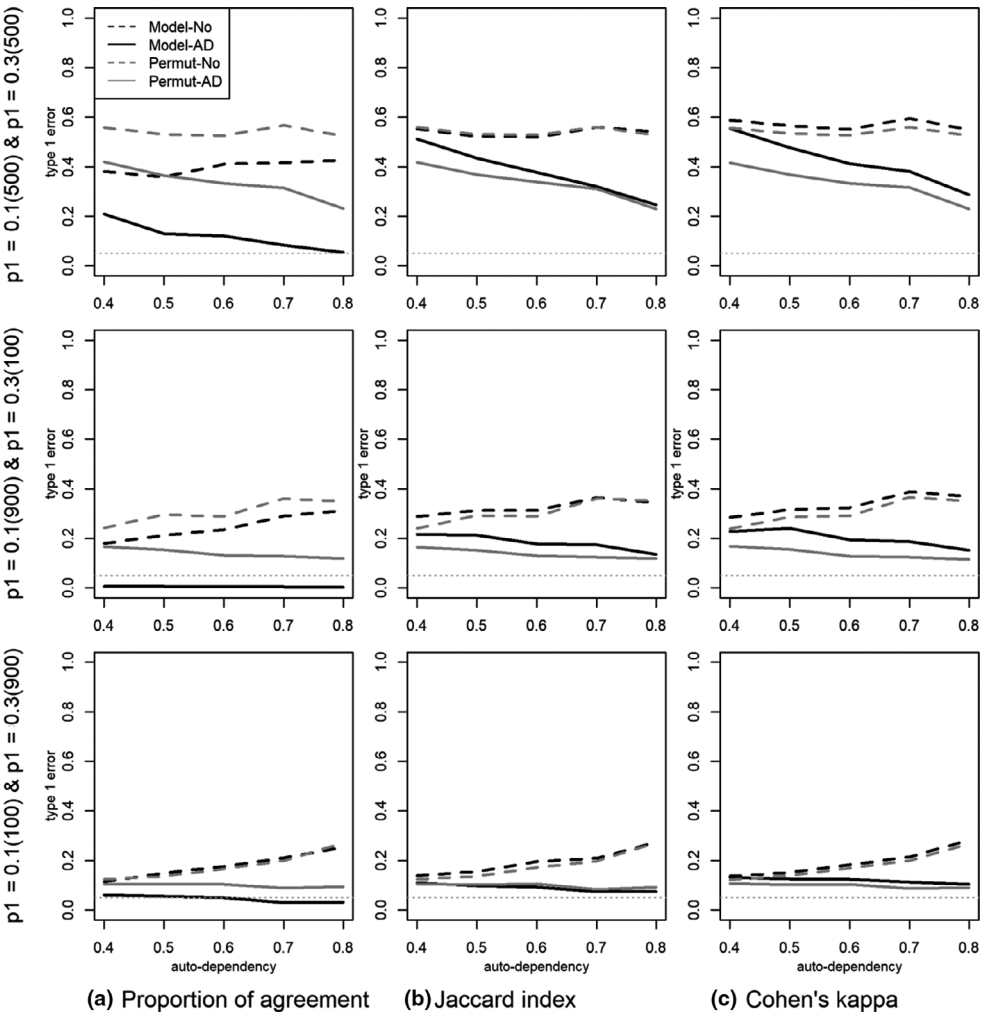


Figure 9. Type I error rates for time series with changing relative frequencies for (a) proportion of agreement, (b) Jaccard index, (c) Cohen's kappa. The change in relative frequency occurs after 500 (first row), 900 (second row) or 1,00 time points (third row). The relative frequencies change from .1 to .3; the auto-dependency is depicted on the *x*-axis. Black lines refer to the model-based tests and grey lines to the permutation-based tests, while solid lines indicate AD tests and dashed lines No tests.

- Relative frequency of X (p_1^X) and Y (p_1^Y), .1, .2 and .3;
- Auto-dependency of X ($p_{1|1}^X$) and Y ($p_{1|1}^Y$), .4, .5, .6, .7, .8 if the relative frequency equals .1 or .2, and .5, .6, .7, .8, .9 if it equals .3;[8]
- Length of time series, 100, 200, 500 or 1,000 time points;

[8] Note that we did not fully cross auto-dependency and relative frequency, because this turned out to be infeasible. Many short time series with an auto-dependency of .9 and a relative frequency of .1 or .2 consisted of 0s only rendering it impossible to compute Jaccard or kappa. Moreover, combining an auto-dependency of .4 with a relative frequency of .3 and a high association strength was also too challenging (see Table S2 in the online supporting materials).

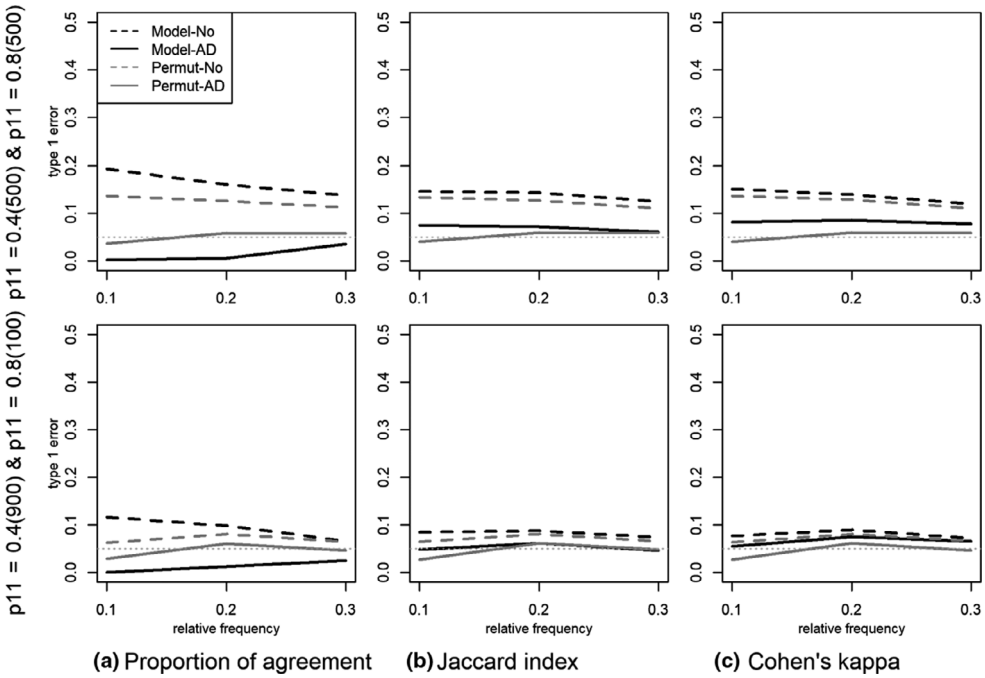


Figure 10. Type I error rates for time series with changing auto-dependencies for (a) proportion of agreement, (b) Jaccard index, (c) Cohen's kappa. The change occurs after 500 (first row) or 900 time points (second row). The auto-dependency changes from .4 to .8; the relative frequency is depicted on the *x*-axis. Black lines refer to the model-based tests and grey lines to the permutation-based tests, while solid lines indicate AD tests and dashed lines No tests.

Table 1. Jaccard values used in the power simulations as a function of the relative frequency of X

Relative frequency p_1^X	Expected Jaccard value	Jaccard values used in the simulation
.1	.0526	.025, .05, .1, .2, .3, .4, .5
.2	.1111	.05, .1, .15, .2, .3, .4, .5
.3	.1764	.1, .15, .2, .3, .4, .5

- Pairwise association strength.

To manipulate the association strength, we relied on the Jaccard similarity index. As the expected Jaccard value (i.e., the Jaccard value for independent X and Y variables) depends on the relative frequencies (see Figure 2), we had to adapt the Jaccard values used based on the relative frequencies under study. Table 1 shows the Jaccard values that we used for each level of relative frequency, as well as the expected Jaccard value.

For each of the 400 design cells, 1,000 observed (X, Y) pairs were generated. To achieve the desired association strength, we introduced cross-conditional probabilities in the data generation scheme. Cross-conditional probabilities express the likelihood that a behaviour is shown at time point t given the states of both variables at the previous time point $t - 1$. For instance, $p^X_{1|1\&1} = P(X(t) = 1|X(t-1) = 1 \wedge Y(t-1) = 1)$ indicates the

probability that behaviour X occurs at time point t, given that both behaviours X and Y occurred at the previous time point. The three other cross-conditional probabilities for variable X, $p^X_{1|1\&0}$, $p^X_{1|0\&1}$ and $p^X_{1|0\&0}$, and the four for variable Y, $p^Y_{1|1\&1}$, $p^Y_{1|1\&0}$, $p^Y_{1|0\&1}$ and $p^Y_{1|0\&0}$, are defined correspondingly. To find the cross-conditional probabilities that match the targeted relative frequencies, auto-dependencies and Jaccard value, we searched all possible values using the differential evolution algorithm DEoptim from the R package *DEoptim* (Ardia, Boudt, Carl, Mullen, & Peterson, 2011; Price, Storn, & Lampinen, 2005). The cross-conditional probabilities obtained were then used to generate the scores on X and Y. First, $X(1)$ and $Y(1)$ were sampled from a Bernoulli distribution with parameter p^X_1. Next, all other $X(t)$ and $Y(t)$ scores were generated time point by time point and drawn from a Bernoulli distribution with parameter the cross-conditional probability that corresponds to the previous states of X and Y (e.g., if $X(1) = 1$ and $Y(1) = 0$, $p^X_{1|1\&0}$ is used for drawing $X(2)$ and $p^Y_{1|0\&1}$ for sampling $Y(2)$).

4.2. Results

Figure 11 shows how power changes as a function of the length of the time series and of association strength, operationalized in terms of the Jaccard index. Since the interpretation of the Jaccard values considered depends on the relative frequencies selected (i.e., higher relative frequencies lead to higher expected Jaccard values; see Table 1) the results are split out across these different relative frequencies. Therefore, it makes sense that, in

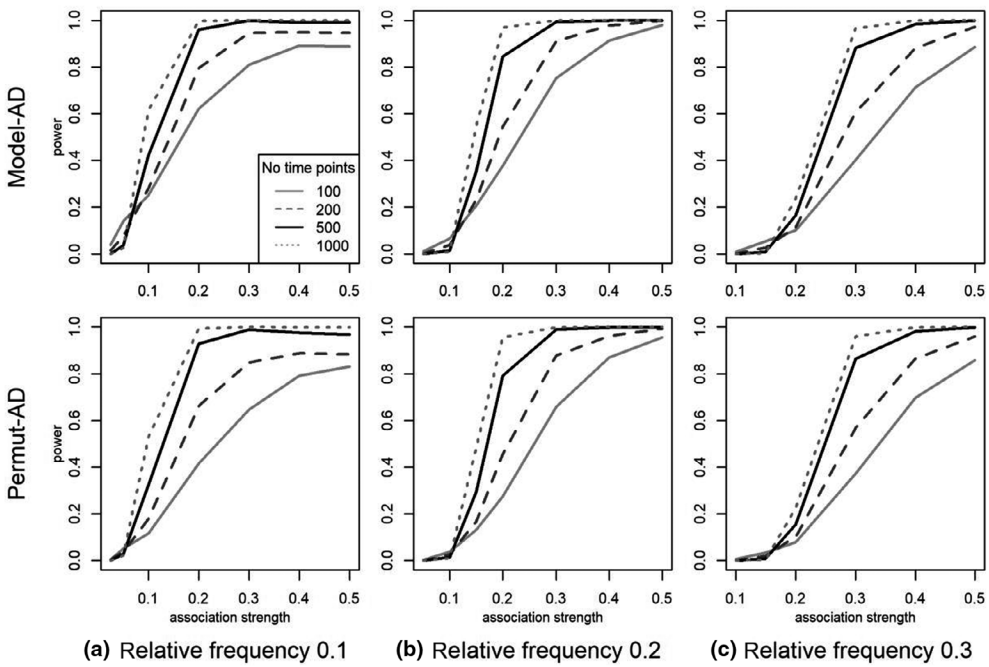


Figure 11. Power for different length of time series for relative frequency (a) .1, (b) .2, (c) .3 in the Model-AD test (first row) and the Permut-AD test (second row). The time series are 100 (solid grey line), 200 (dashed black line), 500 (solid black line) and 1000(dotted gray line) time points long. The results are averaged over auto-dependency and association measure.

each panel, the power curve starts to rise when the manipulated Jaccard value exceeds the expected Jaccard value, which amounts to .0526 for the first panel, .1111 for the second and .1764 for the third (see Table 1). We can conclude that the two tests yield rather comparable results for both association measures, with higher association strength and longer time series leading to higher power.[9]

5. Discussion

Research on the micro-processes that shape interpersonal relations and interactions is gaining ground, often focusing on the behaviours that are displayed in dyadic interactions. To investigate the co-occurrence and temporal sequencing of behaviours, researchers rely on association measures, such as the proportion of agreement, the Jaccard index, and Cohen's kappa. We have demonstrated that the values of these association measures are often ambiguous, however, as they depend on the marginal frequency and auto-dependency of the behaviours under study. To enable researchers to assess whether a certain association value is extreme or not, given the relative frequency and auto-dependency, we investigated the performance of two significance testing paradigms that can take both features into account: a model-based and a permutation-based approach. Note that a beta version of an R package that implements both paradigms can be obtained by emailing the first author. We conducted two simulation studies to investigate the Type I error rate and power of both paradigms, respectively.

5.1. Comparison of the model- and permutation-based approaches

The results of these simulation studies confirm our hypothesis that not accounting for auto-dependency in the significance test inflates the Type I error rate, with more auto-dependency leading to higher Type I error rates. The AD versions of both testing approaches are equally efficient in controlling the Type I error rate and yield comparable power. As could be expected, longer time series lead to steeper power curves.

Researchers should be aware that both testing frameworks assume that the relative frequencies and auto-dependencies are stable across the whole time series. It this assumption is violated, performance of both testing approaches is impacted. Specifically, changes in relative frequency lead to too high Type I error rates, whereas changes in auto-dependency are less detrimental. To deal with this, one can check for the presence of relative frequency change points beforehand, by applying a change point detection framework such as the one proposed by Cabrieto, Tuerlinckx, Kuppens, Hunyadi, and Ceulemans (2018) and Cabrieto, Adolf, Tuerlinckx, Kuppens, and Ceulemans (2019), and split the time series according to the change points obtained into smaller time series that are analysed separately. An alternative is to develop robust tests (e.g., Boker, Rotondo, Xu, & King, 2002; Ombao, Raz, von Sachs, & Malow, 2001; Sanderson, Fryzlewicz, & Jones, 2010).

So far, we have emphasized that both testing approaches yield very similar results. One may wonder whether this will hold in all settings. The main difference between the two approaches pertains to how they generate the new data. Whereas the Model-AD test simulates data from an underlying model, the Permut-AD test (Moulder et al., 2018) cuts

[9] Investigating the power of time series with different relative frequencies and auto-dependencies revealed comparable results. Due to the bad results of the stationarity violations in our first study we did not further investigate this topic as whatever the results we obtained, they would be meaningless given the high degree of Type I error inflation.

the original time series into segments which are subsequently shuffled. Hence, the relative frequency obtained will fluctuate across the simulated data in the Model-AD approach, and the amount of variation will depend on time series length. In contrast, the relative frequency is kept constant in the Permut-AD approach. If relative frequency is very low, the latter feature limits the number of possible co-occurrences of behaviours in the permutation-based test, which may lead to very discrete distributions, as we observed for the example dyad (Figure 5). In the model-based test, very low or high frequencies may also lead to non-symmetric sampling distributions for the proportion of agreement, which in turn result in unusual distributions for Cohen's kappa. Additionally, generating short time series with low relative frequencies and high auto-dependency sometimes resulted in time series where one or both time series only consist only of 0s. In the first case, Cohen's kappa and Jaccard will be zero for many pairs (see Figure 5); in the second case, the two agreement measures cannot be calculated. We handled the second case by not taking these simulated data into account when building the sampling distribution for the model-based test. Yet, this means that the distribution depends on fewer and maybe also less representative cases, implying that the result of the test might be less accurate. Hence, extreme frequency values are challenging for both tests but for different reasons.

In general, both tests are very flexible; considering the use of different association measures and using different association measures with different tests leads to very comparable results. The proportion of agreement used in combination with the model-based approach nevertheless deserves some specific attention. *The fact that proportion of agreement values have a lower limit, which strongly depends on the relative frequency in combination with the variability of this relative frequency in the model-based approach, leads to a deflation of Type I error*,[10] which is absent with the permutation-based tests.

An important distinctive feature of the Permut-AD test, which might make it more robust than the Model-AD test, however, is that permuting segments of the original data also automatically keeps some other characteristics of the time series (cycles, second-order auto-dependencies within the time series, etc.) fairly constant. Such characteristics will largely be removed by the model-based approach, unless they are included in the generation model. Adding more model parameters to the data generation step might be challenging, however, as we already observed when generating our simulation data.

5.2. Limitations of the simulation study

In the simulations we mainly (i.e., except when studying violations of the assumptions) generated data that were in line with the assumptions of our model-based test. Investigating the tests' performance on other simulation data could be informative (e.g., recurring behavioural cycles, influences within a time series that last longer than one time point). Moreover, when investigating power, finding cross-conditional probabilities that adequately fitted specific combinations of relative frequencies, auto-conditional probabilities and association strength with the DEoptim search algorithm was difficult in some cases.

[10] Statistically almost half of the relative frequencies will be smaller than the one observed, leading in many cases to a higher proportion of agreement values. As a consequence, it is almost impossible for the real value to be bigger than 95% of the sampled values.

5.3. Future research directions

We see many directions for future research. First, our test was designed to investigate a one-sided hypothesis by testing whether the observed association value is significantly higher than expected by chance, yielding a one-tailed *p*-value. Building a two-tailed version of the test, though potentially useful to inspect whether observed associations are significantly weaker or stronger than expected by chance, is not straightforward. While checking whether an observed association value is smaller or larger than percentiles of the simulation sampling distribution is easy, obtaining a two-tailed *p*-value is complicated as these simulated sampling distributions can be rather skewed (see Figure 5). To handle this, the work being done on estimating the highest density region of arbitrary distributions (e.g., Hyndman, 1996) may be useful.

Second, both proposed frameworks allow researchers to perform power analysis to optimize study design. One interesting question when designing a study is how long an observed interaction needs to be (i.e., the number of time points) to have sufficient power to test the hypothesis. This question is vital, as micro-coding is a time-consuming and resource-exhausting activity. Based on expected relative frequencies, auto-conditional probabilities and association strength, we can study power as a function of the length of the time series (comparable to what we did in study 2) and thus give useful information regarding this question.

Third, in this paper, we limited ourselves to three association measures, to illustrate the flexibility of the two compared frameworks for significance testing. The test can easily be extended to other association measures, such as the odds ratio. However, as we demonstrated above, each measure comes with its own intricacies (e.g., a limited range of values that strongly depends on relative frequency for the proportion of agreement). Therefore, an interesting question concerns the relative performance of all these measures. It would be nice to gain more insight into which measure works well under which circumstances (Doove, Wilderjans, Calcagnì, & Van Mechelen, 2017).

Fourth, we only considered time-based coding, where the presence or absence of behaviours is rated across time intervals of a fixed length. The length of the intervals automatically impacts auto-dependency, where shorter intervals augment auto-dependency and thus complicate significance testing. Longer intervals, by contrast, might miss out on interesting behavioural patterns. Next to time-based coding approaches, event-based approaches have been proposed (Bakeman & Gottman, 1997; Hollenstein, 2007). The crucial difference is that event-based approaches only take into account the occurrence of a certain behaviour, but ignore its duration. Investigating the advantages and disadvantages of different ranges of intervals, as well as those of event-based data collection for the performance of the significance tests seems to be an important next step.

5.4. Conclusion

To conclude, we have showcased the importance of accounting for relative frequency as well as auto-dependency when testing the association strength of binary dyadic time series, in that ignoring these features can lead to inflated Type I probabilities. Moreover, we compared two test frameworks that allow researchers to take relative frequencies and auto-dependencies into account, demonstrating that both frameworks have an appropriate Type I error rate, while power is reasonable. We hope that our results will ultimately lead to deeper understanding of interpersonal processes, and will further increase awareness that time itself and related effects should never be taken too lightly when working with intensive longitudinal data.

Acknowledgements

The research leading to the results reported in this paper was sponsored in part by two research grants from the Fund for Scientific Research – Flanders (FWO, Project No. G066316N awarded to Eva Ceulemans and Francis Tuerlinckx) and by the Research Council of KU Leuven (C14/19/054). The computational resources and services used in this work were provided by the Flemish Supercomputer Center (VSC), which is funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

Conflict of interest

All authors declare no conflict of interest.

Author contribution

Nadja Bodner (Conceptualization; Formal analysis; Methodology; Software; Visualization; Writing – original draft) Francis Tuerlinckx (Methodology; Supervision; Writing – review & editing) Guy Bosmans (Supervision; Writing – review & editing) Eva Ceulemans (Conceptualization; Methodology; Supervision; Writing – review & editing).

Data availability statement

The data that support the findings of this study are openly available on OSF at https://osf.io/4qbvm.

References

Allen, N.B., Kuppens, P., & Sheeber, L.B. (2012). Heart rate responses to parental behavior in depressed adolescents. *Biological Psychology*, *90*(1), 80–87. https://doi.org/10.1016/j.biopsycho.2012.02.013

Allen, N.B., & L.B. Sheeber (Eds.) (2008). *Adolescent emotional development and the emergence of depressive disorders*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511551963

Ardia, D., Boudt, K., Carl, P., Mullen, K.M., & Peterson, B.G. (2011). Differential evolution with DEoptim. *R Journal*, *3*(1), 27. https://papers.ssrn.com/abstract=1584905

Bakeman, R., & Gottman, J.M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Bakeman, R., McArthur, D., & Quera, V. (1996). Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared. *Behavior Research Methods, Instruments, & Computers*, *28*, 446–457. https://doi.org/10.3758/BF03200524

Bakeman, R., McArthur, D., Quera, V., & Robinson, B.F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*, 357–370. https://doi.org/10.1037/1082-989X.2.4.357

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139017343

Bodner, N., Kuppens, P., Allen, N.B., Sheeber, L.B., & Ceulemans, E. (2018). Affective family interactions and their associations with adolescent depression: A dynamic network approach.

Development and Psychopathology, *30*, 1459–1473. https://doi.org/10.1017/S0954579417001699

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

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR (1) based models do not always outpredict AR (1) models in current psychological applications. *Psychological Methods*, *23*, 740–756. https://doi.org/10.1037/met0000178

Butler, E.A. (2011). Temporal interpersonal emotion systems: The 'TIES' that form relationships. *Personality and Social Psychology Review*, *15*, 367–393. https://doi.org/10.1177/1088868311411164

Butler, E.A., & Randall, A.K. (2013). Emotional coregulation in close relationships. *Emotion Review*, *5*, 202–210. https://doi.org/10.1177/1754073912451630

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2019). An objective, comprehensive and flexible statistical framework for detecting early warning signs of mental health problems. *Psychotherapy and Psychosomatics*, *88*, 184–186. https://doi.org/10.1159/000494356

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., & Ceulemans, E. (2018). Testing for the presence of correlation changes in a multivariate time series: A permutation based approach. *Scientific Reports*, *8*(1), 769. https://doi.org/10.1038/s41598-017-19067-2

Cassidy, J. (1994). Emotion regulation: Influences of attachment relationships. *Monographs of the Society for Research in Child Development*, *59*, 228–249. https://doi.org/10.2307/1166148

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cummings, E.M., Davies, P., & Campbell, S.B. (2000). *Developmental psychopathology and family process: Theory, research, and clinical implications*. New York, NY: Guilford Press.

Doove, L.L., Wilderjans, T.F., Calcagnì, A., & Van Mechelen, I. (2017). Deriving optimal data-analytic regimes from benchmarking studies. *Computational Statistics & Data Analysis*, *107*, 81–91. https://doi.org/10.1016/j.csda.2016.10.016

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Hollenstein, T. (2007). State space grids: Analyzing dynamics across development. *International Journal of Behavioral Development*, *31*, 384–396. https://doi.org/10.1177/0165025407077765

Hollenstein, T., Allen, N., & Sheeber, L. (2016). Affective patterns in triadic family interactions: Associations with adolescent depression. *Development and Psychopathology*, *28*(1), 85–96. https://doi.org/10.1017/S0954579415000309

Holloway, E.L., Wampold, B.E., & Nelson, M.L. (1990). Use of a paradoxical intervention with a couple: An interactional analysis. *Journal of Family Psychology*, *3*, 385–402. https://doi.org/10.1037/h0080552

Hunter, E.C., Hessler, D.M., & Fainsilber Katz, L. (2008). Familial processes related to affective development. In N.B. Allen & L. Sheeber (Eds.), *Adolescent emotional development and the emergence of depressive disorders* (pp. 262–277). Cambridge, UK : Cambridge University Press.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *American Statistician*, *50*, 120–126. http://www.jstor.org/stable/2684423

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin De La Société Vaudoise Des Sciences Naturelles*, *37*, 547–579.

Jaccard, P. (1912). The distribution of the flora in the Alpine Zone. *New Phytologist*, *11*, 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Kobak, R., & Bosmans, G. (2019). Attachment and psychopathology: A dynamic model of the insecure cycle. *Current Opinion in Psychology*, *25*, 76–80. https://doi.org/10.1016/j.copsyc.2018.02.018

Lichtwarck-Aschoff, A., Kunnen, S.E., & van Geert, P.L.C. (2009). Here we go again: A dynamic systems perspective on emotional rigidity across parent–adolescent conflicts. *Developmental Psychology*, *45*, 1364–1375. https://doi.org/10.1037/a0016713

Lloyd, B.P., Yoder, P.J., Tapp, J., & Staubitz, J.L. (2016). The relative accuracy and interpretability of five sequential analysis methods: A simulation study. *Behavior Research Methods*, *48*, 1482–1491. https://doi.org/10.3758/s13428-015-0661-5

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-540-27752-1

Main, A., Paxton, A., & Dale, R. (2016). An exploratory analysis of emotion dynamics between mothers and adolescents during conflict discussions. *Emotion*, *16*, 913–928. https://doi.org/10.1037/emo0000180

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*, *23*, 757–773. https://doi.org/10.1037/met0000172

Ombao, H.C., Raz, J.A., von Sachs, R., & Malow, B.A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, *96*, 543–560. https://doi.org/10.1198/016214501753168244

Price, K.V., Storn, R.M., & Lampinen, J.A. (2005). *Differential evolution: A practical approach to global optimization*. Berlin, Germany: Springer. https://doi.org/10.1007/3-540-31306-0

Randall, A.K., & Schoebi, D. (2015). Lean on me: Susceptibility to partner affect attenuates psychological distress over a 12-month period. *Emotion*, *15*, 201–210. https://doi.org/10.1037/emo0000043

Sanderson, J., Fryzlewicz, P., & Jones, M.W. (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, *97*, 435–446. https://doi.org/10.1093/biomet/asq007

Schwartz, O.S., Sheeber, L.B., Dudgeon, P., & Allen, N.B. (2012). Emotion socialization within the family environment and adolescent depression. *Clinical Psychology Review*, *32*, 447–453. https://doi.org/10.1016/j.cpr.2012.05.002

Sels, L., Ceulemans, E., Bulteel, K., & Kuppens, P. (2016). Emotional interdependence and well-being in close relationships. *Frontiers in Psychology*, 7(283), 1–13. https://doi.org/10.3389/fpsyg.2016.00283

Sels, L., Ceulemans, E., & Kuppens, P. (2018a). A general framework for capturing interpersonal emotion dynamics: Associations with psychological and relational adjustment. In A.K. Randall & D. Schoebi (Eds.), *Interpersonal emotion dynamics in close relationships* (pp. 27–46). Cambridge, UK: Cambridge University Press.

Sels, L., Ceulemans, E., & Kuppens, P. (2018b). All's well that ends well? A test of the peak-end rule in couples' conflict discussions. *European Journal of Social Psychology*, *49*, 794–806. https://doi.org/10.1002/ejsp.2547

Sheeber, L.B., Kuppens, P., Shortt, J.W., Katz, L.F., Davis, B., & Allen, N.B. (2012). Depression is associated with the escalation of adolescents' dysphoric behavior during interactions with parents. *Emotion*, *12*, 913–918. https://doi.org/10.1037/a0025784

Tiedens, L.Z., & C.W. Leach (Eds.) (2004). *The social life of emotions*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511819568

Van keer, I., Ceulemans, E., Bodner, N., Vandesande, S., Van Leeuwen, K., & Maes, B. (2019). Parent-child interaction: A micro-level sequential approach in children with a significant cognitive and motor developmental delay. *Research in Developmental Disabilities*, *85*, 172–186. https://doi.org/10.1016/j.ridd.2018.11.008

Van keer, I., Colla, S., Van Leeuwen, K., Vlaskamp, C., Ceulemans, E., Hoppenbrouwers, K., Desoete, A., & Maes, B. (2017). Exploring parental behavior and child interactive engagement: A study on children with a significant cognitive and motor developmental delay. *Research in Developmental Disabilities*, *64*, 131–142. https://doi.org/10.1016/j.ridd.2017.04.002

Supporting Information

The following supporting information may be found in the online edition of the article:

Table S1 Number of cases in which both simulated time series had a relative frequency of zero.
Table S2 Number of cases in which both time series had a relative frequency of zero, when running power simulations.
Figure S1 Type 1 error rate as a function of the relative frequency and the auto-dependency
Figure S2 Type 1 error rate as a function of the length of the time series (i.e., number of time points) and auto-dependency.

Appendix :

In this appendix we derive the equation

$$p^X_{1|0} \approx \frac{p^X_1 - p^X_1 * p^X_{1|1}}{1 - p^X_1},$$

where $p^X_1 = p(X_t = 1)$.
We have

$$p(\mathrm{X}_t = 1) = p(X_t = 1 \cap X_{t-1} = 1) + p(X_t = 1 \cap X_{t-1} = 0)$$

because $p(\mathrm{A}) = p(A \cap B) + p(A \cap \overline{B})$, with $\overline{B}$ denoting the compliment of B or not B. Then

$$p(X_t = 1) = p(X_t = 1 | X_{t-1} = 1) * p(X_{t-1} = 1) + p(X_t = 1 | X_{t-1} = 0) * p(X_{t-1} = 0)$$

because $p(A|B) = \frac{p(A \cap B)}{p(B)}$ Next

$$p(X_t = 1) \approx p(X_t = 1 | X_{t-1} = 1) * p(X_t = 1) + p(X_t = 1 | X_{t-1} = 0) * p(X_t = 0)$$

because $p(X_{t-1} = 0) \approx p(X_t = 0)$, and

$$p(X_t = 1) \approx p(X_t = 1 | X_{t-1} = 1) * p(X_t = 1) + p(X_t = 1 | X_{t-1} = 0) * [1 - p(X_t = 1)]$$

because $p(X_t = 0) = 1 - p(X_t = 1)$. Rearranging,

$$p(X_t = 1 | X_{t-1} = 0) * [1 - p(X_t = 1)] \approx p(X_t = 1) - p(X_t = 1 | X_{t-1} = 1) * p(X_t = 1)$$

$$p(X_t = 1 | X_{t-1} = 0) \approx \frac{p(X_t = 1) - p(X_t = 1 | X_{t-1} = 1) * p(X_t = 1)}{1 - p(X_t = 1)}$$

if $1 - p(X_t = 1) \neq 0$. Thus $p^X_{1|0} \approx \frac{(p^X_1 - p^X_1 p^X_{1|1})}{(1 - p^X_1)}$, and we are done.