Test–retest reliability measures for curve data: an overview with recommendations and supplementary code

Alessia Pini\(^\text{a,b}\), Jonas L Markström\(^\text{c}\) and Lina Schelin\(^\text{a}\)

\(^\text{a}\)Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden; \(^\text{b}\)Department of Statistical Sciences, Catholic University of the Sacred Heart, Milan, Italy; \(^\text{c}\)Department of Community Medicine and Rehabilitation, Physiotherapy, Umeå University, Umeå, Sweden

ABSTRACT
The purpose of this paper is to provide an overview of available methods for reliability investigations when the outcome of interest is a curve. Curve data, or functional data, is commonly collected in biomechanical research in order to better understand different aspects of human movement. Using recent statistical developments, curve data can be analysed in its most detailed form, as functions. However, an overview of appropriate statistical methods for assessing reliability of curve data is lacking. A review of contemporary literature of reliability measures for curve data within the fields of biomechanics and statistics identified the following methods: coefficient of multiple correlation, functional limits of agreement, measures of distance and similarity, and integrated pointwise indices (an extension of univariate reliability measures to curve data, inclusive of Pearson correlation, intraclass correlation, and standard error of measurement). These methods are briefly presented, implemented (R-code available as supplementary material) and evaluated on simulated data to highlight advantages and disadvantages of the methods. Among the identified methods, the integrated intraclass correlation and standard error of measurement are recommended. These methods are straightforward to implement, enable results over the domain, and consider variation between individuals, which the other methods partly neglect.

ARTICLE HISTORY
Received 18 February 2019
Accepted 5 August 2019

KEYWORDS
Agreement; functional data; kinematics; similarity

Introduction
Biomechanical research often involves the collection of large data sets based on integrated three-dimensional kinematics (e.g., motion capture), kinetics (e.g., force plate data), and muscle activation patterns (e.g., electromyography) during variable functional tasks. Such biomechanical data are high-dimensional and complex (Colyer, Evans, Cosker, & Salo, 2018) but are usually reduced to simple discrete outcome measures. Common examples found in the literature include joint range of motion (Sinsurin, Vachalathiti, Srisangboriboon, & Richards, 2018), maximal angles and moments (Markström, Schelin, & Häger, 2018), ground reaction forces (Elias, 2019).
Hammill, & Mizner, 2015; Sinsurin et al., 2018), values of co-contraction (Elias et al., 2015) and others. These, or similar, outcomes are often collected within specific time frames of interest to enable the use of basic and well-documented statistical methods. However, it is logical to believe, and has been shown, that data showing the entire movement during, e.g., a jump landing, will reveal information that is not evident from pre-selected data points (Donoghue, Harrison, Coffey, & Hayes, 2008; Pataky, Robinson, & Vanrenterghem, 2013; Richter, O’Connor, Marshall, & Moran, 2014). As such, there has been an increased interest and application of methods for analysis of curve data within the field of biomechanics, e.g., using various techniques from functional data analysis (FDA) (Ramsay & Silverman, 2005) or statistical parametric mapping (Friston et al., 1994). Selected research includes topics of sports performance, functional developmental stages for children, and consequences of injury to the anterior cruciate ligament (see, e.g., Harrison, Ryan, & Hayes, 2007; Hébert-Losier et al., 2015; Hébert-Losier, Schelin, Tengman, Strong, & Häger, 2018; Ryan, Harrison, & Hayes, 2006; Warmenhoven et al., 2019a, 2019b). Thus, the variables of interest are no longer restricted to single values but also include functions over a domain. From a theoretical point, such functions are assumed to have an infinite dimension. However, even when analysing it as functions, the data over the domain are typically discretised, either by a basis representation or, as here, by pointwise evaluations of the functions.

The reliability and validity of the data acquired from curve analyses remains an important consideration, particularly given the increasing interest in the analysis of such data. Repetition of trials and/or observations is common practice in scientific literature when measuring task performance over time as well as the performance of the assessors. If such measures are not reliable it is impossible to establish whether an observed effect of an intervention is due to the intervention itself or if it is due instead to the error in the measurement tools or variability in motor performance. Both absolute and relative reliability must be established before interpreting and presenting conclusions of the data. Absolute reliability is the degree of variation in measurements for individuals, whereas relative reliability is the degree of variation in position among individuals over repeated measurements (Bruton, Conway, & Holgate, 2000; Weir, 2005). While several articles have reviewed, discussed and presented available methods for reliability studies for discrete data (Atkinson & Nevill, 1998; Bruton et al., 2000; De Vet, Terwee, Knol, & Bouter, 2006; Hopkins, 2000; Rankin & Stokes, 1998; Weir, 2005), the same interest has not been shown to the topic of reliability for curve data. There is no study to date that presents an overview or review of suitable methods that may be used to evaluate reliability when the outcome of interest is an entire curve. The lack of such overview hinders future studies from adequately analysing and presenting the reliability of more complex data sets before drawing (possibly erroneous) conclusions from the data. Although complex data also include multidimensional curves, the focus of this manuscript is on curves in one dimension.

The purpose of this paper is therefore to provide an overview of available methods that may be used to evaluate reliability of curve data in the common case of a test–retest setting. The characteristics of these methods are described in a language adapted for the wider audience within the field of biomechanics while advantages and disadvantages are highlighted by applying them to different sets of simulated data. R-code (R Core Team,
Overview of suitable reliability methods for curve data

A review of present literature for reliability measures in a test–retest setting related to curve analyses was performed using Google Scholar, Web of Science and Medline. This search was directed to literature related to the analysis of biomechanics and to statistics in general. Since an ambiguity in terminology exists in the literature where alternative terms like repeatability, consistency, agreement, concordance, reproducibility, and stability are used in preference to reliability (Atkinson & Nevill, 1998; Berchtold, 2016; De Vet et al., 2006), these terms were also considered in the searches. The reference lists within the relevant articles were also reviewed.

Within the biomechanics-related literature the most frequently occurring measures used to evaluate relative reliability (for discrete values) were found to be correlation coefficients, including Pearson correlation and different versions of the intraclass correlation (ICC). There are multiple guidelines available that are dedicated to analyses of such data using ICCs (Koo & Li, 2016; McGraw & Wong, 1996; Trevethan, 2017). The most common measures to evaluate absolute reliability includes the standard error of measurement (SEM), the coefficient of variation and limits of agreement (LoA) (Bland & Altman, 1986). Relative and absolute measures for pre-selected data points can be applied to curve data by integrating any point-wise index (see, e.g., Kainz et al., 2017; Sankey et al., 2015; Schwartz, Trost, & Wervey, 2004). This is however not widely adopted in the literature. Bland Altman plots have also been used for gait curves as a visual tool in addition to a univariate reliability study (Meldrum, Shouldice, Conroy, Jones, & Forward, 2014). A formal extension of LoA to curve data called functional LoA (FLoA) has been developed by Roislien, Rennie, and Skaaret (2012). Other identified methods and measures include the coefficient of multiple correlation (CMC), which is based on data from the entire gait cycle, and the coefficient of multiple determination (the square of the CMC) (Kadaba et al., 1989). A systematic review by McGinley, Baker, Wolfe, and Morris (2009) identified and evaluated the evidence for reliability of three-dimensional kinematic gait outcomes, but presented no additional methods for curve data.

Within the statistical literature of curve data analyses, measures of similarity or dissimilarity between curves are common (for example, Dubin & Müller, 2005; Jacques & Preda, 2014; Sangalli, Secchi, Vantini, & Vitelli, 2010), as are measures of distances (Jacques & Preda, 2014). Such measures have not yet been used for the purpose of reliability, but potentially provide new possibilities. Curve similarity has also been assessed using linear regression techniques by Iosa et al. (2014), but the main focus of the paper was on reducing the dimension of the data. Using the technique suggested by Iosa et al. (2014), the parameters describing the similarity can be used to investigate test–retest reliability.

As such, this review presents and discuss the following methods: integrated pointwise index, FLoA, CMC, as well as measures of distance and similarity. To begin with, a few key concepts are defined. These concepts are curve, domain and domain alignment,
presented here in Table 1. These terms will be used throughout the paper and should be addressed whenever working with curve data.

For the presentation of the methods that follow, some common notations are introduced. Let $G$ denote the number of test sessions ($G = 2$ in the test–retest setting). Let $i = 1, \ldots, N$, denote the index of the $N$ individuals. Assume that all individuals participate in all test sessions. Let $Y$ denote the variable of interest, e.g., degree of knee flexion. The notation $Y_{gi}$ refers to the curve corresponding to individual $i$ at test occasion $g$. A curve can mathematically be described with a function over the domain $[a, b]$ (e.g., during a landing phase where $a = $ impact and $b = $ peak knee flexion angle). Thus, the variable of interest will be expressed as a function of the point $t$ of the domain, i.e. $Y_{gi} = Y_{gi}(t), t \in [a, b]$. It is common to discretise the domain using a uniform grid for computational reasons. In such cases, denote with $t_1, t_2, \ldots, t_T$ the discrete points of the domain, where $T$ refers to the total number of discretisation points. In this presentation, the data are assumed to be complete. In practice, there may be individuals with missing data or curves with missing fragments, but such a discussion is outside the scope of this paper.

### Integrated pointwise indices

Reliability measures for univariate data have been proposed and widely studied (Bruton et al., 2000; De Vet et al., 2006; Rankin & Stokes, 1998). The extension of these measures to curve data follows by calculating so called integrated pointwise indices. Even though such an approach is intuitive and has been applied in some works (see, e.g., Kainz et al., 2017; Sankey et al., 2015; Schwartz et al., 2004) it has not been generally described. This may be one reason as to why it has not been widely adopted in papers that evaluate reliability of data originating from curves. Another reason may be that it is non-trivial to perform inference for such measures, since confidence intervals or p-values that are calculated pointwise should be presented with caution and should be adjusted to take multiplicity into account (see Discussion and Implications for more details).

---

**Table 1.** Key concepts of curve data.

| CURVE | Whenever the object of the statistical analysis is a quantity that is a function of a continuous parameter (e.g. time), it can be considered curve data. Curve data are often represented as a collection of values at a finite number of points of the (time) domain, with values at neighbouring points being similar. For a general introduction to functional data analysis the interested reader is referred to Ramsay and Silverman (2005) and Sørensen, Goldsmith, and Sangalli (2013). |
| --- | --- |
| DOMAIN | Each curve is observed on a domain that often is, but not restricted to, time or percentage. For gait data, the domain is often defined over a gait cycle as $[0, 100]$. For other types of curve data such as knee joint angles during a one-leg hop for distance landing phase, the choice of the domain $[a, b]$ is less trivial. In general, the start and end of the domain need to be defined and common for all individuals. |
| ALIGNMENT | Gait data has a natural domain alignment (also referred to as registration or warping) where the start and end of the gait cycle often occurs at approximately the same relative (time) point for most individuals. For other applications, domain alignment should be addressed prior to any reliability analysis. Indeed, it can be used to remove differences in phase shift variation or amplitude variation that are not of interest. The simplest form is referred to as landmark registration and implies a simple transformation of the curves. More advanced methods for alignment also exist, see e.g. Marron et al. (2014) and Ramsay and Silverman (2005), as well as the references therein. |
When discretising the domain in a set of points, a functional data set turns out to be a multivariate data set composed of the values of the functional data at the points. At every fixed point \( t \) of the domain, a univariate data set composed of the values at point \( t \) are obtained. Hence, it is always possible to compute all classical reliability measures for every point of the domain. Such indices will vary over the domain; for every different point a different value of reliability, here denoted as \( R(t) \), are obtained. The functional version of a classical reliability index \( R \), i.e., the integrated pointwise index, is obtained by computing the integral of \( R(t) \) over the domain:

\[
R = \frac{1}{b-a} \int_{a}^{b} R(t) \, dt.
\]

If the discretisation step is constant, such a quantity can be numerically estimated with

\[
R = \frac{\sum_{t=t_1}^{t_T} R(t)}{b-a}.
\]

Every reliability index that is defined for univariate data can be extended to functional data by considering its integrated version. For these integrated pointwise indices, three different measures are considered; the Pearson correlation coefficient, ICC (type: absolute agreement) and SEM. These are not the only possible measures that could be used, but they are by far the most commonly used in biomechanical research. The ICC and SEM are calculated using variance components of an ANOVA. In detail, they are based on a two-way mixed effect model without interactions, following the recommendations in Trevethan (2017) for this test–retest situation. The reader is referred to the supplementary material for details of formulas.

**Functional limits of agreement, FLoA**

Functional limits of agreement (FLoA) (Roislien et al., 2012) is an extension of LoA developed by Bland and Altman (1986). It is based on the assumption that the difference, \( \delta_i(t) = Y_{1i}(t) - Y_{2i}(t), i = 1, \ldots, N \), between pairs of measurements and the corresponding average, \( A_i(t) = (Y_{1i}(t) - Y_{2i}(t))/2 \), are uncorrelated. This assumption can be verified using functional linear regression (Abramowicz et al., 2018; Ramsay & Silverman, 2005, pp 247–258). The FLoA is calculated as \( \overline{\delta(t)} \pm 1.96 \times SD(\delta(t)) \) assuming normality, where \( \overline{\delta(t)} \) is the mean difference across individuals \( i, i = 1, \ldots, N \) and \( SD(\delta(t)) \) the corresponding standard deviation. The FLoA approach acknowledges the nature of the curve data and visualises observed differences.

**Coefficient of multiple correlation, CMC**

The CMC is a frequently used measure of similarity between curves (Ford, Myer, & Hewett, 2007; Kadaba et al., 1989; Milner, Westlake, & Tate, 2011). The CMC value for individual \( i \) is
\[ \text{CMC}(i) = \sqrt{1 - \frac{\sum_{g=1}^{G} \sum_{t=t_i}^{t_f} (Y_{gi}(t) - \bar{Y}_i(t))^2 / T(G-1)}{\sum_{g=1}^{G} \sum_{t=t_i}^{t_f} (Y_{gi}(t) - \bar{Y}_i)^2 / (TG-1)}.} \]

where for every individual, \(i\), \(\bar{Y}_i(t)\) is the mean curve across all test situations and \(\bar{Y}_i\) is the mean value across all test situations and all time points \(t = t_1, t_2, \ldots, t_T\). The overall CMC value is obtained by averaging \(\text{CMC}(i)\) across all individuals \(i = 1, \ldots, N\). An alternative way of calculating CMC is given in Røislien, Skare, Opheim, and Rennie (2012), where the CMC is calculated across all individuals simultaneously. CMC attains values between 0 and 1, with higher values corresponding to a higher degree of similarity between curves.

**Distance measures**

In the FDA literature, the distance between curves is often used to measure how similar they are. The most commonly used measure of distance between curves is the \(L_2\)-distance, the extension of the Euclidean distance to functional data. An overview of other distances can be found in Jacques and Preda (2014) and references therein. In detail, for individual \(i\), the \(L_2\)-distance between test and retest curves \(Y_{1i}(t)\) and \(Y_{2i}(t)\) is

\[ d_{L^2}(i) = \sqrt{\int_a^b (Y_{1i}(t) - Y_{2i}(t))^2 \, dt}. \]

If curve data are discretised on a grid of points on the domain \(t = t_1, t_2, \ldots, t_T\) with a constant discretisation step \(\Delta t = (t_T - t_1)/T\), the distance can numerically be estimated with

\[ d_{L^2}(i) = \sqrt{\Delta t \sum_{t=t_i}^{t_f} (Y_{1i}(t) - Y_{2i}(t))^2}. \]

As for CMC, an overall measure of reliability is obtained by averaging \(d_{L^2}(i)\) across all individuals. If the average distance is large, data are not reliable, and vice-versa.

**Similarity measures**

Similarity measures between curves (from the FDA literature) that do not depend on the unit of measurement and that always attain values in the range \([-1, 1]\) can be found in Dubin and Müller (2005) and Sangalli et al. (2010). The simplest example of a similarity measure between two curves is the functional correlation \(S\). For individual \(i\) the functional correlation between the two curves of test and retest \(Y_{1i}(t)\) and \(Y_{2i}(t)\) is:

\[ S(i) = \frac{\int_a^b Y_{1i}(t) Y_{2i}(t) \, dt}{\sqrt{\int_a^b (Y_{1i}(t))^2 \, dt} \sqrt{\int_a^b (Y_{2i}(t))^2 \, dt}}. \]

Again, if curve data are discretised on a grid of points of the domain with constant discretisation steps, the similarity can be numerically estimated with
Note that in this case, it is not necessary to multiply the similarity measure by the discretisation step $\Delta t$, since it is included in both the numerator and the denominator. For the same reason, $S(i)$ is a dimensionless and standardised quantity. A global measure of reliability is obtained by averaging $S(i)$ over all $i = 1, \ldots, N$. This similarity measure assumes high values (close to 1) in case of very similar curves, and low values (close to −1) in case of very different curves.

Simulation description

Data were generated according to eight different scenarios in the test–retest setting to highlight different properties of the methods under comparison. All scenarios were replicated 1000 times, and one instance (replicate) of simulated data for all scenarios is reported in Figure 1. The same random seed was used for each scenario. For each scenario, test/retest curves are plotted in red/blue, with the same line type corresponding to the same individual. As a complement, Figure 2 displays the difference between test and retest for each individual on the same instance of the data. Simulations were carried out with sample sizes of 10, 20 and 30 individuals. Results for different sample sizes were very similar and as such the sample size of 10 is presented only for the purpose of clearer visualisation. Further, similar results were obtained when selecting a different number of points to discretise the domain; however, too few points are not appropriate, since this might lead to a misrepresentation of the real data. In this paper, results for $T = 100$ are presented, but simulations were carried out for $T = 10, 20, 50, 100, 150, 200$. Starting from $T = 50$ the results are very similar to those presented, with differences within two standard deviations from the average value in all scenarios.

The eight scenarios were divided into two different situations, one where the between-individual variability was higher than the within-individual variability (scenarios A1, B1, C1 and D1) and one where the between–individual variability was lower than the within–individual variability (scenarios A2, B2, C2 and D2). In scenarios A1 and A2, all curve data within each scenario were sampled from the same distribution. To resemble real-life situations, scenarios included cases with simulated patient and control groups, represented in the figures by light and dark colours, respectively. In scenarios B1 and B2, the two groups differed in mean but had the same error distribution. Curves of patients (light) assume higher values with respect to curves of the controls (dark). In scenarios C1 and C2, data of controls were simulated as in scenarios A1 and A2, but data of patients always assumed lower values in the test than in retest. The shape of curve data was consistently similar within and between individuals for these first six scenarios. Finally, scenarios D1 and D2 showed two cases where the shape of the curve data of test and retest was different, with two maxima for the test curves and only one maximum for the retest. There were no simulated group differences in scenario A1, A2, D1 and D2.
Based on the construction of the scenarios, a high reliability is expected in scenarios A1 and B1, characterised by high variability between individuals with respect to the variability within individuals and groups of individuals that follow the same model in test-retest. In scenario B2, two groups of individuals are clearly identified both in the test and retest sessions, where the variability between individuals was low within each group. Hence, a high overall reliability is also expected in this scenario, since group differences are not considered. In all other scenarios, a lower reliability is expected.

**Figure 1.** Simulated data including parameters. The plots show an instance of simulated data for each scenario. The curves of test/retest are plotted in red/blue, and the same line type corresponds to the same individual. Curves of patient/control groups are plotted in light/dark colour, respectively. Note that the two groups differ only in scenarios B and C. Additionally, the parameter setting for each simulated scenario is presented [The reader is referred to the electronic version of this paper for the colour representation.].

| Variability parameters | Mean parameters |
|------------------------|-----------------|
| $m_{\tau} = (0,5,0,-5,0) \forall i$ | $m_{\tau} = (0,5,0,-5,0) \forall i$ |
| $m_{\sigma} = (0,0,0,0) \forall i,g$ | $m_{\sigma} = (0,0,0,0) \forall i,g$ |

| $\sigma_{\tau} = (2.5,5,5,5,2.5)$ | $\sigma_{\tau} = (2.5,5,5,5,2.5)$ |
| $\sigma_{\sigma} = (0.5,1,1,1,0.5)$ | $\sigma_{\sigma} = (0.5,1,1,1,0.5)$ |

| $\sigma_{\tau} = (0.25,0.5,0.5,0.5,0.25)$ | $\sigma_{\tau} = (0.25,0.5,0.5,0.5,0.25)$ |
| $\sigma_{\sigma} = (0.5,1,1,1,0.5)$ | $\sigma_{\sigma} = (0.5,1,1,1,0.5)$ |
Figure 2. Simulated data differences. The difference between test and retest for each individual and each scenario (same instance as in Figure 1). Curves of patient/control groups are plotted in light/dark colour, respectively. Note that the two groups differ only in scenarios B and C.
Technical details of the simulation

Data were simulated according to the model,

\[ Y_{gi}(t) = f_i(t) + \varepsilon_{gi}(t) \]

where \( f_i(t) \) was the true curve of individual \( i \) evaluated at time point \( t \) and \( \varepsilon_{gi}(t) \) was an error function. All functions were obtained by simulating coefficients of a cubic B-spline basis on \([0, 100]\) with five basis functions. The five coefficients of the basis expansion of \( f_i(t) \) were independent and normally distributed with \( \mu_{1i}, \mu_{2i}, \mu_{3i}, \mu_{4i}, \mu_{5i} \) and \( \sigma_B = (\sigma_{1i}, \sigma_{2i}, \sigma_{3i}, \sigma_{4i}, \sigma_{5i}) \). The parameter vector \( \sigma_B \) represent variability between different individuals. Coefficients of the basis expansion of \( \varepsilon_{gi}(t) \) were also assumed to be independent and normally distributed. The corresponding notations for the mean and variances were \( m_{\varepsilon_{gi}} \) and \( \sigma_W \), respectively. The parameter vector \( \sigma_W \) represent variability within each individual.

As mentioned, eight different scenarios were considered in the simulations by varying parameters \( m_{f_i}, m_{\varepsilon_{gi}}, \sigma_B, \) and \( \sigma_W \). Within-individual variability was constant and equal to one in all scenarios. In scenarios A1 and A2 parameters \( m_{f_i} \) and \( m_{\varepsilon_{gi}} \) were the same for each individual and for both test sessions. In scenarios B1 and B2 two groups of individuals were simulated, with different values of \( m_{f_i} \). Difference between groups was equally expressed in the first test session \((g = 1)\) and in the second test session \((g = 2)\). In scenarios C1 and C2 two different groups of individuals react differently in the two test sessions, simulated using different values of both \( m_{f_i} \) and \( m_{\varepsilon_{gi}} \). Finally, in scenarios D1 and D2, the two different groups reacted differently in the two test sessions, but only in a part of the domain, simulated by changing the parameter \( m_{\varepsilon_{gi}} \). All parameter values are presented in Figure 1, in addition to one instance (replicate) of simulated data for all scenarios. The R-code for generating the data is provided as supplementary material.

Advantages and disadvantages of the methods

The reliability outcome results of the simulation are reported in Table 2. In detail, for each considered reliability measure the average index over 1000 simulated scenarios are reported, with its standard deviation in parentheses. In scenarios presenting two distinguished groups of individuals (B1, B2, C1 and C2), both a common measure of reliability (ignoring the groups) and the inter-group reliability were computed. The integrated measures can be visually illustrated pointwise over the domain as illustrated for ICC in Figure 3 for one instance of data. Results for FLoA are presented in Figure 4 since the results are visual and without any numerical index. In the following, each of the presented methods is discussed in relation to the results of the simulation, to highlight the advantages and disadvantages with each method.
### Table 2. Simulation results.

Results of the functional simulation for the eight scenarios, presented as mean (sd) over 1000 replicates. Integrated pointwise indices are computed by considering the whole curves. For scenarios with simulated group differences (scenarios B and C), the results for each group are also presented. The expected reliability result for each scenario is also given, with high and low corresponding to general descriptions of reliability. See Figures 1 and 2 for scenario details.

| Scenario | Expected reliability | Integrated pointwise index |
|----------|----------------------|-----------------------------|
|          | Correlation | ICC    | SEM    | CMC     | Distance | Similarity |
| A1       | High        | 0.96 (0.02) | 0.95 (0.02) | 0.57 (0.09) | 0.95 (0.03) | 7.63 (1.16) | 0.96 (0.03) |
|          | Low         | 0.19 (0.22) | 0.18 (0.21) | 0.57 (0.09) | 0.96 (0.01) | 7.63 (1.16) | 0.94 (0.02) |
| B1       | High        | 0.99 (0.00) | 0.99 (0.00) | 0.57 (0.09) | 0.95 (0.03) | 7.63 (1.16) | 0.98 (0.02) |
|          | Controls    | 0.95 (0.05) | 0.93 (0.05) | 0.56 (0.12) | 0.95 (0.04) | 7.61 (1.55) | 0.96 (0.04) |
|          | Patients    | 0.95 (0.05) | 0.93 (0.05) | 0.56 (0.12) | 0.95 (0.04) | 7.65 (1.60) | 1.00 (0.00) |
| B2       | High        | 0.99 (0.00) | 0.99 (0.00) | 0.57 (0.09) | 0.96 (0.01) | 7.63 (1.16) | 0.97 (0.01) |
|          | Controls    | 0.15 (0.29) | 0.14 (0.28) | 0.56 (0.12) | 0.96 (0.02) | 7.61 (1.55) | 0.94 (0.03) |
|          | Patients    | 0.17 (0.31) | 0.12 (0.30) | 0.56 (0.12) | 0.96 (0.02) | 7.65 (1.60) | 1.00 (0.00) |
| C1       | Low         | -0.56 (0.11) | -0.24 (0.04) | 10.02 (0.14) | 0.95 (0.04) | 104.35 (1.61) | 0.06 (0.03) |
|          | Controls    | 0.95 (0.05) | 0.93 (0.05) | 0.56 (0.12) | 0.95 (0.04) | 7.61 (1.55) | 0.96 (0.04) |
|          | Patients    | 0.95 (0.05) | 0.04 (0.02) | 14.15 (0.20) | NA (NA) | 201.10 (2.81) | -0.84 (0.06) |
| C2       | Low         | -0.98 (0.01) | -0.38 (0.01) | 10.02 (0.14) | 0.96 (0.02) | 104.35 (1.61) | 0.01 (0.02) |
|          | Controls    | 0.19 (0.29) | 0.14 (0.28) | 0.56 (0.12) | 0.96 (0.02) | 7.61 (1.55) | 0.94 (0.03) |
|          | Patients    | 0.17 (0.31) | 0.00 (0.00) | 14.15 (0.20) | NA (NA) | 201.10 (2.81) | -0.92 (0.01) |
| D1       | Low         | 0.96 (0.02) | 0.53 (0.05) | 3.66 (0.12) | 0.57 (0.06) | 66.76 (1.82) | 0.60 (0.13) |
| D2       | Low         | 0.19 (0.21) | 0.06 (0.08) | 3.66 (0.12) | 0.47 (0.03) | 66.76 (1.82) | 0.74 (0.05) |

CMC, coefficient of multiple correlation; ICC, intraclass correlation coefficient; SEM, standard error of measurement.
Figure 3. Pointwise ICC. Pointwise ICC for one instance of simulated data, for all eight scenarios. ICC, intraclass correlation coefficient.
Figure 4. Functional limits of agreement. FLoA for one instance of simulated data in all scenarios. For each scenario, the average of differences $\delta(t)$ and the functional limits of agreement $\delta(t) \pm 1.96 \times SD(\delta(t))$ are displayed. FLoA, functional limits of agreement.
Integrated pointwise indices

When integrating a reliability index over the domain, the range, properties and interpretation of the index remain the same as the original univariate version. This is an advantage since it makes interpretation straightforward. For the same reason, potential disadvantages of the chosen univariate index will transfer to its integrated version.

Integrated pointwise indices perform generally well in the sense that they capture the specific aspects of the different scenarios. Integrated Pearson correlation has the desired features for the first scenarios, but it attains high values (close to 1) in scenario D1, where data are not reliable in the second part of the domain. This is explained by the fact that in this scenario there is a constant mean difference between test and retest that is not captured by Pearson correlation, which only focusses on relative differences between data and the corresponding means. For this reason, the correlation values in scenarios D1 and D2 coincide with values in scenarios A1 and A2, respectively. Hence, Pearson correlation is not recommended for reliability analyses when the actual values of data, the absolute agreement, is of interest.

Integrated ICC displays high values (close to 1) in scenarios A1, B1 and B2 and low values in all other cases (see Table 2). This means that it is sensitive to changes in variability between individuals (scenarios A1 and A2), to changes in at least one group of individuals between test and retest (scenarios C1 and C2) and to changes over parts of the domain (scenarios D1 and D2, see Figure 3).

Integrated SEM behaves differently with respect to ICC, but differences are in line with its definition and interpretation. Indeed, SEM is a measure of variability and is expected to be high and low, respectively, in unreliable and reliable situations. The implication of high and low depends on the application. In addition, while changing the between variability parameter as shown for scenarios A1 compared to A2 and B1 compared to B2, the same SEM values are expected. SEM attains low values in scenarios A1, A2, B1 and B2. The SEM has the same unit of measurement as the data and, as such, it has to be compared with the magnitude of data. This is one reason as to why it is recommended and commonly presented in addition to reliability measures such as the ICC.

It should be noted that the indices ICC and SEM are based on variance components from an ANOVA, with an underlying assumption that the model is suitable. Explicit formulas used in the calculations are provided as supplementary material. If formulas for ICC type consistency were used instead of absolute agreement, a high reliability would be observed for scenarios with a constant shift in mean but the same relative differences between test and retest (such as for Pearson correlation coefficient). However, a high reliability (ICC close to one and a relatively small SEM) is not warranted in those scenarios.

For integrated reliability indices, it can also be useful to visualise the pointwise values of the index (Figure 3). The value of ICC is nearly constant over the whole domain for five of the scenarios. Further, the expected change over the domain is visible in scenario D1, with high values only in the first part of the domain. In such cases, the choice of the
domain of the functional data is important since it has a clear impact on the results and its conclusions. To complement the pointwise index with pointwise confidence bands is straightforward from the univariate case. However, producing functional confidence bands taking into consideration the nature of the data is less straightforward and outside the scope of this paper.

**Functional limits of agreement, FLoA**

The FLoA results reported in Figure 4 show a high reliability in scenarios A1, A2, B1 and B2, as well as in parts of the domain for scenarios D1 and D2. The method is not able to distinguish between scenarios where the between-individual variability is higher than the within-individual variability (e.g., C1 compared to C2). FLoA is visually easy to interpret since it provides pointwise bands for the difference between test and retest data, but no global index is provided. Deciding whether the observed difference is within acceptable limits is therefore up to the interpreter. The method relies on the assumption that $\delta(t)$ and $A(t)$ are uncorrelated, which has to be checked through functional linear regression. This may however not be trivial to check and requires additional computational knowledge.

**Coefficient of multiple correlation, CMC**

The CMC is very similar (and close to its maximum value of 1) for the first six scenarios in Table 2 (A1, A2, B1, B2, C1, C2). Several methodological shortcomings have been reported for the CMC (Chia & Sangeux, 2017; Røislien et al., 2012; Sangeux, Passmore, Graham, & Tirosh, 2016). The method does not work for curves with a small ROM, since in such cases the term under the square root can be negative and CMC is undefined. This is illustrated in scenarios C1 and C2, where the CMC for the second group is undefined. In such cases, the common CMC value is computed as the average among individuals of the first group only. This is why the indices for C1 and C1 group 1 are identical and high. This computational disadvantage may result in the mistake of concluding high reliability where it is not. Further, the statistical properties of CMC are related both to the number of individuals and to the number of test situations (Røislien et al., 2012). It does not take into account the variability between individuals. As such, CMC is not able to distinguish between low and high between-individual variation, as in scenarios A1 and A2. Also, the CMC provides an overall value, thus with no information of the reliability throughout the domain.

**Distance measures**

The distance between two curves is always defined and so, unlike CMC, it is always possible to provide a measure of reliability with this method. The same distance value is observed between the first four scenarios (A1, A2, B1, B2), between scenarios C1 and C2, and between scenarios D1 and D2 (Table 2). This occurs because the distance between test and retest curves only depends on the within-individual variability, i.e., the difference
between test and retest measurements. As constructed, the variability between individuals is ignored, as for the CMC. In the mentioned scenarios the parameters characterising this difference are the same and the same random seed is used; hence, the same distance is observed. The average distance is very low in scenarios A1, A2, B1 and B2, quite high in scenarios D1 and D2, and very high in scenarios C1 and C2.

Although this method seems intuitive it suffers from a few problems. First, the distance between two curves depends on the unit of measurement of data. The distance can be standardised with respect to the unit of measurement, but no thresholds can be defined. The observed distance is also dependent on the domain, since the discretisation step is a multiplying factor in the estimation. It is also highly influenced by the low reliability in one of the two groups as presented in Table 2 (see scenarios C1 and C2). The distance measure provides an overall value, thus with no information of the reliability throughout the domain.

Similarity measures

The similarity measure is an always well-defined dimensionless quantity measured on a specific range, making its interpretation easy. However, similar to CMC and distance measures, it neglects variability between individuals. As a result, very similar similarity measures are obtained for very different scenarios, which for instance is displayed in scenarios A1 and A2. Similar to the distance measure, the similarity measure provides no information about the reliability throughout the domain.

Discussion and implications

This review provides an overview of available methods for test–retest reliability when the data of interest are curves (also called functional data). Overall, we recommend the use of the integrated pointwise indices, specifically ICC and SEM, to investigate reliability of curve data. These methods provide advantages of being straightforward to implement, enable results over the domain, and consider variation between individuals. As such, the methods are relevant for both the researcher and the sports coach.

It is important to realise that an observed curve consists of two parts: a true curve and an error component. The error component corresponds to all different sources of variation, both systematic and random. In the test–retest situation the variation can originate from the individual (fatigue, mood), the test leader (instructions), the test situation (practicalities around the test), the instrument (technical, the approach), and the processing (coding) (Chau, Young, & Redekop, 2005; Gorton, Hebert, & Gannotti, 2009; Sankey et al., 2015; Srinivasan & Mathiassen, 2012). Some sources of variation should always be controlled, e.g., by utilising clear instructions and rigorous routines. Applicable to data collected from humans, Kottner et al. (2011) argued that reliability may be defined as the ratio of variability between persons to the total variability of all measurements in the sample. Put differently, reliability may be defined as the ability of a measurement to separate between these persons. Since biomechanical research often aims to evaluate effects of injury, diagnosis, or level of expertise, a suitable statistical method employed to investigate reliability should incorporate variability both within
and between individuals. Indeed, motor variability in human tasks is considered an important factor theorised to be regulated by the brain to, e.g., avoid overuse injuries (Srinivasan & Mathiassen, 2012). The variability of curve data may in itself be relevant for comparisons between different cohorts. As an example, greater variability for intralimb couplings has been displayed among persons with an injury to the anterior cruciate ligament when compared to asymptomatic persons (Gribbin et al., 2016; Pollard, Stearns, Hayes, & Heiderscheit, 2015). Further, not considering the variability between individuals may increase the risk of concluding that data is reliable when it is in fact not. The evaluation of such erroneously reliable data may result in non-significant findings that may be real if evaluated more appropriately. Such conclusions may be devastating for the targeted population of interest and hinders the development of future knowledge.

Another important aspect that needs special attention is the integrated version of the reliability index along the domain, which is essentially an average of the index values for the considered time points. The interpretation of the integrated pointwise ICCs and SEMs is the same as the original quantities for discrete data. Hence, it is possible to use the same thresholds. However, the choice of appropriate thresholds for discrete data is often somehow arbitrary, and supporting theoretical results are lacking. In addition to the integrated pointwise index, a figure that shows the pointwise index over the curve provides, at least visually, descriptive local information about the shape of the function $R(t)$ along the domain. A problem may arise if data are highly reliable in one part of the domain and highly unreliable in another part of the domain, as shown by the simulated example (see Figures 3 and 4, scenario D1 and D2). Indeed, greater variability in knee moments have been displayed in the weight acceptance phase (0–25% of landing) compared to later periods of the landing phase during a side-cutting task (Sankey et al., 2015). As such, the initial choice of the domain of interest can strongly impact the result and need careful consideration. Applied sport scientists should rather discuss on the obtained results using graphs of pointwise indices instead of only checking if the integrated value is above or below a certain threshold. These graphs highlight particular phases over the domain that may be more or less reliable, thus aiding sports biomechanists and coaches in the evaluation of biomechanical data for movement technique or performances for athletes. Note however that such figures are only descriptive, and not suited for inferential identification of reliable-unreliable phases.

Another issue to consider is the possible misalignment of curve data that may negatively affect the results of a reliability analysis (e.g., phase shift in the start of the swing phase during a gait cycle). Therefore, the key concepts of domain and domain alignment for curve data (Table 1) deserve special attention and should be clearly stated. Therefore, the key concepts of domain and domain alignment for curve data (Table 1) deserve special attention and should be clearly stated. Here, the data are assumed to be aligned prior to the reliability analysis. This is naturally the case for curve data of gait cycles, where the domain is typically expressed as cycle percentage. In other cases, however, the original data are not aligned, and one should carefully perform domain alignment prior to the analysis. For many applications, landmark registration is a suitable method for domain alignment (Moudy, Richter, & Strike, 2018). Indeed, in most cases, it is possible to identify landmarks
on the curves, which often correspond to events of interest. Using landmark registration in connection to pointwise reliability indices has some sound properties. For instance, the pointwise reliability index at the landmark correspond to the value of the reliability index that would be computed on univariate data that only refer to the specific landmark. With careful considerations more flexible methods may also be used (Marron, Ramsay, Sangalli, & Srivastava, 2014). However, the more flexible the alignment technique is, the less interpretable the results tend to be. With a flexible alignment, the concept of domain itself risks being difficult to interpret, since after aligning the curves, the domain is not time nor percentage of the task (except the case when landmark alignment is performed). It is important to emphasise that these key concepts be comprehensively discussed and decided upon in advance of data analysis, and in relation to the specific research question to attain the best possible results.

Finally, it is important to point out that no inferential procedure is presented here. Inference (tests with p-values or confidence bands) for, e.g., the ICC in each point of the domain would require adjustments due to multiple comparison issues. Within the FDA-framework techniques for adjusting p-values have been presented (Abramowicz et al., 2018; Pataky et al., 2013; Vsevolozhskaya et al., 2013). For future studies, such methods may be used in the reliability setting.

Conclusion

This paper presented advantages and disadvantages of the available methods for assessing reliability of curve data in a test–retest setting, in order to provide recommendations for future research. A summary of these findings is presented in Table 3. It is recommended to use the integrated pointwise indices, specifically ICC and SEM, since the methods evaluate reliability of curve data in an appropriate way. The use of such analyses is encouraged in order to avoid drawing potentially erroneous conclusions from the data. R-code are provided as supplementary material to facilitate the implementation of the recommended analysis methods.

Table 3. Summary of the reliability methods. Principal characteristics of the presented reliability measures that can be applicable to curve data.

| Method                          | Advantages                                                                 | Disadvantages                                                                 |
|---------------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Integrated pointwise index      | Straightforward from univariate analysis. Possible to visualise results over the domain. | Potential disadvantages for the chosen univariate measure will transfer to this index. |
| Functional limit of agreement   | Acknowledges the nature of the curve data and enables results over the domain. | The user must decide on the acceptable limits. Does not take into account the variation between individuals. It may be computationally demanding to check assumptions. |
| Coefficient of multiple correlation | Dimensionless measure attaining values in [0,1]. | Computational errors and possibly not defined when the range of motion is small. Does not take into account the variation between individuals. Only presented as a global measure. |
| Distance measure                | Easy to compute and always well-defined.                                     | The user must decide on the acceptable limits. Does not take into account the variation between individuals. Only presented as a global measure. |
| Similarity measure              | Easy to compute, always well-defined and attains values in [−1,1].            | Does not take into account the variation between individuals. Only presented as a global measure. |
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by Vetenskapsrådet [2016-02763].

ORCID

Alessia Pini http://orcid.org/0000-0001-9235-3062
Jonas L Markström http://orcid.org/0000-0002-6339-9544
Lina Schelin http://orcid.org/0000-0001-7917-5687

References

Abramowicz, K., Häger, C. K., Pini, A., Schelin, L., Sjöstedt de Luna, S., & Vantini, S. (2018). Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. Scandinavian Journal of Statistics, 45, 1036–1061. doi:10.1111/sjos.12454

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine, 26, 217–238. doi:10.2165/00007256-199826040-00002

Berchtold, A. (2016). Test–retest: agreement or reliability? Methodological Innovations, 9, 1–7. doi:10.1177/2059799116672875

Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327, 307–310. doi:10.1016/S0140-6736(86)90837-8

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: what is it, and how is it measured? Physiotherapy, 86, 94–99. doi:10.1016/S0031-9406(05)61211-4

Chau, T., Young, S., & Redekop, S. (2005). Managing variability in the summary and comparison of gait data. Journal of Neuroengineering and Rehabilitation, 2, 22. doi:10.1186/1743-0003-2-22

Chia, K., & Sangeux, M. (2017). Quantifying sources of variability in gait analysis. Gait & Posture, 56, 68–75. doi:10.1016/j.gaitpost.2017.04.040

Colyer, S. L., Evans, M., Cosker, D. P., & Salo, A. I. (2018). A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. Sports Medicine-Open, 4, 1–15. doi:10.1186/s40798-017-0116-x

De Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. Journal of Clinical Epidemiology, 59, 1033–1039. doi:10.1016/j.jclinepi.2005.10.015

Donoghue, O. A., Harrison, A. J., Coffey, N., & Hayes, K. (2008). Functional data analysis of running kinematics in chronic Achilles tendon injury. Medicine & Science in Sports & Exercise, 40, 1323–1335. doi:10.1249/MSS.0b013e31816c4807

Dubin, J. A., & Müller, H. G. (2005). Dynamical correlation for multivariate longitudinal data. Journal of the American Statistical Association, 100, 872–881. doi:10.1198/016214504000001989

Elias, A. R., Hammill, C. D., & Mizner, R. L. (2015). Changes in quadriceps and hamstring cocontraction following landing instruction in patients with anterior cruciate ligament reconstruction. Journal of Orthopaedic & Sports Physical Therapy, 45, 273–280. doi:10.2519/jospt.2015.5335
Ford, K. R., Myer, G. D., & Hewett, T. E. (2007). Reliability of landing 3D motion analysis: implications for longitudinal analyses. *Medicine and Science in Sports and Exercise*, 39, 2021–2028. doi:10.1249/mss.0b103e318149332d

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2, 189–210. doi:10.1002/hbm.460020402

Gorton, G. E., III, Hebert, D. A., & Gannotti, M. E. (2009). Assessment of the kinematic variability among 12 motion analysis laboratories. *Gait & Posture*, 29, 398–402. doi:10.1016/j.gaitpost.2008.10.060

Gribbin, T. C., Slater, L. V., Herb, C. C., Hart, J. M., Chapman, R. M., Hertel, J., & Kuenze, C. M. (2016). Differences in hip-knee joint coupling during gait after anterior cruciate ligament reconstruction. *Clinical Biomechanics*, 32, 64–71. doi:10.1016/j.clinbiomech.2016.01.006

Harrison, A. J., Ryan, W., & Hayes, K. (2007). Functional data analysis of joint coordination in the development of vertical jump performance. *Sports Biomechanics*, 6, 199–214. doi:10.1080/14763140701323042

Hébert-Losier, K., Pini, A., Vantini, S., Strandberg, J., Abramowicz, K., Schelin, L., & Häger, C. K. (2015). One-leg hop kinematics 20 years following anterior cruciate ligament rupture: data revisited using functional data analysis. *Clinical Biomechanics*, 30, 1153–1161. doi:10.1016/j.clinbiomech.2015.08.010

Hébert-Losier, K., Schelin, L., Tengman, E., Strong, A., & Häger, C. K. (2018). Curve analyses reveal altered knee, hip, and trunk kinematics during drop–jumps long after anterior cruciate ligament rupture. *The Knee*, 25, 226–239. doi:10.1016/j.knee.2017.12.005

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, 30, 1–15. doi:10.2165/00007256-200030010-00001

Iosa, M., Cereatti, A., Merlo, A., Campanini, I., Paolucci, S., & Cappozzo, A. (2014). Assessment of waveform similarity in clinical gait data: the linear fit method. *BioMed Research International*. doi:10.1155/2014/214156

Jacques, J., & Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8, 231–255. doi:10.1007/s11634-013-0158-y

Kadaba, M. P., Ramakrishnan, H. K., Wootten, M. E., Gainey, J., Gorton, G., & Cochran, G. V. B. (1989). Repeatability of kinematic, kinetic, and electromyographic data in normal adult gait. *Journal of Orthopaedic Research*, 7, 849–860. doi:10.1002/jor.1100070612

Kainz, H., Graham, D., Edwards, J., Walsh, H. P., Maine, S., Boyd, R. N., . . . Carty, C. P. (2017). Reliability of four models for clinical gait analysis. *Gait & Posture*, 54, 325–331. doi:10.1016/j.gaitpost.2017.04.001

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155–163. doi:10.1016/j.jcm.2016.02.012

Kottner, J., Audigé, L., Borron, S., Donner, A., Gajewski, B. J., Hrøbjartsson, A., . . . Steiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64, 96–106. doi:10.1016/j.jclinepi.2010.03.002

Markström, J. L., Schelin, L., & Häger, C. K. (2018). A novel standardised side hop test reliably evaluates landing mechanics for anterior cruciate ligament reconstructed persons and controls. *Sports Biomechanics*. doi:10.1080/14763141.2018.1538385

Marron, J. S., Ramsay, J. O., Sangalli, L. M., & Srivastava, A. (2014). Statistics of time warpings and phase variations. *Electronic Journal of Statistics*, 8, 1697–1702. doi:10.1214/14-EJS901

McGinley, J. L., Baker, R., Wolfe, R., & Morris, M. E. (2009). The reliability of three-dimensional kinematic gait measurements: a systematic review. *Gait & Posture*, 29, 360–369. doi:10.1016/j.gaitpost.2008.09.003

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. doi:10.1037/1082-989X.1.1.30

Meldrum, D., Shouldice, C., Conroy, R., Jones, K., & Forward, M. (2014). Test–retest reliability of three dimensional gait analysis: including a novel approach to visualising agreement of gait
cycle waveforms with Bland and Altman plots. *Gait & Posture*, 39, 265–271. doi:10.1016/j.gaitpost.2013.07.130

Milner, C. E., Westlake, C. G., & Tate, J. J. (2011). Test–retest reliability of knee biomechanics during stop jump landings. *Journal of Biomechanics*, 44, 1814–1816. doi:10.1016/j.jbiomech.2011.04.005

Moudy, S., Richter, C., & Strike, S. (2018). Landmark registering waveform data improves the ability to predict performance measures. *Journal of Biomechanics*, 78, 109–117. doi:10.1016/j.jbiomech.2018.07.027

Pataky, T. C., Robinson, M. A., & Vanrenterghem, J. (2013). Vector field statistical analysis of kinematic and force trajectories. *Journal of Biomechanics*, 46, 2394–2401. doi:10.1016/j.jbiomech.2013.07.031

Pollard, C. D., Stearns, K. M., Hayes, A. T., & Heiderscheit, B. C. (2015). Altered lower extremity movement variability in female soccer players during side-step cutting after anterior cruciate ligament reconstruction. *American Journal of Sports Medicine*, 43, 460–465. doi:10.1177/0363546514560153

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. New York, NY: Springer.

Rankin, G., & Stokes, M. (1998). Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation*, 12, 187–199. doi:10.1191/026921598672178340

Richter, C., O’Connor, N. E., Marshall, B., & Moran, K. (2014). Comparison of discrete-point vs. dimensionality-reduction techniques for describing performance-related aspects of maximal vertical jumping. *Journal of Biomechanics*, 47, 3012–3017. doi:10.1016/j.jbiomech.2014.07.001

Roislien, J., Rennie, L., & Skaaret, I. (2012). Functional limits of agreement: a method for assessing agreement between measurements of gait curves. *Gait & Posture*, 36, 495–499. doi:10.1016/j.gaitpost.2012.05.001

Roislien, J., Skare, Ø., Opheim, A., & Rennie, L. (2012). Evaluating the properties of the coefficient of multiple correlation (CMC) for kinematic gait data. *Journal of Biomechanics*, 45, 2014–2018. doi:10.1016/j.jbiomech.2012.05.014

Ryan, W., Harrison, A., & Hayes, K. (2006). Functional data analysis of knee joint kinematics in the vertical jump. *Sports Biomechanics*, 5, 121–138. doi:10.1080/14763141.2006.9628228

Sangalli, L. M., Secchi, P., Vantini, S., & Vitelli, V. (2010). K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54, 1219–1233. doi:10.1016/j.csda.2009.12.008

Sangeux, M., Passmore, E., Graham, H. K., & Tiros, O. (2016). The gait standard deviation, a single measure of kinematic variability. *Gait & Posture*, 46, 194–200. doi:10.1016/j.gaitpost.2016.03.015

Sankey, S. P., Azidin, R. M. F. R., Robinson, M. A., Malfait, B., Deschamps, K., Verschueren, S., . . . Vanrenterghem, J. (2015). How reliable are knee kinematics and kinetics during side-cutting manoeuvres? *Gait & Posture*, 41, 905–911. doi:10.1016/j.gaitpost.2015.03.014

Schwartz, M. H., Trost, J. P., & Wervey, R. A. (2004). Measurement and management of errors in quantitative gait data. *Gait & Posture*, 20, 196–203. doi:10.1016/j.gaitpost.2003.09.011

Sinsurin, K., Vachalathiti, R., Srisangboriboon, S., & Richards, J. (2018). Knee joint coordination during single-leg landing in different directions. *Sports Biomechanics*. doi:10.1080/14763141.2018.1510024

Sorensen, H., Goldsmith, J., & Sangalli, L. M. (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine*, 32, 5222–5240. doi:10.1002/sim.5989

Srinivasan, D., & Mathiassen, S. E. (2012). Motor variability in occupational health and performance. *Clinical Biomechanics*, 27, 979–993. doi:10.1016/jclinbiomech.2012.08.007

Trevethan, R. (2017). Intraclass correlation coefficients: clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, 17, 127–143. doi:10.1007/s10742-016-0156-6
Vsevolozhskaya, O. A., Greenwood, M. C., Bellante, G. J., Powell, S. L., Lawrence, R. L., & Repasky, K. S. (2013). Combining functions and the closure principle for performing follow-up tests in functional analysis of variance. *Computational Statistics & Data Analysis, 67*, 175–184. doi:10.1016/j.csda.2013.05.005

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A., Bargary, N., & Smith, R. (2019a). Bivariate functional principal components analysis: considerations for use with multivariate movement signatures in sports biomechanics. *Sports Biomechanics, 18*, 10–27. doi:10.1080/14763141.2017.1384050

Warmenhoven, J., Cobley, S., Draper, C., Harrison, A., Bargary, N., & Smith, R. (2019b). Considerations for the use of functional principal components analysis in sports biomechanics: Examples from on-water rowing. *Sports Biomechanics, 18*, 317–341. doi:10.1080/14763141.2017.1392594

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research, 19*, 231–240. doi:10.1519/15184.1