MEASUREMENT IN PHYSICAL EDUCATION AND EXERCISE SCIENCE
2025, VOL. 29, NO. 4, 422–433
https://doi.org/10.1080/1091367X.2025.2494998

Routledge
Taylor & Francis Group

The Impact of Sample Size on Reliability Metrics Stability in Isokinetic Strength Assessments: Does Size Matter?

Konstantin Warneke[a,b,c], Michael Keiner[d], Sebastian Wallot[a], Stanislav D. Siegel[b], Christian Günther[e], Klaus Wirth[f], and Sebastian Puschkasch-Möck[e]

[a]Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany; [b]Institute of Human Movement Science, Sport and Health, University of Graz, Graz, Austria; [c]Institute of Human Movement and Exercise Physiology, University of Jena, Jena, Germany; [d]Department of Sport Science, German University of Health and Sport, Ismaning, Germany; [e]Department of Exercise Science, Olympic Training and Testing Center of Hessen, Frankfurt am Main, Germany; [f]Department of Exercise Science, University of applied Sciences Wiener Neustadt, Wiener Neustadt, Austria

ABSTRACT

The ability to reliably capture performance parameters must be considered as crucially important to produce valid study results. The ICC and the inclusion of the calculation of the standard error of measurement and the minimal detectable change became the most common way to justify subsequent testing procedures to be reliable. However, early studies around the new millennium identified weaknesses of the ICC and proposed the implementation of more elaborate procedures, including the quantification of the systematic bias and the quantification of the random error via the mean absolute error or mean absolute percentage error. According to the law of large number and earlier research indicating that relative indices such as correlation coefficients necessitate a minimum sample size to stabilize, it was hypothesized that reliability indices follow an optimal sample size trend. In accordance with previous studies in correlation coefficients, this study highlights the importance of including high numbers of participants to receive stable reliability measures. The random error was not significantly affected by increased samples while providing important information about the performed standardization success in the testing, the study also underlines the relevance of reporting not only ICC-based reliability statistics but also the quantification of random errors.

KEYWORDS

reliability; repeatability; intraclass correlation coefficient; law of large numbers; measurement errors

Introduction

Testing an athlete's strength capacity is an important component of almost every performance test in elite sports practice and rehabilitation to track the effectivity of the performed training (Tanner & Gore, 2012; Warneke et al., 2023). One of the several methods to measure neuromuscular force output is isokinetic dynamometry. This method is frequently used in clinical and rehabilitation settings (Gleeson & Mercer, 1996) as well as high-performance sports (Blazquez et al., 2013; Cometti et al., 2001). While strength and conditioning research majorly focus on improving and optimizing training routines (French & Torres Ronda, 2022), the lowest border that must be crossed to evaluate training effects is sufficient reliability of the testing routine used in maximum strength evaluation (Atkinson & Nevill, 1998, 2000; Barnhart et al., 2007; Warneke et al., 2025). To provide practitioners with helpful information whether sufficient reliability, validity and objectivity of testing are fulfilled, the reporting of reliability coefficients was implemented. The most common ones are the intraclass correlation coefficient (ICC) for consistency or agreement (Koo & Li, 2016) and the variability coefficient (CV) (Chen et al., 2020; Hauser et al., 2012), while some authors supplement their statistical baseline check with the quantification of the standard error of measurement (SEM). Additionally, to account for the precision of the testing protocol and to attribute potential effects to the intervention by surpassing measurement errors, the minimal detectable change (MDC) as well as the smallest worthwhile change (SWC) (Haugen et al., 2019; Willigenburg & Poolman, 2023) were calculated in literature (Impellizzeri et al., 2008; Lienhard et al., 2013; Martins et al., 2017; Wollin et al., 2016).

However, this procedure was frequently criticized in the past. Atkinson and Nevill (1998); Barnhart et al. (2007); Hopkins (2000) and Warneke et al. (2025) discussed limitations of these relative statistical measures

CONTACT Konstantin Warneke Konstantin.Warneke@icloud.com Institute of Psychology, Leuphana University Lüneburg, Universitätsallee 1, Lüneburg 21335, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/1091367X.2025.2494998

to judge a measurement device as reliable. A major concern relates to the differentiation between systematic error and random error, which are crucial factors in measurement reliability. Systematic error refers to consistent and predictable biases in measurements, often introduced by external influences such as calibration errors, tester bias, or environmental conditions. Random error, on the other hand, represents unpredictable fluctuations in measurements due to biological variability, momentary inconsistencies in the testing process, or equipment sensitivity. These errors create scatter around the true value and reduce measurement precision (Hopkins, 2000). To account for such error facets, several authors strongly suggested the implementation of Bland Altmann analyses (BA) in addition to other reliability statistics to account for potential systematic errors and to state the limits of agreement (LoA) (Atkinson & Nevill, 1998, 2000; Lamb, 1998; Nevill & Atkinson, 1997). The LoAs can be considered a qualitative illustration of the random scattering around the systematic bias. The BA analysis is performed by plotting the differences of value 1 and value 2 in dependency of the respective mean of measurements (Doğan, 2018), while the LoAs cover 95% of the values. To supplement this subjective inspection of the graphical measurement error illustration with a quantification of the actual measurement error, C. Willmott and Matsuura (2005, 2006) and Kim and Kim (2016) suggested the calculation of the mean absolute error (MAE) and mean absolute percentage error (MAPE) (C. Willmott & Matsuura, 2005; C. J. Willmott & Matsuura, 2006) between trial 1 and trial 2 as a means to account for individual measurement errors (Warneke et al., 2025).

Especially for correlation-based statistics (which is also true for ICCs), earlier research indicated that validity of these coefficients is biased by instability when applying in small sample sizes (Schönbrodt & Perugini, 2013). This analysis, however, focused on correlation coefficients, but was never performed for reliability ICCs. A stable determination of reliability, with its facets of precision and accuracy, can therefore serve as a benchmark to reasonably justify the sample size, as possibly larger sample sizes are required to reach a stable measurement reliability. This assumption stems from the law of large numbers (LLN) (Yao & Gao, 2016) stating that the average of results obtained from a large number of observations leads to the random scattering of individual courses being neglectable for the generalizability of study results. Accordingly, the random error could be inversely related with the sample size, as the impact of individual scattering would decrease. In this case, the ICC would increase in its validity as the sample size has crucial impact on the variance, on which ICC calculations are based, making the quantification of the MAPE unnecessary.

Consequently, this study was conducted to investigate the impact of the sample size on the ICC, the MAPE and the ICC/MAPE ratio, which was hypothesized to increase with an increasing sample size if the MAPE would drop automatically. To better account for both relative reliability and absolute measurement error, the ICC/MAPE ratio could provide a reasonable combined metric allowing for a more comprehensive evaluation of test consistency and accuracy.

Material and methods

To account for the influence of different sample sizes on the reliability, in this case, on isokinetic strength testing sessions for the lower limb and trunk, the relative reliability supplemented by absolute reliability indices and accounting for the systematic and random measurement error was calculated. Furthermore, in different sample sizes and test protocols, the influence was investigated by adding $n = 5$ randomly picked data points out of the overall sample to evaluate the courses of the reliability metrics (ICC, MAPE, ICC/MAPE ratio) to allow conclusions whether there is a crucial sample size to stabilize these metrics as shown by Schönbrodt and Perugini (2013) for the correlation coefficient.

Subjects

Isokinetic dynamometer testing can be considered a frequently applied performance test in sports medicine and exercise science. However, it requires habituation of participants to testing conditions, as unfamiliar performance tests can be considered invalid to assess maximal strength, per se (Warneke et al., 2023). Therefore, to ensure validity of using isokinetic dynamometer tests and ensure minimal habituation effects, overall, 430 high-performance athletes of various sports (national league competitive athletes) (318 males, 112 females, mean age = 21.1 ± 5.4 years, mean height = 179.7 ± 8.6 cm, mean weight = 75.3 ± 11.3 kg) participated in the performed strength testing sessions. The athletes originated from sports such as track and field, table tennis, field hockey, bobsleigh, karate and volleyball. Since these sports have different needs in their regular performance testing, not all athletes performed all the applied tests to avoid different habituation states between the athletes, as those could affect the sample size/reliability relationship, while simultaneously reflecting realistic and

frequently applied study designs with heterogeneous sample sizes.

Each subject was informed of the experimental risks involved with the research and signed informed written consent. The experimental protocols and procedures were approved by the Internal Review Board (GZ.39/225/63ex2023/24) and conformed to the standards set by the Declaration of Helsinki.

Isokinetic peak torque measurements

The isokinetic measurements were performed for reciprocal flexion and extension of the knee, the hip, the ankle and the trunk using the ISOMED2000 isokinetic dynamometer (D&R Ferstl GmbH, Hemau, Germany) with a measuring rate of 200 hz. All the tests were performed in two sets of three concentric repetitions with an inter-set rest of 1 min. Trunk flexion and extension were measured at 60°/s ($n = 336$), while knee ($n = 430$ for 60°/s, $n = 324$ for 180°/s) and hip flexion and extension were measured at 60°/s ($n = 166$) and 180°/s ($n = 156$) and ankle flexion and extension were measured at 30°/s ($n = 276$) and 120°/s ($n = 114$), respectively. The movement velocities were used as they represent commonly applied velocities in strength diagnostics (Blazquez et al., 2013; Koutras et al., 2016; Menzel et al., 2013; Möck & Wirth, 2024; Möck et al., 2023; Roth et al., 2017).

The measurements at the knee joint were performed in a seated position on the dynamometer with the back rest set at an angle of 75° (0° referring to a full horizontal decline), while the measurements at the hip and ankle joints were performed in a supine position on the dynamometer. The participants were fixed with straps and cushions according to the manufacturer's recommendations (Figure 1) and instructed to contract as hard as possible throughout the full range of motion and strong verbal encouragement during the tests was provided. The range of motion was 90–170° of the respective angle for the knee and hip flexions and extensions (180° referring to full extension) and 70–125° at the ankle joint (90° referring to the neutral position). Peak torque was calculated for the strongest repetition of each set using LabView 2018 (National Instruments, Austin, TX, USA). The order of the tests as well as the order of the legs was randomized, and the lower velocity was tested first according to the guidelines proposed by Perrin (1993).

Statistical procedure

Data analysis was performed with JASP (Version 0.18.3). Descriptive statistics are stated as mean (M) and standard deviation (SD) as well as the number (n) of participants. To test reliability, the ICC (3,1) including the 95% confidence intervals (CI) was calculated, using the formula described by (Koo & Li, 2016).

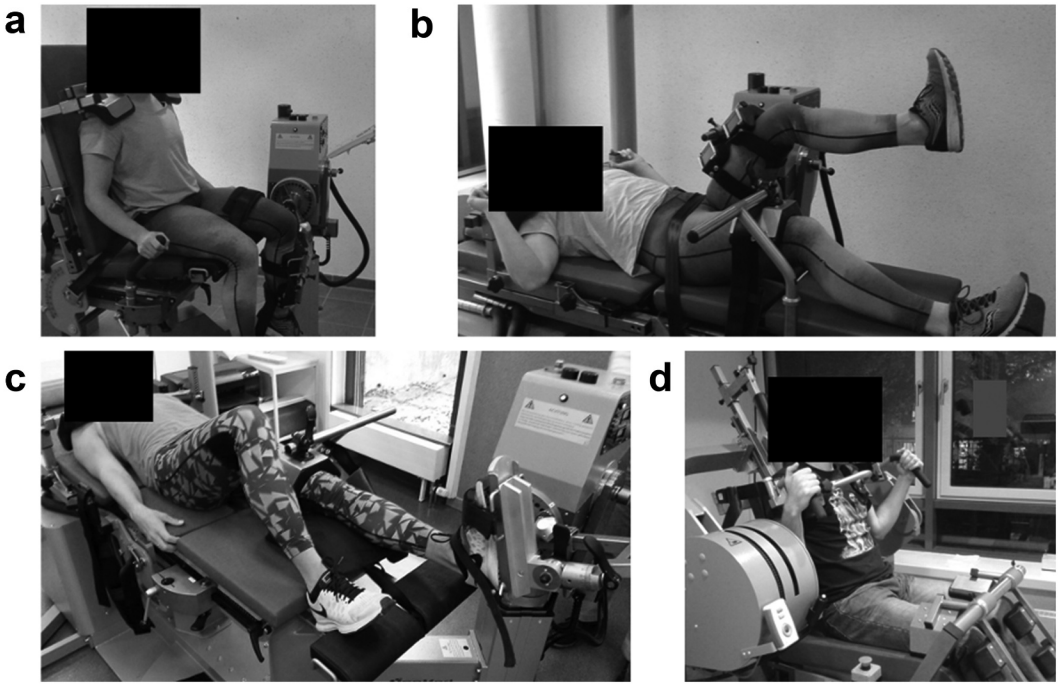

Figure 1. Positioning and fixation of the participants on the isokinetic dynamometer. (a) Knee measurement, (b) hip measurement, (c) ankle measurement, (d) trunk measurement.

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

with ICC = intraclass correlation coefficient, MS_C = mean square for columns, MS_E = mean square for error, MS_R = mean square for rows, n = number of subjects,

ICC thresholds were adopted and classified as follows: < 0.50 = poor, 0.50–0.75 = moderate, 0.75–0.90 = good, and > 0.90 = excellent (Koo & Li, 2016). As the SEM (Tighe et al., 2010) and the MDC (Seamon et al., 2022) are considered standard reliability parameters as well, we calculated those using the following formulas:

$$SEM = SD * \sqrt{1 - ICC}$$

with SEM = standard error of measurement, SD = standard deviation of the mean difference between trial 1 and 2, ICC = intraclass correlation coefficient

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

with MDC = minimal detectable change, SEM = standard error of measurement

Additionally, BA agreement analyses were plotted, and the mean difference (systematic error) was calculated using the dependent two sample t-test (Atkinson & Nevill, 1998) with the respective LoAs provided to account for random scattering (Bland & Altman, 1999; Bland & Altmann, 1986; Doğan, 2018). Additionally, the mean absolute error (MAE) (C. Willmott & Matsuura, 2005; C. J. Willmott & Matsuura, 2006) as well as the mean absolute percentage error (MAPE) (Kim & Kim, 2016) were calculated using the following formulas:

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

with n = number of data points, i = index for each (paired) data point, x_i = i-th data point in variable x, y_i = i-th data point in variable y.

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

with n = number of data points, i = index for each (paired) data point, x_i = i-th data point in variable x,y_i = i-th data point in variable y.

To assess the influence of large sample sizes on the ICC and a potential decrease in MAPE with an increasing sample size, the MAPE was plotted as a function of the ICC. Additionally, the MAPE/ICC ratio was used to evaluate potential dependencies regarding sample size. Additionally, Kendall's tau (due to lack of assumption of normal distribution) was calculated to investigate the relationship between the ICC and MAPE. To further validate the findings, individual data points were randomly drawn from the total sample of each respective test, and MAPE and ICC were plotted against the sample size. Starting with $n = 5$ per subsample, the number of participants was progressively increased by $n = 1$ to plot the ICC-sample size, MAPE-sample size and ICC/MAPE-sample size ratios for each individual measurement.

Results

Ranging between 0.94 and 0.98, all ICCs are classified as excellent (≥0.9) (Koo & Li, 2016), with corresponding SEM and MDC ranging between 0.25–4.20 Nm and 0.68–11.64 Nm, respectively. With MAE from 1.72 to 22.05 (4.1–9.9%), the MAE was higher than the MDC in all cases. Furthermore, all systematic bias calculations showed a significant systematic increase from the test to the retest (p < .001–0.008). Applying the Bonferroni-Holm Correction decreased the alpha level threshold to p = .004, causing only the plantar flexor strength systematic error to become non-significant. The results are shown in Table 1, while Figure 2 graphically illustrates the systematic bias and the random error via BA plots, including the LoAs for the knee joint testing (knee flexion and extension with 60°/s and 180°/s). The remaining BA plots can be reviewed in the Supplemental material (A–C).

Regarding the initial hypothesis that the validity of the ICC would increase with an increased sample size (as evident by a simultaneous decrease in the MAPE), the relationship between the ICC and the MAPE is plotted in Figure 3(a) as a function of sample size. Since there are currently no standard procedures to evaluate this relationship, the MAPE/ICC ratio was calculated and plotted as a function of the sample size (Figure 3(b)) to qualitatively evaluate whether there is a systematic association.

In the population investigated in this study, the ICC remains stable even with an increasing sample size (r = 0.24, p = .27), whereas the MAPE exhibits a moderate dependence (r = −0.46, p = .03). In contrast to the assumed inverse relationship between the ICC and MAPE, Figure 3(a) shows that high ICCs were unsystematically accompanied by high and small MAPEs (e.g. Hip Flex 60), while lower ICCs showed moderate MAPEs (e.g. dorsiflex 30) as well. When plotting the ICC/MAPE ratio as a function of sample size, however, visual inspection suggests a tendency of the ICC/MAPE-ratio to increase with sample size (increasing ICC is accompanied by a decreasing MAPE, Figure 4). This relationship in these 14 data points can be described with Kendall's tau of r = 0.44 (p = .04), thus explaining 19.4% of variance.

Table 1. Results of the reliability and error calculations.

Parameter	N	M ± SD Trial 1 in Nm	M ± SD Trial 2 in Nm	ICC (95% CI)	CV (%)	SEM in Nm	MDC in Nm	LoA (l-u) in Nm	Syst bias (p-value)	MAE in Nm	MAPE (in %)	MPE (in %)
Hip Flex 60	165	156.61 ± 48.81	162.67 ± 50.56	0.96 (0.94–0.97)	5.06	2.89	8.02	− 34.41–22.29	− 6.06 (<.001)*	11.23	7.5	47.2
Hip Ext 60	165	354.54 ± 89.15	369.32 ± 92.94	0.97 (0.95–0.97)	4.45	4.20	11.64	− 62.29–32.73	− 14.78 (<.001)*	22.05	6.6	47.7
Hip Flex 180	156	129.45 ± 46.48	134.18 ± 48.09	0.95 (0.94–0.96)	6.11	3.29	9.11	− 33.53–24.07	− 4.73 (<.001)*	10.89	9.9	62.2
Hip Ext 180	156	316.15 ± 86.73	322.74 ± 89.05	0.98 (0.97–0.98)	3.52	2.81	7.78	− 45.50–32.30	− 6.57 (<.001)*	14.49	5.1	41.2
Knee Flex 60	430	135.74 ± 32.12	139.23 ± 32.46	0.97 (0.97–0.98)	3.51	1.35	3.75	− 18.79–11.82	− 3.5 (<.001)*	6.56	6.0	30.9
Knee Ext 60	430	199.67 ± 50.23	202.29 ± 48.85	0.96 (0.95–0.97)	4.15	2.87	7.96	− 30.75–25.52	− 2.6 (<.001)*	5.48	5.1	62.0
Knee Flex 180	324	117.05 ± 28.24	120.19 ± 28.71	0.98 (0.97–0.98)	3.26	0.87	2.41	− 15.21–8.93	− 3.14 (<.001)*	5.30	4.7	22.0
Knee Ext 180	324	150.24 ± 33.94	152.85 ± 34.18	0.98 (0.97–0.98)	2.88	1.03	6.69	− 16.91–11.68	− 2.61 (<.001)*	6.0	4.1	27.0
Plantar Flex 30	274	145.51 ± 40.96	149.61 ± 42.16	0.97 (0.96–0.98)	4.49	1.76	4.89	− 24.04–15.86	− 6.66 (<.001)*	8.70	6.4	31.0
Dorsi Flex 30	274	31.14 ± 7.7	31.60 ± 7.76	0.95 (0.94–0.96)	4.61	0.54	1.50	− 5.21–4.28	− 3.18 (.002)*	1.95	6.6	35.3
Plantar Flex 120	114	109.42 ± 32.76	112.27 ± 32.18	0.94 (0.92–0.96)	6.16	3.16	8.75	− 24.87–19.16	− 2.85 (.008)	9.43	8.9	36.0
Dorsi Flex 120	114	21.26 ± 7.33	22.42 ± 7.16	0.97 (0.96–0.98)	6.28	0.25	0.68	− 4.65–2.32	− 1.16 (<.001)*	1.72	9.5	67.0
Back Flex 60	336	137 ± 42.52	141.50 ± 44.81	0.97 (0.97–0.98)	3.99	1.81	5.01	− 24.61–16.29	− 4.16 (<.001)*	8.05	5.8	29.1
Back Ext 60	336	278.73 ± 100.94	291.45 ± 104 ± 07	0.98 (0.97–0.98)	5.12	3.06	13.89	− 55.10–29.66	− 12.72 (<.001)*	19.25	7.6	44.5

Back Flex = back flexion testing, Back Ext = back extension testing, Dorsi Flex = dorsiflexion testing, Hip Flex = hip flexion testing, hip ext = hip extension testing, Knee Ext = knee extension testing, Knee Flex = knee flexion testing, Plantar Flex = plantar flexion testing, 30 = testing velocity of 30°/s, 60 = testing velocity of 60°/s, 120 = testing velocity of 120°/s, 180 = testing velocity of 180°/s.
*=significant considering adjusted threshold via Bonferroni-correction.

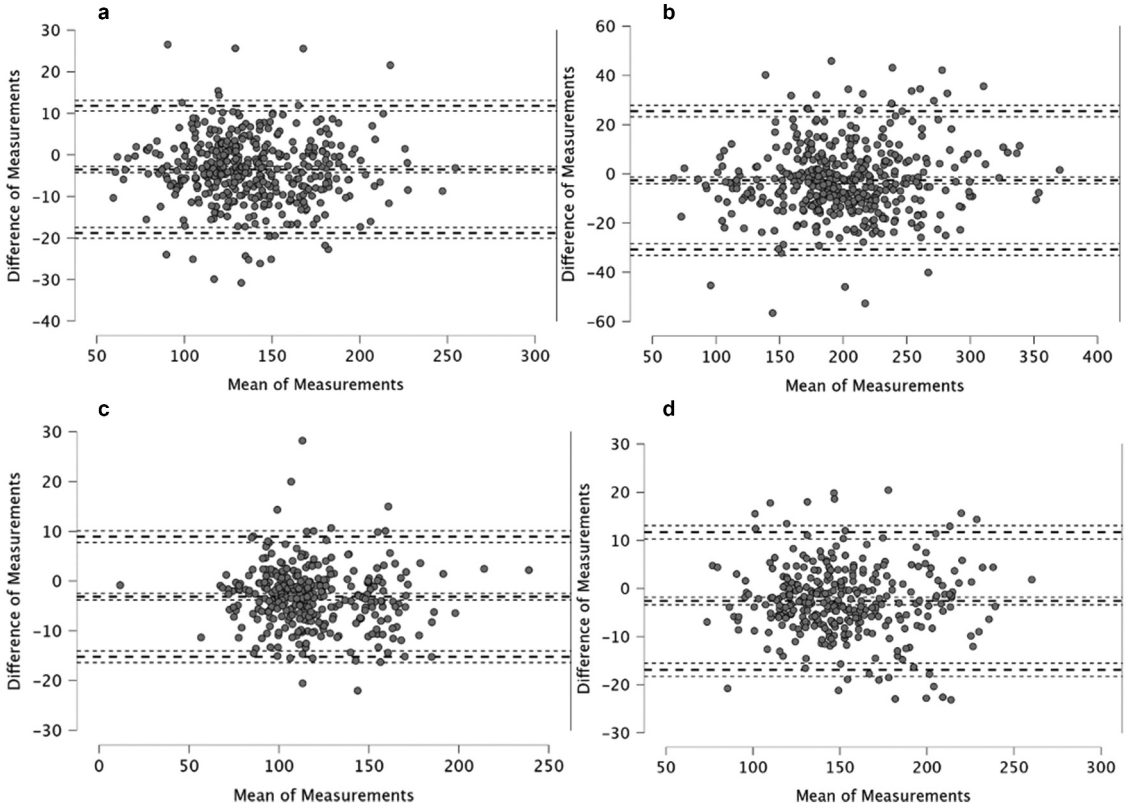

Figure 2. BA plots for knee flexion and extension (60°/s (a,b) and 120°/s (c,d)), with LoAs ranging between −18.8–11.8 Nm (a), −30.8–25.5 Nm (b) around a mean of measurements of 135.7 Nm and 199.7 Nm, respectively. In the higher velocities (180°/s), BA analyses show means of 117.1 Nm and 150.2 Nm with LoAs of −15.2–8.9 Nm (c) and −16.9–11.7 Nm (d), respectively. Plots stems from $n = 430$ and $n = 324$.

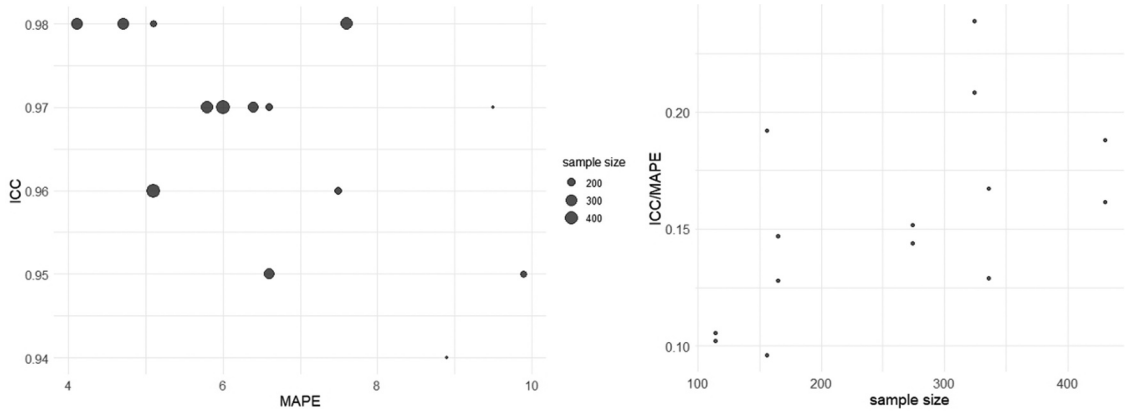

Figure 3. ICC in dependency of the MAPE (left) as well as the ICC/MAPE ratio in dependency on the sample size (right).

For these calculations, it is not possible to exclude the role of the different movements/test joint segments and the problem of confounds that might be introduced by the different types of data on the analysis. Therefore, Figure 5 graphically illustrates the relationship of the MAPE and ICC in dependency on the sample size for each movement. Starting with $n = 5$, the sample size was progressively increased. The courses that show the progression of the ICC (A) and MAPE (B) (y-axis) and the sample size (x-axis) for each testing condition are illustrated in Figure 5

In Figure 5, the following color codes were used: hip180 = red, knee180 = magenta, knee60 = cyan, back60 = yellow, plantar flexion30 = blue, plantar flexioin120 = green, with continous lines representing the extension data and dashed lines referring to the flexion data. As can be seen, different patterns emerge for the ICC and the MAPE within each data set: We can see that

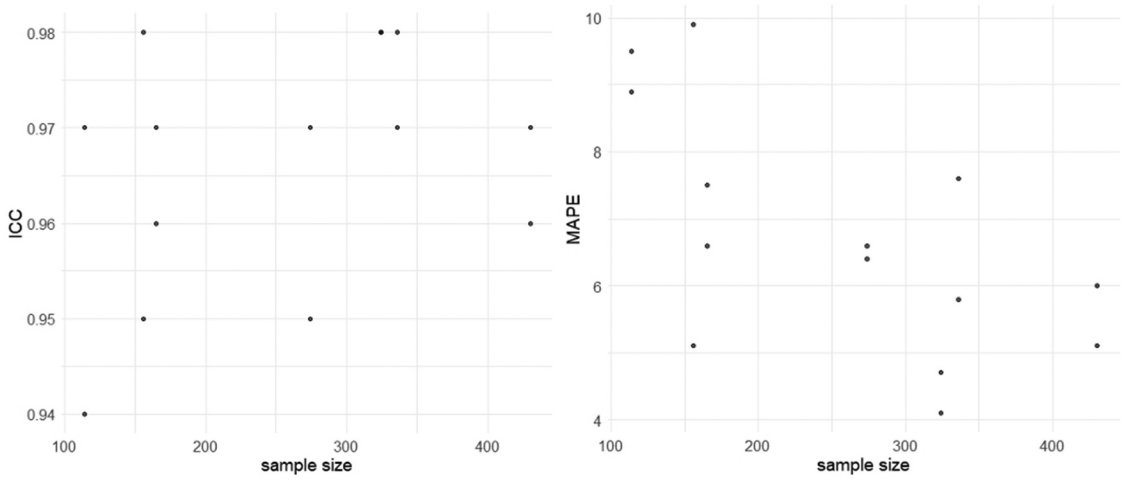

Figure 4. ICC (left) and the MAPE (b) in dependency on the sample size (right).

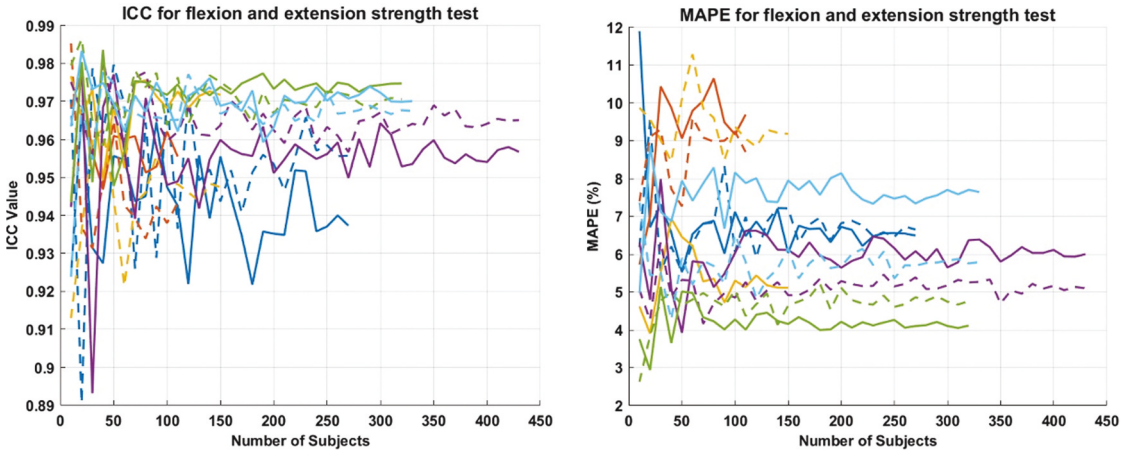

Figure 5. Progression of the ICC and MAPE for the intraday reliability in dependency on the number of included subjects for each of the tested movements.

both measures fluctuate and stabilize only at higher sample sizes, in line with previous investigations of the stability of correlation coefficients as a function of sample size (Schönbrodt & Perugini, 2013) suggesting that a substantial sample size is needed for stable coefficients. For the ICC/MAPE ratio, a pattern of stabilization with increasing sample size can be observed as well without a systematic in- or decrease in the relationship between these two measures (Figure 6). Compared to the between-sample data presented in Figures 3 and 4, this underlines the assumption that different types of training/testing data are subject to measurement-specific factors. Consequently, relations between ICC and MAPE for one type of test situation cannot be generalized and transferred to other situations. Otherwise, Figures 2–5 should have shown converging pattern.

Discussion

With this study, we aimed to investigate whether an increase in the sample size would affect the validity of the ICC in quantifying measurement errors according to the LLN. Here, validity refers to the extent to which ICC reliably quantifies measurement consistency without being overly influenced by sample size fluctuations. If ICC estimates vary significantly when adding only a few data points, this suggests instability and challenges its practical validity in small samples. This stabilization of reliability metrics could contribute to an improved sample size estimation, as it provides viable information on how precise and accurate an ability was assessed.

It was first hypothesized that an increase in the sample size would automatically result in a more stable

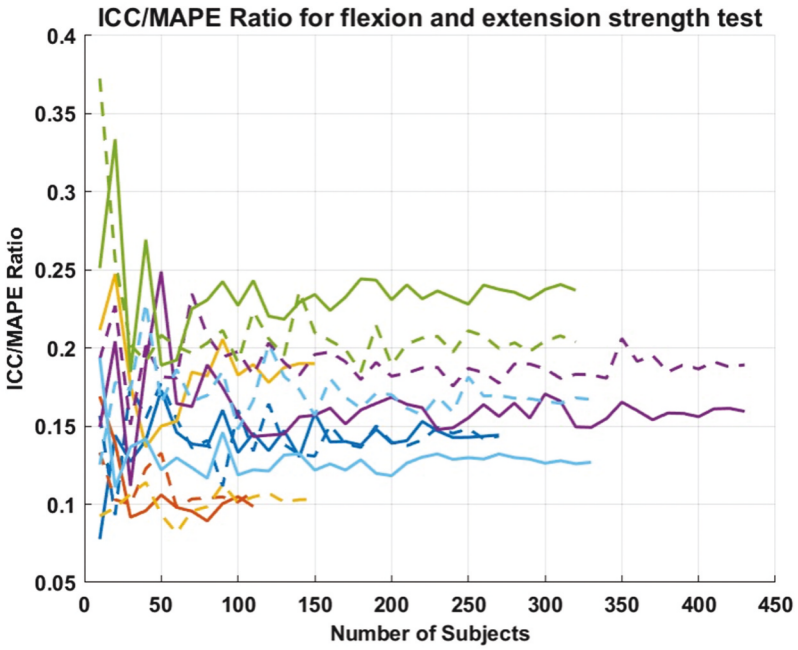

Figure 6. Progression of the ICC/MAPE ration for intraday reliability in dependency on the number of included subjects for each testing movement.

reliability determination as described for correlation statistics by Schönbrodt and Perugini (2013). Furthermore, in accordance with the LLN (Yao & Gao, 2016), we hypothesized that the relevance of quantifying the random measurement error would decrease, as increasing the sample size would narrow a measured value, decreasing the variance in the test–retest relationship.

While the first part of this paper indicated excellent relative reliability with ICC for all measurements ranging between 0.94 and 0.98, it must be noted that these were accompanied by non-negligible random (e.g., MAPE: 4.1–9.9 Nm, SEM: 0.25–4.20 Nm, 0.68–13.89 Nm) as well as significant systematic errors. Nevertheless, this paper provides two main statements regarding the ICC and random error validity. First, in accordance with the LLN, larger sample sizes narrow the confidence interval of the ICC, leading to more stable estimates. However, the absolute magnitude of random error (e.g., MAPE, SEM) did not change, only its variance decreased. This means that while ICC estimates become more consistent, the underlying random fluctuations in individual test scores persist and are not mitigated by a larger sample. As illustrated in Figure 5, the measurement error quantification via the MAPE as well as the relative reliability level was dependent on the nature of the test as well as the actual measurement error. Second, systematic errors remain independent of sample size and were detected across almost all testing conditions. This indicates that systematic biases, such as learning effects or tester-related inconsistencies, must be controlled through protocol standardization, rather than statistical adjustments. This demonstrates that ICC alone does not fully capture measurement reliability and that additional error quantifications (e.g., MAPE, BA analysis) are necessary. This was extensively discussed in a recent review article showing that even ICCs classified as excellent were accompanied by a substantial level of random errors (20% MAPE) (Warneke et al., 2025). Furthermore, outliers are a common challenge in reliability analyses, particularly when working with small sample sizes, where they can significantly impact ICC. In this study, we intentionally retained all data points, as variability is an inherent characteristic of sport and exercise science measurements. Our findings confirm that ICCs remained stable as sample size increased, suggesting that the chosen sample size was sufficient to mitigate potential biases introduced by outliers.

Influence of the sample size

In the first part, we investigated whether we could detect a systematic increase of the ICC or decrease of the MAPE, which was summarized by reviewing the ICC/MAPE ratio. Although, from visual inspection, a slight tendency was observable that the relationship improved with an increasing sample size, the level of significance only barely reached the level of significance ($p = .04$). However, since not only the sample size per test, but also the nature of the test itself differed, it was not clear

whether this progression could be attributed to the different sample sizes, or the differences and specificities of the performed tests. Therefore, we performed a separated statistical analysis and plotted the ICC and MAPE in dependency of the sample size for each strength test individually to standardize the results for testing specificity. The study results underline the key message associated with the LLN (Yao & Gao, 2016), that an increasing number of participants narrows down the actual value, while being also in accordance with previous results from Schönbrodt & Perugini (Schönbrodt & Perugini, 2013). In accordance with the authors who recommended the inclusion of a minimum sample size of $n = 250$ as a general suggestion, our results merely confirm these suggestions, as we see that the stepwise addition of single participants does not seem to affect ICC and MAPE at a number of participants approximating the one proposed. The lower the sample size, the larger the variations in the determination of the reliability (adding only few participants may cause an increase or decrease of the MAPE by several percent). To provide a merged reliability coefficient that included the random scattering as well as the relative reliability including the variance of the differences, the ICC/MAPE ratio was calculated. The rationale for this was that if the ICC increases its validity to account for reliability, as initially hypothesized, the random error should decrease with an increased sample size. Here, more consistent results were observed at slightly lower sample sizes of around 100–150 participants. The variation observed in smaller sample sizes persists, albeit in a more consistent manner with increasing sample size. Our second aim was to investigate if the MAPE decreases automatically when increasing the sample size. This would automatically result in an increase of the ICC/MAPE ratio per testing condition, while we would see a decrease in all testing conditions in the MAPE plot that illustrates the random error with respect to the sample size. However, increasing the sample size did not reduce the MAPE, but only its variance. In accordance with previous research that underlined the relevance of reporting systematic and random errors in addition to the ICC (Lohmann et al., 2024), an increase in the number of participants did not solve the problems of random scattering around the actual measurement error. Therefore, as the random error provides important information on how successful the authors standardized their measurement protocol, the isolated reporting of the ICC still neglects information, and this problem cannot be solved by increasing the sample size. For both, the ICC and the MAPE, a least sample size is necessary to receive a valid and stable quantification of reliability, however, a large sample size does not substitute for quantifying random measurement errors. Previous research, however, not only suggested the quantification of random errors and relative reliability indices, but also requested the implementation of systematic test–retest differences to rule out potential learning effects that would indicate invalid testing procedures in several cases. For instance, if the goal is to determine the ability of a participant to produce a high force output, it is crucial that the results are not biased by habituation and warm-up effects. Therefore, participants must be familiar with testing conditions (otherwise, we would probably measure learning effects instead of the actual force output), and testing must occur after a sufficient warm-up (otherwise, we do not measure the actual force but warm-up effects). Both cases would bias the validity of the measurement, if the aim was to explore force production capacity. To check for systematic effects (habituation/Warm-up), the test–retest relationship for systematic mean differences can be utilized (using a t-test (Atkinson & Nevill, 1998, 2000; Barnhart et al., 2007).

Systematic bias and the sample size

An excellent reliability as indicated by ICCs of ≥0.9 (Koo & Li, 2016) would inherently mean that subsequent testing of one and the same parameter results in the same value. Reliability quantification should therefore account for the deviation between the test and retest value, with a difference of zero between the test values indicating maximal reliability. While the random error, or "noise" (Hopkins, 2000), can be attributed to lack of standardization (e.g. in ultrasound testing due to applying unsystematically more or less probe pressure, while the daily performance level of an athlete might randomly affect the interday reliability in strength and performance tests), measurements can only be considered valid to reflect, for instance, the strength level, if participants are familiar with the measurement protocol. If learning effects occur, the used strength testing procedure must be considered invalid as not the force production ability, but learning effects were monitored. To account for such learning effects, literature suggests performing two-sampled t-tests to check for significant mean differences between the test–retest means (including, per definition, the sample size, mean difference and pooled standard deviation).

Initially, we hypothesized that an increase in the sample size would also positively affect the validity of the ICC by considering possible changes of the random error depending on the sample size. However, we observed that the bias of the ICC's validity due to the systematic error was more pronounced, while the

magnitude of random error was not significantly affected by an increased sample size. On the one hand, all ICCs calculated indicated excellent reliability while, on the other hand, however, almost all parameters showed significant systematic errors as well. Although the mean differences were comparatively small with, for instance, mean differences of 1.3% (knee extension with 60°/s), this difference was considered significant ($p < .05$). This example, again, underlines the necessity to account for several facets of measurement errors when exploring reliability of testing and to consider the individual population (which includes the sample size). When measuring small samples, the ICC is biased by comparatively high variance (adding single individuals to the sample size can cause comparatively large changes in the reliability measure), while in large samples, even stable ICCs and MAPEs on a high level do not automatically indicate truly repeatable measurements, as even small mean differences can indicate significant mean differences that could be attributed to learning effects for example.

Therefore, our results are in line with previous research questioning the validity of the exclusive reporting of the ICC as a justification of test reliability. Furthermore, it is a frequently performed practice to refer to other studies with a similar testing protocol (e.g. using isokinetic tests to assess strength or other performance tests) to justify the own protocol as reliable. Since we showed that random errors seem to be specific to the test, systematic errors can be reasonably assumed to be specific to the population (high-performance athletes might be more familiar with strength testing compared to untrained participants).

In sum, this investigation shows that the MAE and MAPE (C. Willmott & Matsuura, 2005; C. J. Willmott & Matsuura, 2006) as well as the calculation of systematic errors challenge the reporting of ICC-based reliability. Our results indicate that ICCs classified as excellent and MAPEs of up to 67% can occur simultaneously as well as significant systematic errors. Additionally, this problem cannot be solved by an increase in sample size, which underlines our request for developing a more meaningful reliability classification that includes systematic and random measurement errors in sports medicine and exercise science.

Limitations

As the increase in the sample size highlighted measurement errors that are specific to the device and movement used and therefore need to be quantified, it seems obvious that our results cannot be transferred to other populations, sample sizes and testing protocols. However, this reflects exactly the main message of our study: it is paramount to determine the reliability appropriately for each data collection session for intra, and inter-day reliability by using a sample size as large as possible to figure out the actual random measurement error (MAPE). Although this will cause test–retest differences reaching the level of significance earlier, this might cause more effort in standardization of testing protocols. However, systematic learning effects must be avoided to perform a valid testing protocol (Warneke et al., 2023). When interpreting the results, the reader must note the specificity of the included population. The included participants provided a comparatively high homogeneity, as they were all recruited from an Olympic training and testing center. This circumstance might also affect reliability indices, and, consequently, the relationship between the ICC/MAPE and the sample size as well as learning effects. Since correlation statistics are not transferable between testing protocols with different samples, per se, this limitation also applies to any other reliability measure. Additionally, it must be noted that the participants were not equally distributed across the specific sports involved. This could possibly influence ICC estimates, especially in small sample sizes. Lastly, although it might be a minor aspect, several different ways to calculate the SEM have been proposed. While Weir (2005) uses the typical error introduced by Hopkins (2000) interchangeably with the SEM (thus the MDC would be calculated with another SEM), we used the SD of the difference of trial one and trial two for further SEM calculation (Lohmann et al., 2024) (see formula provided in this manuscript). Since there is no generally accepted way, both calculations might be correct; however, the reader should note that the results could slightly differ.

Conclusion

The present study highlights the limitations of relying solely on ICC for reliability assessments and underscores the importance of considering random and systematic errors. Our findings show that ICC stability is highly dependent on sample size, with estimates remaining unstable below $n = 150$, reinforcing prior research on correlation-based reliability measures. While increasing sample size reduces the variability of ICC and MAPE estimates, it does not decrease the absolute magnitude of random error, which remains test-specific. This means that while larger samples improve statistical precision, they do not eliminate random fluctuations in individual measurements. Furthermore, systematic errors persist regardless of sample size, emphasizing the need for methodological

standardization rather than statistical corrections. Despite ICC values indicating excellent reliability (≥0.9), we observed significant systematic biases in almost all strength testing conditions, highlighting the influence of learning effects, calibration inconsistencies and protocol variability. Given these limitations, increasing sample size alone does not ensure accurate reliability assessments. These findings reinforce the need for more comprehensive approach to reliability analysis, integrating ICC for relative reliability, MAPE and SEM for absolute error quantification and paired t-tests or Bland-Altman analyses for systematic bias detection. Future studies should ensure adequate sample sizes, report random and systematic errors alongside ICC and adopt rigorous test standardization to enhance the validity of strength assessments in sports science and rehabilitation research, as larger sample sizes were less vulnerable to the influence of outliers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Konstantin Warneke http://orcid.org/0000-0003-4964-2867
Michael Keiner http://orcid.org/0000-0002-1817-1743
Sebastian Puschkasch-Möck http://orcid.org/0000-0001-6277-2696

Author contribution statement

KoW provided the first draft of the manuscript and performed the statistics in collaboration with SW and MK. CG and SPM performed the data collection. SDS and MK performed the graphical illustration. KlW provided his expertise in strength and performance testing. All authors discussed and revised the manuscript and agreed to the final version.

Data availability statement

Original data can be provided from the authors due to reasonable request.

References

Atkinson, G., & Nevill, A. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(5), 375–381. https://doi.org/10.2165/00007256-200030050-00005

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine*, *26*(4), 217–238. https://doi.org/10.2165/00007256-199826040-00002

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, *17*(4), 529–569. https://doi.org/10.1080/10543400701376480

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bland, J. M., & Altmann, D. G. (1986). Statistical methods of assessing agreement between two methods of clinical measurement. *Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Blazquez, I. N., Warren, B. L., O'Hanlon, A. M., & Silvestri, L. R. (2013). An adequate interset rest period for strength recovery during a common isokinetic test. *The Journal of Strength & Conditioning Research*, *27*(7), 1981–1987. https://doi.org/10.1519/JSC.0b013e3182764d70

Chen, A., Kirkland, M. C., Wadden, K. P., Wallack, E. M., & Ploughman, M. (2020). Reliability of gait and dual-task measures in multiple sclerosis. *Gait & Posture*, *78*, 19–25. https://doi.org/10.1016/j.gaitpost.2020.03.004

Cometti, G., Maffiuletti, N. A., Pousson, M., Chatard, J.-C., & Maffulli, N. (2001). Isokinetic strength and anaerobic power of elite, subelite and Amateur French soccer players. *International Journal of Sports Medicine*, *22*(1), 45–51. https://doi.org/10.1055/s-2001-11331

Doğan, N. Ö. (2018). Bland-Altman Analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine*, *18*(4), 139–141. https://doi.org/10.1016/j.tjem.2018.09.001

French, D., & Torres Ronda, L. (2022). NSCA's essentials of sport science. In D. French & L. T. Ronda (Eds.), *NSCA's essentials of sport science* (1st ed., pp. xviii–xix). Human Kinetics.

Gleeson, N. P., & Mercer, T. H. (1996). The utility of isokinetic dynamometry in the assessment of human muscle function. *Sports Medicine*, *21*(1), 18–34. https://doi.org/10.2165/00007256-199621010-00003

Haugen, T., Seiler, S., Sandbakk, Ø., & Tønnessen, E. (2019). The training and development of elite sprint performance: An integration of scientific and best practice literature. *Sports Medicine - Open*, *5*(1), 44. https://doi.org/10.1186/s40798-019-0221-0

Hauser, T., Bartsch, D., Baumgärtel, L., & Schulz, H. (2012). Reliability of maximal lactate-steady-state. *International Journal of Sports Medicine*, *34*(3), 196–199. https://doi.org/10.1055/s-0032-1321719

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

Impellizzeri, F. M., Bizzini, M., Rampinini, E., Cereda, F., & Maffiuletti, N. A. (2008). Reliability of isokinetic strength imbalance ratios measured using the cybex NORM dynamometer. *Clinical Physiology and Functional Imaging*, *28*(2), 113–119. https://doi.org/10.1111/j.1475-097X.2007.00786.x

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koutras, G., Bernard, M., Terzidis, I. P., Papadopoulos, P., Georgoulis, A., & Pappas, E. (2016). Comparison of knee flexion isokinetic deficits between seated and prone positions after ACL reconstruction with hamstrings graft: Implications for rehabilitation and return to sports decisions. *Journal of Science & Medicine in Sport*, *19*(7), 559–562. https://doi.org/10.1016/j.jsams.2015.07.018

Lamb, K. (1998). Test-retest reliability in quantitative physical education research: A commentary. *European Physical Education Review*, *4*(2), 145–152. https://doi.org/10.1177/1356336X9800400205

Lienhard, K., Lauermann, S. P., Schneider, D., Item-Glatthorn, J. F., Casartelli, N. C., & Maffiuletti, N. A. (2013). Validity and reliability of isometric, isokinetic and isoinertial modalities for the assessment of quadriceps muscle strength in patients with total knee arthroplasty. *Journal of Electromyography and Kinesiology*, *23*(6), 1283–1288. https://doi.org/10.1016/j.jelekin.2013.09.004

Lohmann, L. H., Hillebrecht, M., Schiemann, S., & Warneke, K. (2024). Stressing the relevance of differentiating between systematic and random measurement errors in ultrasound muscle thickness diagnostics. *Sports Medicine - Open*, *10*(1). https://doi.org/10.1186/s40798-024-00755-z

Martins, J., da Silva, J. R., da Silva, M. R. B., & Bevilaqua-Grossi, D. (2017). Reliability and validity of the belt-stabilized handheld dynamometer in hip- and knee-strength tests. *Journal of Athletic Training*, *52*(9), 809–819. https://doi.org/10.4085/1062-6050-52.6.04

Menzel, H.-J., Chagas, M. H., Szmuchrowski, L. A., Araujo, S. R. S., de Andrade, A. G. P., & de Jesus-Moraleida, F. R. (2013). Analysis of lower limb asymmetries by isokinetic and vertical jump tests in soccer players. *The Journal of Strength & Conditioning Research*, *27*(5), 1370–1377. https://doi.org/10.1519/JSC.0b013e318265a3c8

Möck, S., Happ, K., & Wirth, K. (2023). The evaluation of strength imbalances as risk factor for contactless injuries of the knee and thigh: A critical review. *The Journal of Sports Medicine and Physical Fitness*, *63*(5). https://doi.org/10.23736/S0022-4707.23.14501-4

Möck, S., & Wirth, K. (2024). Bilateral differences of isokinetic knee extensor strength are velocity- and task-dependent. *Sports Biomechanics*, *23*(12), 3641–3653. https://doi.org/10.1080/14763141.2024.2315260

Nevill, A. M., & Atkinson, G. (1997). Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *British Journal of Sports Medicine*, *31*(4), 314–318. https://doi.org/10.1136/bjsm.31.4.314

Perrin, D. H. (1993). *Isokinetic exercise and assessment* (Vol. 1). Human Kinetics Publishers.

Roth, R., Donath, L., Kurz, E., Zahner, L., & Faude, O. (2017). Absolute and relative reliability of isokinetic and isometric trunk strength testing using the IsoMed-2000 dynamometer. *Physical Therapy in Sport*, *24*, 26–31. https://doi.org/10.1016/j.ptsp.2016.11.005

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Seamon, B. A., Kautz, S. A., Bowden, M. G., & Velozo, C. A. (2022). Revisiting the concept of minimal detectable change for patient-reported outcome measures. *Physical Therapy*, *102*(8). https://doi.org/10.1093/ptj/pzac068

Tanner, R., & Gore, C. (2012). Physiological tests for elite athletes. In K. T. Rebecca & J. G. Christopher J. Gore (Eds.), *Physiological tests for elite athletes* (2nd ed.). Human Kinetics.

Tighe, J., McManus, I., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). 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 Medical Education*, *10*(1), 40. https://doi.org/10.1186/1472-6920-10-40

Warneke, K., Gronwald, T., Wallot, S., Magno, A., Hillebrecht, M., & Wirth, K. (2025). Discussion on the validity of commonly used reliability indices in sports medicine and exercise science - a critical review with data simulations. *European Journal of Applied Physiology*. https://doi.org/10.1007/s00421-025-05720-6

Warneke, K., Wagner, C.-M., Keiner, M., Hillebrecht, M., Schiemann, S., Behm, D. G., Wallot, S., & Wirth, K. (2023). Maximal strength measurement: A critical evaluation of common methods-a narrative review. *Frontiers in Sports and Active Living*, *5*, 1105201. https://doi.org/10.3389/fspor.2023.1105201

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231. https://doi.org/10.1519/15184.1

Willigenburg, N. W., & Poolman, R. W. (2023). The difference between statistical significance and clinical relevance. The case of minimal important change, non-inferiority trials, and smallest worthwhile effect. *Injury*, *54*, 110764. https://doi.org/10.1016/j.injury.2023.04.051

Willmott, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science*, *20*(1), 89–102. https://doi.org/10.1080/13658810500286976

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wollin, M., Purdam, C., & Drew, M. K. (2016). Reliability of externally fixed dynamometry hamstring strength testing in elite youth football players. *Journal of Science & Medicine in Sport*, *19*(1), 93–96. https://doi.org/10.1016/j.jsams.2015.01.012

Yao, K., & Gao, J. (2016). Law of large numbers for uncertain random variables. *IEEE Transactions on Fuzzy Systems*, *24*(3), 615–621. https://doi.org/10.1109/TFUZZ.2015.2466080