


ORIGINAL RESEARCH ARTICLE

Open Access



# Stressing the Relevance of Differentiating between Systematic and Random Measurement Errors in Ultrasound Muscle Thickness Diagnostics

Lars Hubertus Lohmann<sup>1,2\*</sup> , Martin Hillebrecht<sup>1</sup>, Stephan Schiemann<sup>3</sup> and Konstantin Warneke<sup>3,4</sup>

## Abstract

**Background** The majority of studies that explore changes in musculature following resistance training interventions or examine atrophy due to immobilization or sarcopenia use ultrasound imaging. While most studies assume acceptable to excellent reliability, there seems to be unawareness of the existing absolute measurement errors. As early as 1998, methodological research addressed a collective unawareness of the random measurement error and its practical indications. Referring to available methodological approaches, within this work, we point out the limited value of focusing on relative, correlation-based reliability indices for the interpretability in scientific research but also for clinical application by assessing 1,512 muscle thickness values from more than 400 ultrasound images. To account for intra- and inter-day repeatability, data were collected on two consecutive days within four testing sessions. Commonly-stated reliability values (ICC, CV, SEM and MDC) were calculated, while evidence-based agreement analyses were applied to provide the accompanied systematic and random measurement error.

**Results** While ICCs in the range of 0.832 to 0.998 are in accordance with the available literature, the mean absolute percentage error ranges from 1.34 to 20.38% and the mean systematic bias from 0.78 to 4.01 mm (all  $p \leq 0.013$ ), depending on the measurement time points chosen for data processing.

**Conclusions** In accordance with prior literature, a more cautious interpretation of relative reliability values should be based on included systematic and random absolute measurement scattering. Lastly, this paper discusses the rationale for including different measurement error statistics when determining the validity of pre-post changes, thus, accounting for the certainty of evidence.

## Key Points

- While reliability of a testing protocol is most often determined via relative reliability indices such as the intraclass correlation coefficient, further reliability values such as the systematic bias and the random error have been described as valuable for results interpretation.

\*Correspondence:  
Lars Hubertus Lohmann  
lohmannlars@gmx.de

Full list of author information is available at the end of the article

- This study used the most frequently employed procedure to determine muscle hypertrophy (ultrasound) as an example linking relative reliability values to absolute measurement errors under special consideration of appropriate calculation models using three scenarios (best case, worst case, stability).
- Overall, 504 ultrasound images were examined showing excellent relative reliability, but the corresponding measurement errors suggest that caution must be exercised when interpreting pre-post settings in cases where the measurement error exceeds the expected changes.

**Keywords** Ultrasound, Reliability, Training intervention studies, Mean absolute percentage error, Expected effect

## Background

Due to its high importance in rehabilitation and prevention, several exercise training programs were designed to induce muscle hypertrophy in healthy participants or after injury [1], while the muscle thickness/cross-sectional area are considered of utmost importance when quantifying age-related sarcopenia [2].

Assuming training-induced muscle mass increases of  $7.6 \pm 1.2\%$  ( $d = 0.47 \pm 0.08$ ) in intervention periods of up to 13 weeks [3], a highly sensitive, and therefore reliable as well as reproducible procedure for data collection is strongly recommended in sports medicine and science to preclude measured differences being the result of measurement errors [4]. While described as the gold standard method, magnetic resonance imaging [5] is frequently substituted by ultrasound muscle thickness evaluations as the literature suggests high validity and reliability, while being portable and cost- as well as time-efficient [5–7].

Notwithstanding, concerns arose regarding the objectivity of using ultrasound due to applied pressure to soft tissue, lack of probe angle standardization and lack of agreement with muscle cross-sectional area values from magnetic resonance imaging [8]. As early as 1998, Atkinson & Nevill [9] as well as Lamb [10] drew attention to unsatisfactory reliability when validating measurement procedures. Additionally, de Vet et al. [11] as well as Kottner et al. [12] highlighted that the context of a given measurement set-up is of utmost importance, stressing the relevance of using agreement and not reliability measures to quantify the magnitude of measurement error when evaluating changes over time.

Even though 25 years have passed since Atkinson & Nevill [9] as well as Lamb [10] published their respective papers, it appears that an unawareness of the detailed quantification and evaluation of systematic and random measurement errors still exists in sports medicine and science. This is because reliability and repeatability are most often solely stated on the basis of correlations (i.e. intraclass correlation coefficient (ICC)) and its derivatives (such as the standard error of the mean (SEM) or minimal detectable change (MDC)) [4, 13] as can be seen in a systematic review concerned with ultrasound reliability by Nijholt et al. [7]. Relative reliability indices, expressed as correlation coefficient-based statistical

parameters, focus on the relationship between two values with or without accounting for variance and do not distinguish (in a sense of separate quantification) between systematic and random error [4, 9].

For all users of a specific measurement method, practitioners such as therapists and medical staff or researchers, it is of paramount importance to be able to distinguish systematic bias (error arising from, e.g., habituation, familiarization or in ultrasound from muscle swelling or water content increases) from random error (unsystematic scattering from, for example, different probe pressure or angle) when interpreting results [4, 13]. While commonly-used reliability indices seem relevant for assessing relative reliability [4], Lamb [10] has impressively delineated the limitations of correlation-based reliability calculation methods for interpretability regarding the repeatability of the testing procedure.

Therefore, in this study, we aimed to apply the commonly used (also considered standard) methods for reliability calculations in sports science and medicine research and oppose these methods to those proposed, inter alia, in the articles by Barnhart et al. [4], Hopkins [13] and Atkinson & Nevill [9]. Therefore, after firstly calculating the ICC, SEM and MDC with those formulas most commonly employed in current original sports science and medicine research, secondly, the corresponding systematic and random errors, arising from test-retest performance (i.e., intra- and inter-day reliability), will be provided to raise awareness of the strengths and weaknesses of the commonly used reliability reporting methods.

Accordingly, to provide a well-balanced perspective on the repeatability and stability of ultrasound muscle thickness data collection, three different scenarios of measurement error calculation will be presented, stressing the relevance of reporting the random error when performing diagnostics.

## Methods

### Experimental Set-up

Data collection was performed on two consecutive days, including 2 test sessions each day (4 test sessions in total), while assuming no meaningful exercise-induced morphological adaptations within 48 h. Muscle thickness images were acquired via B-mode ultrasound once

in the morning and once in the (late) afternoon of both these days. The muscles investigated are the vastus lateralis (VL), the lateral head of the gastrocnemius (GL) and the medial head of the gastrocnemius (GM) – chosen as these exhibit some of the highest ICC values stated in the literature for ultrasound muscle thickness measurements and are frequently investigated in training intervention studies [6, 14, 15].

In total, 504 images from 21 participants (see *Image acquisition* and *Participants* sections) and thus 168 images per muscle were used for the calculation (21 participants  $\times$  4 measurement time points  $\times$  3 muscles  $\times$  2 images per muscle). Since all images comprised three muscle thickness determinations across the width of the image (left, middle, right), the calculations are based on a total of 1,512 muscle thickness determinations. Data were collected by the same experienced investigator (LHL) who has been involved in extensive B-mode ultrasound image acquisition for muscle thickness determination in various chronic static stretching intervention studies [16, 17]. Figure 1 shows a flow-chart illustrating how the experiment was conducted.

### Image Acquisition

The B-mode ultrasound images were acquired using a MyLab™ Gamma ultrasound device with a 5 cm wide SL1543 linear probe (Esaote Biomedica DE GmbH, Cologne, Germany) operating at a frequency range of 3 to 13 MHz with image acquisition in the longitudinal direction. To ensure using the same spots for the repeated measurements, all spots were marked with a water-resistant sharpie and re-painted in each session.

For VL measurements (in the right leg), the participants adopted a seated position with the knees slightly over the edge of a massage bench to ensure no contraction in the quadriceps musculature. For GM and GL measurements (in the left leg), the participants assumed a prone position on the same massage bench with their feet hanging slightly over the edge of the bench to ensure no contraction in the calf muscles.

Muscle thickness was defined as the distance between the superficial and deep aponeuroses of a muscle. The spots used for the ultrasound muscle thickness measurements on the right VL as well as left GL and GM were determined following two criteria: (1) clear image and (2) superficial and deep aponeuroses as parallel as possible to ensure that the measurement point was not close to a muscle-tendon junction.

To counteract potential variations within a single ultrasound picture and minimize assessment limitations, for each muscle, muscle thickness was calculated as the mean of three distances between the upper and lower fascia in each picture, leading to 1,512 muscle thickness values (504 pictures  $\times$  3 muscle thickness determinations).

Image processing was performed via ImageJ (version 1.53t, National Institutes of Health, Bethesda, MD, USA) which is illustrated in Fig. 2.

### Participants

To account for the widespread application in different clinical settings including heterogeneous performance level, sexes/gender and anthropometric parameters, the included participants' attributes ranged from sedentary lifestyle with no training history to strength training seven days per week (bodybuilder with a body mass of 125 kg). Therefore, the age, height, mass and body mass index ranged from 20 to 65 years ( $33.9 \pm 14.2$  years), 168 to 195 cm ( $180.71 \pm 7.34$  cm), 66 to 130 kg ( $86.95 \pm 19.5$  kg) and 21.46 to 40.12 kg/m<sup>2</sup> ( $26.48 \pm 4.9$  kg/m<sup>2</sup>), respectively, for the 13 male and 8 female participants. All participants provided written informed consent for participation in the study which was conducted in accordance with the Declaration of Helsinki and approved by the Oldenburg Medical Ethics Committee (2021-089).

### Data Analysis

In the first step, commonly-used reliability parameters were calculated using SPSS 29 (IBM Deutschland GmbH, Ehningen, Germany) and Microsoft Excel (Microsoft Corp., Redmond, WA, USA). These include:

- 1) ICC two-way mixed model for consistency [18].

$$ICC = (MS_R - MS_E) / MS_R \quad (1)$$

- 2) ICC two-way mixed effects model for absolute agreement [18].

$$ICC = (MS_R - MS_E) / (MS_R + (MS_C - MS_E) / n) \quad (2)$$

- 3) The coefficient of variability (CV).

$$CV = \frac{SD}{Mean} * 100 \quad (3)$$

- 4) The standard error of measurement for consistency ( $SEM_{consistency}$ ).

$$SEM_{consistency} = SD * \sqrt{1 - ICC_{consistency}} \quad (4)$$

- 5) The SEM for agreement ( $SEM_{agreement}$ ).



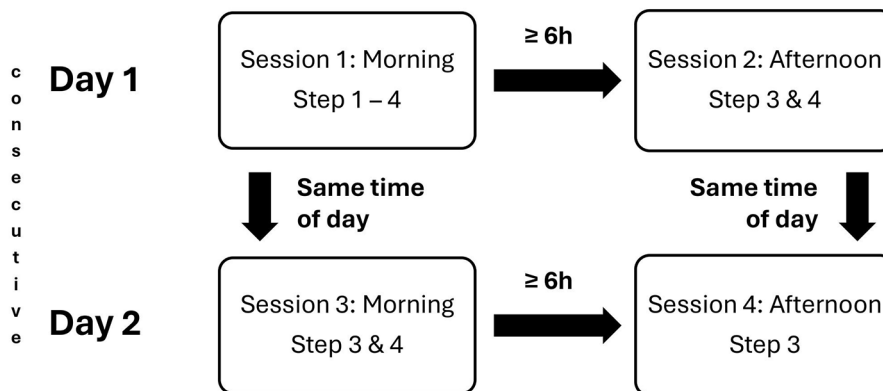
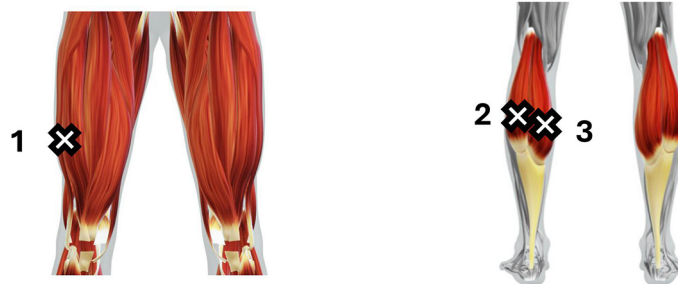
**Recruitment**

Recruiting convenience sample of 21 subjects. Allocating four time slots per subject ensuring sessions are ≥ 6h apart and both morning and afternoon sessions take place during the same time of day on two consecutive days.



**Data collection**

Step 1: Determining the spots used for ultrasound image acquisition  
 Step 2: Marking the spots with water-resistant sharpie  
 Step 3: Image acquisition in following order: VL (1), GL (2), GM (3)  
 Step 4: Re-marking the spots with water-resistant sharpie



**Data processing & analysis**

Thickness measurements and statistical analysis conducted after data collection was completed for all subjects.

**Fig. 1** Flow-chart showing how the experiment was conducted

$$SEM_{agreement} = \sqrt{(\sigma_{observations}^2 + \sigma_{residual}^2)} \tag{5}$$

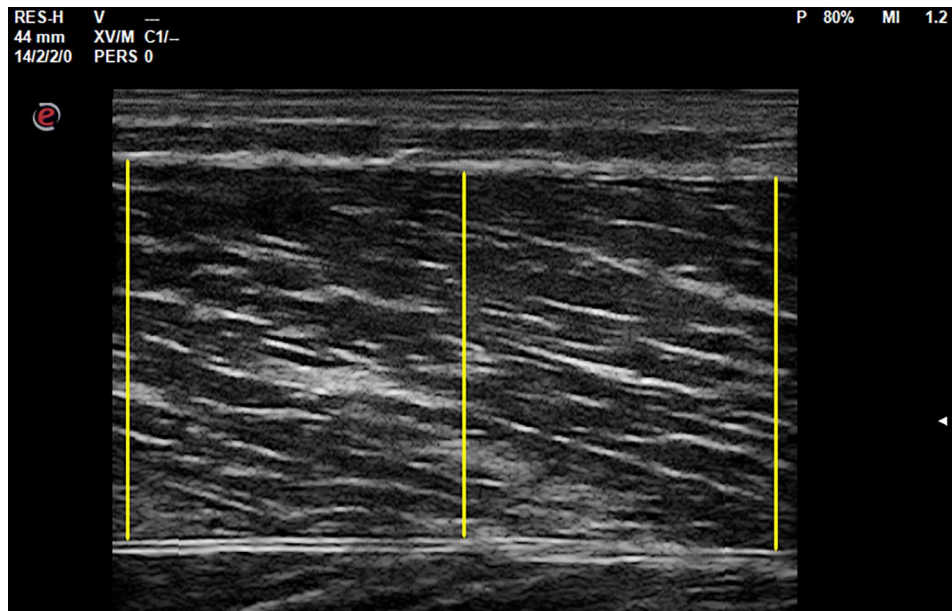
6) The minimal detectable change (MDC).

$$MDC = SEM * 1.96 * \sqrt{2} \tag{6}$$

columns,  $MS_E$  = mean square for error,  $MS_R$  = mean square for rows,  $n$  = number of subjects,  $SD$  = standard deviation,  $\sigma_{observations}^2$  = the variance in observations,  $\sigma_{residual}^2$  = residual variance being the interaction between subjects and observations

Noteworthy, the SEM and MDC share most of the measures with the ICC as they are all based on variability values that stem from analysis of variance (ANOVA) calculations. While the  $SEM_{consistency}$  and the “consistency”-based MDC are not suitable to assess agreement between

The terms used in the above equations are: ICC=intra-class correlation coefficient,  $MS_C$  = mean square for



**Fig. 2** Illustration of how three distances between the upper and lower fascia of the respective muscles across the width of an image (left, middle, right) were used to determine the mean muscle thickness for each acquired image

measurements, they are still commonly used within sports science and medicine research (see, e.g., [19–22]). Therefore, these parameters are included in the following analyses and supplemented by the “agreement”-based variations. The MDC is generally calculated with the same formula irrespective of its use being within consistency or agreement settings and will be listed separately.

In advance of the following calculations, the construct of reliability must be discussed. We aimed to explore repeatability as the basis of all further reliability models, meaning that the same investigator assessed the same parameter on the same subject, just at a different time point. Assuming no further variation in the testing conditions, using a reliable and valid measurement tool, maximal *agreement* between the values can reasonably be assumed. Accordingly, Barnhart, Haber & Lin [4] provide an overview of different assumptions and calculation models to assess repeatability in measurements. To account for the random error, including the variance of individual courses providing a range of the typical error, Hopkins [13] described it as the mean of the standard deviations (SD) divided by  $\sqrt{2}$ . Assuming heteroscedasticity in most sports science and medicine-related parameters, the absolute typical error (TE) usually increases with higher performance levels [23], and the statement of the percentage of the TE can thus be assumed beneficial [13]. Therefore, the TE as well as the CV of the TE ( $CV_{TE}$ ) are also provided in Table 1. A further agreement analysis considering the individual deviations of the mean was provided by Bland & Altman [24], graphically illustrating the systematic bias (which is equal to the mean differences of the paired t-test applied to the

data of interest) with the scatter of individual plots. Furthermore, the limits of agreement (LoA) are included to this graphical evaluation. Consequently, to assign the TE, mean absolute error (MAE), the mean absolute percentage error (MAPE) as well as the graphical illustration of the random error to the commonly stated reliability measures, these values were additionally added to Table 1. The level of significance for the mean systematic bias via paired t-test was set at  $p < 0.05$ .

$$MAE = \frac{1}{n} * \sum_{i=1}^n |x_i - y_i| \quad (7)$$

$$MAPE = \frac{1}{n} * \sum_{i=1}^n \left| \frac{x_i - y_i}{x_i} \right| * 100 \quad (8)$$

$$TE = SD(x_i - y_i) / \sqrt{2} \quad (9)$$

$$CV_{TE} = \frac{TE}{Mean} * 100 \quad (10)$$

The terms used in the above equation are:  $n$ =number of subjects,  $x_i$  = value of measurement 1,  $y_i$  = value of measurement 2,  $SD$ =standard deviation,  $TE$ =typical error

The Bland-Altman plot stems from the JAMOVI software (version 2.3.28) using the ‘blandr’ module.

To control the data for a possible influence of bodyfat on imaging quality, the Pearson product-moment correlation coefficient values for the subgroups *normal body-mass-index* vs. *overweight* as well as *male* vs. *female* were z-transformed according to the Fisher method. The

**Table 1** Absolute error statistics based on the ultrasound-derived muscle thickness values acquired during four measurement time points on two consecutive days

Muscle	Comparison	CV (in %)	TE (in mm)	CV <sub>TE</sub> (in %)	MAE (in mm)	MAPE (in %)	Mean systematic bias (95% CI) (in mm)	LoA of mean systematic bias (95% CI) (in mm)
VL	Highest vs. sec. highest	1.96	0.92	3.45	0.77	2.67	0.78 (0.18, 1.37) <i>p</i> =0.013*	-1.78 (-2.81 – -0.75) – 3.33 (2.3–4.36)
	Highest vs. lowest	11.79	1.35	5.06	4.06	15.11	4.01 (3.19, 4.93) <i>p</i> <0.001**	0.32 (-1.19–1.83) – 7.8 (6.29–9.31)
	Highest vs. mean	5.61	0.923	3.46	2.06	7.54	2.06 (0.28, 1.47) <i>p</i> <0.001**	-0.5 (-1.53–0.53) – 4.62 (3.59–5.65)
GL	Highest vs. sec. highest	3.36	0.358	2.37	0.69	4.58	0.69 (0.46, 0.92) <i>p</i> <0.001**	-0.3 (-0.7–0.1) – 1.68 (1.28–2.08)
	Highest vs. lowest	16.27	0.729	4.83	3.23	20.38	3.24 (2.77, 3.7) <i>p</i> <0.001**	1.21 (0.4–2.03) – 5.26 (4.44–6.07)
	Highest vs. mean	8.2	0.459	3.04	1.71	10.86	1.71 (1.41, 2) <i>p</i> <0.001**	0.43 (-0.08–0.95) – 2.98 (2.47–3.5)
GM	Highest vs. sec. highest	0.96	0.239	1.14	0.31	1.34	0.31 (0.15, 0.46) <i>p</i> <0.001**	-0.36 (-0.62 – -0.09) – 0.97 (0.7–1.24)
	Highest vs. lowest	6.25	0.942	4.51	1.91	8.39	1.91 (1.3, 2.51) <i>p</i> <0.001**	-0.71 (-1.76–0.35) – 4.52 (3.46–5.57)
	Highest vs. mean	3.04	0.544	2.6	0.97	4.18	0.97 (0.62, 1.32) <i>p</i> <0.001**	-0.54 (-1.15–0.07) – 2.48 (1.87–3.09)

ICC=Intraclass correlation coefficient, CV=Coefficient of variation, TE=Typical error, CV<sub>TE</sub>=Typical error expressed as a coefficient of variation, MAE=Mean absolute error, MAPE=Mean absolute percentage error, LoA=Limits of agreement, (95% CI)=95% confidence interval, Highest=Highest muscle thickness value, sec. highest=Second highest muscle thickness value, lowest=Lowest muscle thickness value, mean=Mean muscle thickness value across eight images, \* = significant *p*<0.05, \*\* = significant *p*≤0.001

**Table 2** Muscle thickness measurement characteristics

Muscle	N	Total number of images	Min (in mm)	Max (in mm)	M (95% CI) ± SD (in mm)
Vastus lateralis	21	168	11.7	53.4	26.66 (25.25, 28.07) ± 9.26
Lateral head of gastrocnemius	21	168	6.8	30.1	15.1 (14.26, 15.93) ± 5.55
Medial head of gastrocnemius	21	168	14.3	37.2	20.89 (20.14, 21.64) ± 4.91

N=Number of participants, Min=Minimal muscle thickness value, Max=Maximal muscle thickness value, mm=Millimeter, M±SD=Mean±standard deviation

Benjamini-Hochberg procedure was used to control the study-wise false discovery rate with a significance value of 0.05 [25]. The analysis yielded no significant differences in relationships of these parameters for the subgroups. Additionally, the review by Nijholt et al. [7] found no differences in the reliability of ultrasound measurements between older and younger participants.

### Three Scenario Calculation

Commonly, reliability values are calculated using the best and the second-best value available (scenario 1). However, when aiming to provide objectively reported results for practical useful information [26], this procedure can be exclusively performed in very stable measurement procedures, especially if the real muscle thickness is not known and can exclusively be determined by using the performed procedure [4]. To illustrate, when measuring a muscle thickness of 5 mm in trial 1 and 6 mm in trial 2, can we assume the real value to be 5 mm, 6 mm (20% increase compared to trial 1) or 5.5 mm? Since the real values are unknown, the stability of the measurement

provides a range of the true measurement errors, leading to a statement about the precision of the measurement. However, from a scientific point of view, we cannot exclusively use the best-case scenario but should also consider the probability of personal errors. Thus, a well-balanced perspective requires providing the worst-case scenario as well (scenario 2). Additionally, accounting for the stability of the measurement (and to weaken the worst-case scenario), our third scenario provides the best measurement value compared to the mean across all measurement values (scenario 3). The best and worst measurements represent the highest and lowest muscle thickness values, respectively.

### Results

Table 2 reports the descriptive statistics for the muscle thickness measurements.

#### Best-case Scenario (Scenario 1)

Using the best and second-best value, the best-case scenario exhibits ICCs for agreement and consistency that

would be classified as excellent according to the current literature, ranging from 0.988 to 0.998 with CV values from 0.96 to 3.36% and SEMs ranging from 0.01 to 0.12 mm for consistency and from 1.03 to 2.67 mm for agreement measures. Since the MDC is strongly related to the SEM, just multiplied by a fixed factor, we do not additionally list this value but report it in Table 3 only. The listed relative reliability values correspond, exemplarily for the VL, to an absolute, systematic measurement error (mean systematic bias) based on the test-retest procedure of 0.78 mm ( $p=0.013$ ) with LoAs ranging from  $-1.78$  to  $3.33$  mm. Therefore, accounting for the random individual scattering, the typical error occurring from repeated measurements calculated in the best-case scenario for VL, GL and GM shows noise of 0.239–0.92 mm for TE, 0.31–0.77 mm for MAE and 1.34–4.58% for MAPE. Test-retest values with LoAs for GL and GM as well as 95% confidence intervals are reported in Tables 1 and 3.

#### Worst-case Scenario (Scenario 2)

In contrast to the best-case scenario, scenario 2 compare the highest and lowest muscle thickness values, which, logically, exhibits the largest deviation in muscle thickness measured within the two days and four measurement time points. These comparisons still result in relative reliability with ICCs ranging from 0.965 to 0.982 for consistency and 0.832 to 0.903 for agreement with CV values from 6.25 to 16.27% and SEMs ranging from 0.14 to 0.27 mm for consistency and from 6.25 to 13.23 mm

for agreement. The systematic bias calculated via test-retest procedure and Bland-Altman analysis yields a 3.24 mm ( $p<0.001$ ) difference with LoA of 2.77–3.7 mm for the GL. The calculated TE shows an expected range of 0.729–1.35 mm, while MAE and MAPE are quantified in ranges of 1.91–4.06 mm and 8.39–20.38%, respectively. Test-retest values with LoA for GL and GM as well as 95% confidence intervals are reported in Tables 1 and 3.

#### Measurement Stability (Scenario 3)

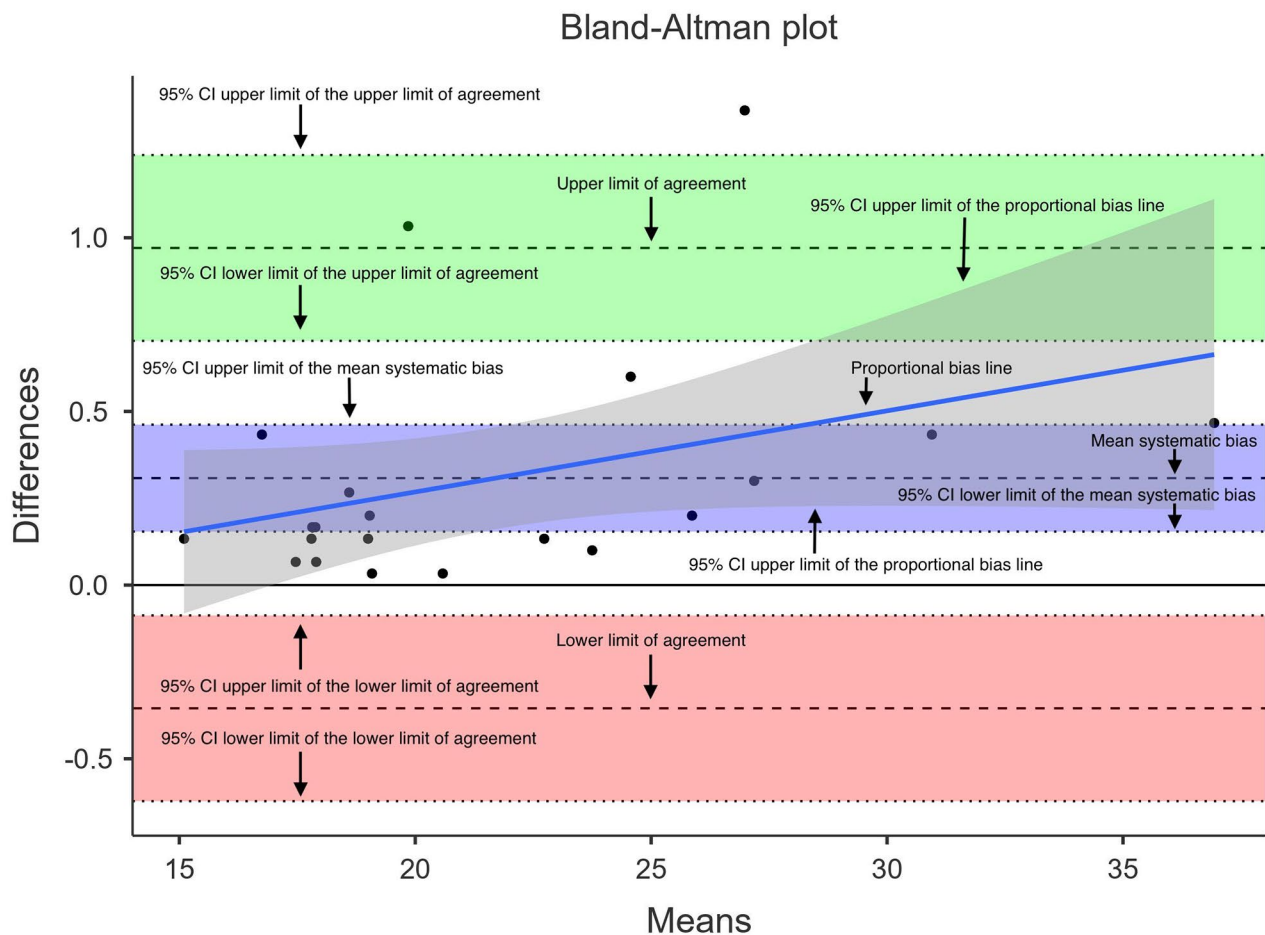
Hypothesizing both previous scenarios to be unrealistic and in an attempt to include measurement stability, the third scenario uses the best value and the mean of the measurements, resulting in ICCs ranging from 0.989 to 0.993 for consistency and 0.947 to 0.973 for agreement with CV values from 3.04 to 8.2% and SEMs ranging from 0.05 to 0.12 mm for consistency and from 3.19 to 6.74 mm for agreement. The test-retest procedure and Bland-Altman analysis states a systematic measurement bias of 2.06 mm ( $p<0.001$ ) with LoA of 0.28–1.47 mm. For VL, GL and GM, the TEs range from 0.239 to 1.35 mm, MAEs from 0.97 to 2.06 mm and MAPEs from 4.18 to 10.86%. Test-retest values with LoA for GL and GM as well as 95% confidence intervals are reported in Tables 1 and 3.

Figure 3 provides an example for the best-case scenario for GM as a Bland-Altman plot.

**Table 3** Relative, correlation-based reliability based on the ultrasound-derived muscle thickness values acquired during four measurement time points on two consecutive days

Muscle	Comparison	ICC <sub>consistency</sub> (95% CI)	ICC <sub>agreement</sub> (95% CI)	SEM <sub>consistency</sub> (in mm)	SEM <sub>agreement</sub> (in mm)	MDC <sub>consistency</sub> (in mm)	MDC <sub>agreement</sub> (in mm)
VL	Highest vs. sec. highest	0.991 (0.978, 0.996)	0.988 (0.96, 0.996)	0.12	2.67	0.34	7.41
	Highest vs. lowest	0.98 (0.951, 0.992)	0.898 (0, 0.978)	0.27	13.23	0.75	36.66
	Highest vs. mean	0.991 (0.977, 0.996)	0.968 (0.347, 0.993)	0.12	6.74	0.34	18.68
GL	Highest vs. sec. highest	0.996 (0.99, 0.998)	0.988 (0.75, 0.997)	0.03	2.26	0.09	6.27
	Highest vs. lowest	0.982 (0.955, 0.993)	0.832 (0, 0.964)	0.14	10.51	0.38	29.13
	Highest vs. mean	0.993 (0.983, 0.997)	0.947 (0.019, 0.99)	0.05	5.55	0.15	15.39
GM	Highest vs. sec. highest	0.998 (0.995, 0.999)	0.997 (0.974, 0.999)	0.01	1.03	0.04	2.84
	Highest vs. lowest	0.965 (0.916, 0.986)	0.903 (0.132, 0.976)	0.25	6.25	0.69	17.32
	Highest vs. mean	0.989 (0.973, 0.996)	0.973 (0.609, 0.993)	0.08	3.19	0.22	8.84

ICC=Intraclass correlation coefficient, SEM=Standard error of measurement, MDC=Minimal detectable change, (95% CI)=95% confidence interval, Highest=Highest muscle thickness value, sec. highest=Second highest muscle thickness value, lowest=Lowest muscle thickness value, mean=Mean muscle thickness value across eight images



**Fig. 3** Bland-Altman plot for the best-case scenario of the medial head of the gastrocnemius. The quantification of the systematic error (mean difference) as well as the random scattering that illustrates the random error/secondary variance provide crucial information beyond information on relative reliability. In accordance with Carstensen et al. [27], the limits of agreement provide a range in which 95% of the measurements could be expected when repeating the measurement via the same devices in the same population. Especially the random error should be considered as highly important in ultrasound as it might indicate unsystematic standardization problems (e.g., different probe angles, different measurement spots, differences in applied pressure [8]), while the systematic error could be attributed to, for example, muscle swelling or increased water content in measurements conducted in the evening. Systematic bias=mean difference between mean 1 and mean 2, random error=scattering around the systematic error, lower and upper limit of agreement=reference interval or normal range for the test-retest differences expected for 95% of individuals causing a probability statement for expected values [28]

## Discussion

Around 20 to 25 years ago, several authors [4, 10] had already stressed the paramount importance of not focusing solely on relative errors, considering means and standard deviations, but rather shifting the focus to random measurement errors, especially when addressing clinical and practical applicability. However, the majority of the literature still almost exclusively reports ICCs (sometimes the CV), the SEM/MDC, while collectively neglecting the random scattering of individual value pairs, arising from repeated measurements. Consequently, the present study was designed to evaluate the commonly-used relative error values and to additionally provide recommended random error parameters. With ICCs ranging from 0.832 to 0.998, the data collection showed

comparable reliability to the current ultrasound literature [7]. Depending on the scenario, we found significant (all  $p < 0.05$  and all but one  $p < 0.001$ ) systematic error as represented by the mean systematic bias and the corresponding LoA despite the small sample size.

In scientific settings as well as in clinical practice, the use of precise and accurate measurements is of critical importance. The criteria objectivity, validity and reliability are commonly known as preconditions for the further use of collected data. However, there seems to be no consensus about the classification of the aforementioned criteria. While mostly referring to Cohen's [29] classifications, it seems that authors neglect important aspects. Firstly, the suggested classifications are based on assumptions from mostly behavioral and psychological sciences.



Secondly, it is clearly described that classifications of reliability must always be viewed in the light of the setting in which they are applied [30, 31]. Using correlation-based reliability values, it seems reasonable to consider two aspects. On the one hand, as mentioned above, the true value is not known. Consequently, the better the reliability, the closer the LoA of the measurements and, thus, the scatter range of individual deviations decreases. On the other hand, the expected or measured pre-post change of a measurement tool provides the relevance of the random measurement error, as the systematic measurement error (a mean error shift over- or under-calculating the real value by repeating the measurement without surrounding scattering) could be solved by adding a fixed factor to the formula. Therefore, relating to reliability, repeatability (intra- and inter-day reliability) can be described as a value of measurement precision and vice versa measurement precision a value of repeatability [4]. Additionally, as already described by Lamb [10], using a measurement tool with a systematic and random error can not be assumed to be either objective or valid. Nevertheless, it is still mandatory to attribute measured noise as well as systematic bias to the related circumstances and context.

Random errors in ultrasound include differences in water content in the musculature due to variations in hydration, activity level on the measurement day and possibly the days prior but also the applied pressure with which the transducer is placed on the skin, sub-optimal standardization of the measurement point etc [8]. This list is not exhaustive but already highlights many different possible influences affecting the outcome.

Regardless of the resulting measurement error, the further relevance of the variability magnitude in sports science arises from the expected increase in intervention studies. The literature indicates muscle hypertrophy effects at around  $7.6 \pm 1.2\%$  in response to up to 13 weeks of resistance training [3] while Goodpaster et al. [32] quantified the age-related loss of skeletal muscle mass to be around 1% per year within a 3-year span in a study sample of 1,880 subjects with a mean age of  $73.5 \pm 2.8$  years. Even though most intervention studies are controlled via a passive control group and assuming no statistically significant changes from pre- to post-test (in which the same measurement error could be assumed), the repeatability values might not be sufficient to prove a difference between groups in general, implicating that a more cautious interpretation of increases is needed. Therefore, contrasting the measurement errors of the best-case, worst-case and stability scenario to changes of  $7.6 \pm 1.2\%$  in resistance-training studies and 1% per year in sarcopenia-related atrophy, the question arises about the real pre-post changes.

When drawing conclusions on a bigger scale, this would encourage rating reliability on agreement measures

(such as absolute agreement ICCs) and adjusting the classification based on the expected effect (size) as well as the expected measurement error, which would make the assessment more meaningful. A similar approach is already in effect in meta-analyses and other review articles when quality of evidence and strength of recommendation are judged based on a framework. A good example is the renowned GRADE framework [33] that first grades the level of evidence as high for randomized trials, low for observational studies and very low for any other evidence, after which the level is adjusted, decreasing, e.g., with serious limitations in study quality, imprecise data or high probability of reporting bias, but also increasing *inter alia* with strong evidence of association or evidence of a dose-response gradient.

Grading the ICC values based on measures of error makes it a necessity to consider the setting in which the measurement takes place. A 7.54% MAPE should be considered too high when assessing muscle thickness/cross-sectional area via ultrasound for pre-post-comparisons in short-lasting training interventions but could be negligible, e.g., when measuring the maximal strength in the squat in a one-year strength training study in previously untrained subjects where much higher effects are to be expected. Potentially, this could contribute to researchers critically questioning and appraising reliability classifications and their own work instead of unreflectively following the current conventions. Additionally, in turn, whether a measurement error is high or low might be relativized by the magnitude of the effect. Therefore, the LoAs in Bland-Altman analyses should be defined prior to an investigation when determining a tolerable range. When referring back to their original application to evaluate the agreement between blood pressure devices, Bland & Altman [24] performed exactly this procedure. In regard to reliability, Wright & Royston [28] defined the LoAs as the reference interval for test-retest differences expected for 95% of individuals. Thus, it can be considered the range most of the measurement errors will fall into when repeating the testing procedure under equal conditions in the same population [27]. Consequently, the evaluated LoA span can be used to check if testing was performed under suitable conditions meaning the error did not surpass the pre-defined ranges. Currently, it seems that these parameters are regularly determined without any consequence for the interpretation of the following results.

#### Limitations

This study's operator (LHL) acquired and rated all ultrasound images with utmost care. However, it cannot be precluded that investigator-dependent errors occurred, which might be present in any ultrasound investigation. Indeed, this underlines the relevance of determining

the random error, as investigator-related scattering would also contribute to this kind of error. Additionally, Bates, Dufek & Davis [34] and Dufek, Bates & Davis [35] stressed the role of the sample size for reliability values as well as a lack of generalizability when a testing procedure (in contrast to the reliability analysis of a device) is evaluated. Therefore, the results of this study are not transferable to other studies.

However, the results presented in this study underline the importance of not focusing solely on systematic and relative measurement errors, but rather adopting a more careful and balanced repeatability analysis for different measurements to realistically interpret the study results.

Reliability includes a broad range of indices, including intra- and inter-day repeatability (same conditions, same investigators, different time point), reproducibility (same conditions for the procedure, but different laboratories, investigators etc.), inter-investigator reliability/objectivity (almost the same time point, but different assessors or investigators). Also, validity analyses mostly use the same statistical approaches, comparing values from different measurement systems (e.g., ultrasound vs. magnetic resonance imaging being the gold standard). Given all of these different criteria, an uncertainty regarding the real muscle thickness arises that depends on the magnitude of the calculated value. In this study, the exclusive focus was placed on repeatability, which is just one potential error source, neglecting all other sources. Another origin of variance which might be expressed as secondary variance can be determined between different raters/investigators. A combined investigation approach with multiple test sessions for which data are collected from at least two different investigators was provided by Carstensen [36] and Carstensen et al. [27]. This approach should be applied to follow-up studies to account for further measurement error explorations and with that lead to improvements for future standardization of ultrasound investigations. Unfortunately, these more complex approaches were not suitable in this study, as our data were generated by just one investigator. Additionally, while the 95% confidence bands for the LoA are preferably derived via the exact method [37], we used the approximate approach for a better comparison across scenarios.

Another limitation in this paper stems from the use of horizontal LoAs in the Bland-Altman plot. Heteroscedasticity of data implies that deviations increase when measurement values increase which must be assumed for most sports science and medicine-related parameters [23] and can also be seen in this data collection (see the proportional bias line in Fig. 3). Thus, ideally, the LoAs should adapt to this trend shift and not be completely horizontal (see [38]). However, since this is commonly not done in current sports science and medicine research, this paper also used the simplified, horizontal

LoAs. This was *inter alia* done to improve the comparability to other studies' results as the focus of this study was to illustrate the shortcomings of current quantifiable parameters in reliability reporting.

The limitations mentioned above should be understood as an outlook and call for future original research to incorporate the latest statistical methods to improve reliability reporting.

## Conclusions

Researchers and clinicians should pay closer attention to random errors when using and referring to pre-post measurement changes using ultrasound-based data collection. The interpretations and derived recommendations should consider the random and systematic measurement error to provide a more careful and reliable statement. Even after accounting for the repeatability measurement source, there is no common classification that relates the different sources of the error to the expected or measured pre-post change, e.g. the magnitude of downscaling for the reported effect sizes or classification of the uncertainty arising from these error sources.

## Abbreviations

CV	Coefficient of variation
CV <sub>TE</sub>	Coefficient of variation of the typical error
ICC	Intraclass correlation coefficient
GL	Lateral head of the gastrocnemius
GM	Medial head of the gastrocnemius
LoA	Limits of Agreement
M	Mean
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MDC	Minimal detectable change
SD	Standard deviation
SEM	Standard error of measurement
VL	Vastus lateralis
TE	Typical error

## Acknowledgements

We thank the Medical School Hamburg for providing the ultrasound device and Daniel Jochum for his help with Fig. 1.

## Author Contributions

LHL and KW developed the idea for the study. LHL carried out the experiment and provided the first draft. The statistical analysis was performed by LHL and KW. The manuscript was discussed and revised by MH and SS. All authors contributed to the manuscript and agreed to the final version.

## Funding

Open Access funding enabled and organized by Projekt DEAL. No funding was received for conducting this study.

Open Access funding enabled and organized by Projekt DEAL.

## Data Availability

Data can be provided by the corresponding author on reasonable request.

## Declarations

### Ethics approval and consent to participate

All participants provided written informed consent for participation in the study which was conducted in accordance with the Declaration of Helsinki and approved by the Oldenburg Medical Ethics Committee (2021-089).

### Consent for publication

Not applicable.

### Competing Interests

Lars Hubertus Lohmann, Martin Hillebrecht, Stephan Schiemann and Konstantin Warneke declare that they have no competing interests.

### Author details

<sup>1</sup>University Sport Center, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

<sup>2</sup>Department of Human Movement Science and Exercise Physiology, Institute of Sport Science, Friedrich Schiller University, Jena, Germany

<sup>3</sup>Institute of Exercise, Sport and Health, Leuphana University, Lüneburg, Germany

<sup>4</sup>Institute of Sport Sciences, University of Klagenfurt, Klagenfurt am Wörthersee, Austria

Received: 5 January 2024 / Accepted: 22 July 2024

Published online: 15 August 2024

### References

1. Wada T, Tanishima S, Kitsuda Y, Osaki M, Nagashima H, Hagino H. Preoperative low muscle mass is a predictor of falls within 12 months of surgery in patients with lumbar spinal stenosis. *BMC Geriatr*. 2020;20:1–8.
2. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age Ageing*. 2019;48:16–31.
3. Schoenfeld BJ, Grgic J, Ogborn D, Krieger JW. Strength and hypertrophy adaptations between low- vs. high-load resistance training: a systematic review and meta-analysis. *J Strength Cond Res*. 2017;31:3508–23.
4. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. *J Biopharm Stat*. 2007;17:529–69.
5. Franchi MV, Raiteri BJ, Longo S, Sinha S, Narici MV, Csapo R. Muscle architecture assessment: strengths, shortcomings and new frontiers of in vivo imaging techniques. *Ultrasound Med Biol*. 2018;44:2492–504.
6. Betz TM, Wehrstein M, Preisner F, Bendszus M, Friedmann-Bette B. Reliability and validity of a standardized ultrasound examination protocol to quantify vastus lateralis muscle. *J Rehabil Med*. 2021;53.
7. Nijholt W, Scafoglieri A, Jager-Wittenaar H, Hobbelen JSM, van der Schans CP. The reliability and validity of ultrasound to quantify muscles in older adults: a systematic review. *J Cachexia Sarcopenia Muscle*. 2017;8:702–12.
8. Warneke K, Keiner M, Lohmann LH, Brinkmann A, Hein A, Schiemann S et al. Critical evaluation of commonly used methods to determine the concordance between sonography and magnetic resonance imaging: a comparative study. *Front Imaging*. 2022;1.
9. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to Sports Medicine. *Sport Med*. 1998;26:217–38.
10. Lamb K. Test-retest reliability in quantitative physical education research: a commentary. *Eur Phys Educ Rev*. 1998;4:145–52.
11. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59:1033–9.
12. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48:661–71.
13. Hopkins WG. Measures of reliability in sports medicine and science. *Sport Med*. 2000;30:1–15.
14. Warneke K, Brinkmann A, Hillebrecht M, Schiemann S. Influence of long-lasting static stretching on maximal strength, muscle thickness and flexibility. *Front Physiol*. 2022;13:1–13.
15. Cleary CJ, Nabavizadeh O, Young KL, Herda AA. Skeletal muscle analysis of panoramic ultrasound is reliable across multiple raters. *PLoS ONE*. 2022;17:1–12.
16. Warneke K, Keiner M, Wohlant T, Lohmann LH, Schmitt T, Hillebrecht M, et al. Influence of long-lasting static stretching interventions on functional and morphological parameters in the plantar flexors: a randomized controlled trial. *Head print: J Strength Cond Res*; 2023.
17. Warneke K, Wirth K, Keiner M, Lohmann LH, Hillebrecht M, Brinkmann A, et al. Comparison of the effects of long-lasting static stretching and hypertrophy training on maximal strength, muscle thickness and flexibility in the plantar flexors. *Eur J Appl Physiol*. 2023;123:1773–87.
18. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–63.
19. Bramah C, Preece SJ, Gill N, Herrington L. The between-day repeatability, standard error of measurement and minimal detectable change for discrete kinematic parameters during treadmill running. *Gait Posture*. 2021;85:211–6.
20. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Med Educ*. 2010;10:1–9.
21. Grönkvist R, Vixner L, Ång B, Grimby-Ekman A. Measurement error, minimal detectable change, and minimal clinically important difference of the short form-36 health survey, hospital anxiety and depression scale, and pain numeric rating scale in patients with chronic pain. *J Pain*. 2024;104559.
22. Högelin ER, Thulin K, von Walden F, Fornander L, Michno P, Alkner B. Reliability and validity of an Ultrasound-based protocol for measurement of quadriceps muscle thickness in children. *Front Physiol*. 2022;13:1–8.
23. Nevill AM, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med*. 1997;31:314–8.
24. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–10.
25. Ferreira JA, Zwiderman AH. On the Benjamini-Hochberg method. *Ann Stat*. 2006;34:1827–49.
26. French D, Torres Ronda L. Preface. *NSCA's essentials Sport Sci*. 1st ed. Human Kinetics; 2021. pp. XVIII–XIX.
27. Carstensen B, Simpson J, Gurrin LC. Statistical models for assessing agreement in method comparison studies with replicate measurements. *Int J Biostat*. 2008;4.
28. Wright EM, Royston P. Calculating reference intervals for laboratory measurements. *Stat Methods Med Res*. 1999;8:93–112.
29. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. New York: Lawrence Erlbaum Associates; 1988.
30. Hebert JJ, Koppenhaver SL, Parent EC, Fritz JM. A systematic review of the reliability of rehabilitative ultrasound imaging for the quantitative assessment of the abdominal and lumbar trunk muscles. *Spine (Phila Pa 1976)*. 2009;34:848–56.
31. Liljequist D, Elfving B, Roaldsen KS. Intraclass correlation – a discussion and demonstration of basic features. *PLoS ONE*. 2019.
32. Goodpaster BH, Park SW, Harris TB, Kritchevsky SB, Nevitt M, Schwartz AV, et al. The loss of skeletal muscle strength, mass, and quality in older adults: the Health, Aging and Body Composition Study. *Journals Gerontol - Ser Biol Sci Med Sci*. 2006;61:1059–64.
33. GRADE working group. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328.
34. Bates BT, Dufek JS, Davis HP. The effect of trial size on statistical power. *Med Sci Sports Exerc*. 1992;24:1059–68.
35. Dufek JS, Bates BT, Davis HP. The effect of trial size and variability on statistical power. *Med Sci Sports Exerc*. 1995;27:288–95.
36. Carstensen B. Comparing and predicting between several methods of measurement. *Biostatistics*. 2004;5:399–413.
37. Carkeet A. Exact parametric confidence intervals for bland-altman limits of agreement. *Optom Vis Sci*. 2015;92:e71–80.
38. Carstensen B. Introduction to the MethComp package. 2012. <https://bendix-carstensen.com/MethComp/introMethComp.pdf>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.