Brief Communication

Ger J Exerc Sport Res
https://doi.org/10.1007/s12662-025-01084-3
Received: 20 February 2025
Accepted: 13 November 2025

Konstantin Warneke[1,2] · Stanislav D. Siegel[2] · José Afonso[3] · Sebastian Wallot[1]

[1] Department of Human Movement and Exercise Physiology, Friedrich Schiller University Jena, Jena, Germany

[2] Institute of Sustainability Psychology, Leuphana University Lüneburg, Lüneburg, Germany

[3] Centre of Research, Education, Innovation, and Intervention in Sport (CIFI2D), Faculty of Sport, University of Porto, Porto, Portugal

What the mean absolute percentage error (MAPE) should adopt from Bland–Altman analyses

Background

Behavioural and social sciences rely on human involvement, whether as test subjects or assessors, which introduces unique challenges to measurement protocol accuracy (e.g. requirement of highly standardized procedures, skilled assessors and habituated participants). These challenges must be considered when interpreting collected data, as human behaviour and responses are inherently variable and influenced by numerous factors. These include, among others, measurement errors stemming from instruments (device reliability) and measurement errors appearing within the testing procedures (protocol reliability) (Warneke et al., 2025a; Kottner et al., 2011; de Vet, Terwee, Knol, & Bouter, 2006).

Even if it was possible to standardize all surrounding factors (e.g. temperature, time of the day, protocol conditions) while using 100% accurate devices, secondary variance of testing results would still not approach 0%, as biological variability (e.g. motivation, circadian rhythms, learning effects) influences the reliability of the collected data (de Vet et al., 2006; Patel et al., 2021; Ozturk et al., 2021). Moreover, especially in subjective tests where skilled and experienced assessors are required for collecting data (e.g. ultrasound (Warneke et al., 2022; Warneke et al., 2025b), myoton (Brandl, Acikalin, Bartsch, Wilke, & Schleip, 2024; Warneke et al., 2025c), goniometer (Pazira, Rostami Haji-Abadi, Zolaktaf, Sabahi, & Pazira, 2016)), the human-related variability must not only be considered for participants, but also for the assessors. For example, experience moderates ultrasound investigations, even if the ultrasound is considered reliable (Warneke et al., 2025b). These errors can accumulate and contribute to biased measurement results that hamper meaningful conclusions drawn from research (Srinivasan, Westover, & Bianchi, 2012). This contrasts with the consensus regarding the need for data to be collected under reliable, objective and valid conditions (Hopkins, 2000). What are the consequences of all of this?

Quality criteria in data collection—short summary of terms and definitions

In its broadest sense, reliability is often understood as the repeatability (repeating test under same conditions) and reproducibility (reproducing the same test but in a different population or in, for instance, another laboratory) of measurements as one essential precondition for valid data collection (Hopkins, 2000; Atkinson & Nevill, 2000; Barnhart, Haber, & Lin, 2007; Vetter & Schober, 2018). However, although often reported to denote that the test was conducted without meaningful measurement errors, reliability and measurement error analyses refer to very different concepts. Reliability indices quantified as excellent allow no statement about measurement errors (Warneke et al., 2025a; de Vet et al., 2006). This difference was previously illustrated by Lamb (Lamb, 1998), de Vet et al. (de Vet et al., 2006) and Warneke et al. (Warneke et al., 2025a), showing that excellent reliability values such as the ICC can be accompanied by meaningful measurement errors. Tests can have high reliability values while containing considerable measurement error, simultaneously (Warneke et al., 2025a). Furthermore, the ICC oftentimes neglected systematic errors, meaning that there were significant mean differences between the test and the retest. For example, if participants were unfamiliar with a test, they habituate with the test protocol and progressively increase from test-to-test. For instance, several habituation sessions were necessary to reach a stabile testing niveau in strength tests (Ploutz-Snyder & Giamis, 2001) or neurocognitive tasks (Warneke et al., 2025d; Reimers & Maylor, 2005), although the ICC already indicated excellent reliability between the first and second test session.

A lack of awareness for the difference between excellent ICCs and small measurement errors can have serious implications for data interpretation of empirical

Published online: 05 December 2025

Abbreviations	
BA analysis	Bland–Altman analysis
BPM	Beats per minute
95% CI	95% confidence interval
ICC	Intraclass correlation coefficient
LoA	Limits of agreement
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MDC	Minimal detectable change
MPE	Maximal percentage error
SEM	Standard error of measurement
SWC	Smallest worthwhile change

results. If a repeated test does not lead to the same result even when performed under the same conditions (intraday reliability), any conclusions derived from this data must be questioned (Hopkins, 2000; Atkinson & Nevill, 1998). Although reliable measurements can be seen as the baseline for internal data validity there are further aspects to be considered. Even if an assessment is highly repeatable and reproducible, reliability should still be considered a necessary, but insufficient condition for validity (e.g. the countermovement jump is not valid to assess shoulder range of motion).

Definitions from the literature

Given the paramount importance of reliability for all subsequently applied calculations and interpretation of data, the reader must note that no overall consensus is available in the literature. Some concepts are illustrated here, while others can be reviewed elsewhere (Warneke et al., 2025a; de Vet et al., 2006; Hopkins, 2000; Atkinson & Nevill, 2000; Atkinson & Nevill, 1998; Nevill & Atkinson, 1997; Hopkins, 2004; Mokkink et al., 2020; Weir, 2005; de Vet, Terwee, Mokkink, & Knol, 2011). As Barnhart and colleagues put it, "In the simplest intuitive terms, reliable and accurate measurement may simply mean that the new measurement is the same as the truth or agrees with the truth" (Barnhart et al., 2007, p. 530). A ground truth is not always available or is only measurable with some degree of error. Reliability metrics attempt to account for deviations between two measurements over the investigated population to assess how *precisely* the truth can be determined. The Food and Drug Administration (Food and Drug Administration, 2001) suggested "precision as the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under the prescribed conditions" (Barnhart et al., 2007, p. 532). Conversely, accuracy reflects the closeness of mean test results obtained by the method to the true value (concentration) of the analysis (Schmidt et al., 2018). The deviation of the mean from the true value, i.e. systematic bias, serves as the measure of accuracy (Barnhart et al., 2007). In simple terms, precision refers to hitting *a* spot under the same conditions (reflecting random error), while accuracy requires hitting the *correct* spot consistently (accounting for both random and systematic error; Warneke et al., 2025a).

Therefore, accuracy refers to the closeness of the average of a result to the true value, with minimal deviations (thus includes validity and precision). However, in a pre–post study design with a control group, systematic bias may be mitigated by study design, making precision more critical than accuracy (i.e. as long as the error is consistent in direction and magnitude, we can still confidently assess pre- to post-changes). These definitions suggest an ample relevance of measurement error quantifications reaching beyond common reliability quantification when referring to accuracy and precision to underscore internal data validity. Although the relevance of a valid quantification is obvious, the literature provides numerous metrics without a clear consensus regarding the gold standard method. To move forward, we propose an improvement of measurement error quantification, but first several common metrics must be introduced.

Reliability quantification in literature

Assessment of measurement errors dates back to at least the late 19th or early 20th centuries. Pearson (1896–1901) contributed with articles of mathematical approaches to the theory of evolution to assess the similarity and variability in children with the same parents. Quoting Pearson: "(…), the degree of correlation between brothers and the variability of an array of brothers due to the same parentage, so we can determine the correlation, i.e., the degree of resemblance between the undifferentiated like organs of the individual and the degree of variability within the individual. This determination is the answer to our first fundamental problem, that of the ratio of individual to racial variability" (Pearson, 1901, p. 287).

Correlations provide meaningful information (e.g. the strength of the relationship between two variables) but were not designed to quantify reliability. Making a long story short: the popular Pearson correlation does not account for systematic or random error as it represents the relation between two parameters, not the agreement of measurements (Lin, 1989; Bland & Altman, 1999). Therefore, large correlations may sometimes mask systematic bias. Nevertheless, authors commonly provide the intraclass correlation coefficient (ICC), usually classified according to Koo and Li (Koo & Li, 2016) ICCs ≥ 0.9 as excellent, ≥ 0.7 as good and < 0.7 as insufficient. However, the ICC does not provide sufficient information on measurement errors (other works extensively discussed this issue (Warneke et al., 2025a; Hopkins, 2000; Atkinson & Nevill, 2000; Lamb, 1998; Atkinson & Nevill, 1998; Nevill & Atkinson, 1997)) as it is simply a correlation coefficient extended by an analysis of variance of the included values. To account for random and systematic errors, Atkinson and Nevill (Atkinson & Nevill, 2000) suggested the Bland–Altman (BA) analysis (Doğan, 2018; Bland & Altmann, 1986). This was taken up by some current papers outlining the need for implementing BA analyses (Grgic, Lazinica, Schoenfeld, & Pedisic, 2020; Zöger, Nimmerichter, Baca, & Wirth, 2025; Berchtold, 2016), as agreement between the test–retest was recognized as more important than simply exhibiting a large correlation.

There are several formulas to quantify the standard error of measurement (SEM; Mokkink et al., 2020; Weir, 2005; Palta et al., 2011; Harvill, 1991), leading to different baseline values for follow-up calculations such as the minimal detectable change (MDC) (Seamon, Kautz, Bowden, & Velozo, 2022; Powden, Hoch, & Hoch, 2015) or the smallest worthwhile change (SWC; de Vet et al., 2006; Willigenburg & Poolman, 2023). Interestingly, some formulas to calculate the SEM seem identical to the typical error (TE; Hopkins, 2000). These additional metrics were introduced to dampen the ICCs' limitations of just providing a value on reliability, not accounting for the measurement errors (Warneke et al., 2025a; de Vet et al., 2006). There are noteworthy caveats when implementing these supplementary metrics. If calculating the SEM and the MDC based on the ICC (see formulas provided in **Table 1**) these provide only a minor benefit for measurement error quantification when assuming the ICC lacks validity to account for precision and accuracy, per se (Warneke et al., 2025a). The SWC is often applied by performing 0.2*SD (Marocolo et al., 2019) under the assumption of 0.2 as the threshold between small and trivial effects (Warneke et al., 2025a; Cohen, 1988), while also SEM formulas without the inclusion of the ICC should be preferred (de Vet et al., 2006; Mokkink et al., 2020).

The BA analysis was introduced by Bland and Altman (Bland & Altmann, 1986) to evaluate the agreement between blood pressure devices. The measurement error in the BA analysis is graphically illustrated by plotting the difference between device value 1 and device value 2 on the y-axis, and the mean of the difference on the x-axis. This provides a graphical illustration of the absolute measurement error scattering in relationship to the mean of the tests (also graphically illustrated via the limits of agreement). Also, the mean difference is graphically illustrated and stated as the systematic bias line. If one device systematically measures higher/lower values than the other device, the systematic bias line will deviate from zero. This measurement error analysis was suggested to be included into reliability analysis by plotting the test–retest difference in dependency of the test–retest mean (Warneke et al., 2025a; Atkinson & Nevill, 1998).

The BA analysis and the TE also have meaningful limitations. First, transforming test–retest data organization to second best–best order affects precision and accuracy, casting doubts on the objectivity and validity of the BA analysis and the TE (for more details, see Warneke et al. 2025a). Second, interpreting the limits of agreement (LoA) from BA analyses as a measure for the random scattering requires appropriate reporting and interpretation to provide practically relevant conclusions (Gerke, 2020). This is rarely performed in the literature. For instance, Grgic et al. (Grgic et al., 2020) showed that only a handful of the included studies (12 in total) performed Bland–Altman analysis for reliability evaluation in maximal strength testing. They reported LoA from the literature, ranging from ± 3–5 kg for the bench press, ± 5–8 kg for the power clean, ± 8–13 kg for the leg press, and ± 10–15 kg for the squat. These LoAs (as stand-alone information) cannot be interpreted without the respective reference means, underscoring the relevance of key information that are practically relevant. Whether these values are of practical relevance depends on the mean which is surrounded by these differences. To illustrate: a LoA range from –5 to 5 kg might be acceptable if referring to a mean of (e.g.) 500 kg. If the mean strength was only 10 kg, these estimated LoAs must be interpreted as unacceptable. This reporting problem, however, is not limited to sports-related parameters, but can also reviewed for clinical/medical outcomes such as stiffness (Brandl et al., 2024). The question arises whether there are better ways to quantify and report reliability so that practitioners and clinicians can easily interpret the metrics.

Another often-overlooked issue in the literature is that device-related reliability may be highly specific not only for the metric being assessed, but also sensitive to dose. Perhaps the most well-known example derives from reliability studies of lactate analysers. Assessment of multiple devices (i.e. different brands and different models for a given brand) across blood samples with different lactate concentrations showed that the devices' reliability is not uniform and instead varies depending on the metabolite concentration (Bonaventura et al., 2015).

Ways forward—providing an easy interpretable absolute error quantification

The BA analysis is an appropriate way to explore the agreement between devices

Abstract

Ger J Exerc Sport Res
https://doi.org/10.1007/s12662-025-01084-3

K. Warneke · S. D. Siegel · J. Afonso · S. Wallot

What the mean absolute percentage error (MAPE) should adopt from Bland–Altman analyses

Abstract

Reporting reliability with precision and accuracy is of paramount importance in empirical data collections to ascertain whether data is trustworthy. Reliability is often quantified using the intraclass correlation coefficient (ICC), from which the standard error of measurement (SEM) and the minimal detectable change (MDC) can be calculated. However, the literature outlined limited validity of the ICC to account for systematic and random measurement errors stemming from learning or fatiguing effects or a lack of standardization or normal biological variability, respectively. Therefore, the Bland–Altman analysis was introduced to illustrate the systematic bias and quantify the random error via the limits of agreement, originally used to evaluate agreement between devices. Unfortunately, the literature presents common interpretation problems, including missing reference values or misunderstanding of the message transported by the upper and lower border of the Bland–Altman analysis. In this communication paper, we introduce a modified reporting of the mean absolute percentage error that can solve this interpretation problem as it provides a percentage reporting of the mean random error and orientate on the random error calculation performed in the Bland–Altman analysis.

Keywords

Random error · Systematic bias · Precision · Accuracy · Reliability

Table 1 Formulas for statistical metrics associated with reliability calculations.

Metric	Formula
ICC (general)—intraclass correlation coefficient	$ICC = \frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2}$ σ_B^2 = between-subject variance σ_w^2 = within subject variance
ICC(1,1)—one-way random effects model	$ICC(1,1) = \frac{MS_B - MS_W}{MS_B+(k-1)MS_W}$ MS_B = between-subject mean square MS_W = within subject mean square k = number of raters per subject
ICC(2,1)—two-way random effects model	$ICC(2,1) = \frac{MS_B - MS_E}{MS_B+(k-1)MS_W+k\frac{MS_E}{n}}$ MS_B = between-subject mean square MS_W = within subject mean square MS_E = residual mean square k = number of raters per subject n = number of subjects
ICC(3,1)—two-way mixed effects model	$ICC(3,1) = \frac{MS_B - MS_E}{MS_B+(k-1)MS_E}$ MS_B = between-subject mean square MS_E = residual mean square k = number of raters per subject n = number of subjects
Mean absolute error (MAE)	$MAE = \frac{1}{n} \times \sum_{i=1}^{n} \lvert x_i - y_i \rvert$ 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
Mean absolute percentage error (MAPE)	$\text{MAPE} = \frac{1}{n} \times \sum_{i=1}^{n} \left\lvert \frac{x_i - y_i}{x_i} \right\rvert * 100$ 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
Smallest worthwhile change (SWC)	$SWC = 0.2 \times SD_{baseline}$ 0.2 = based on Cohen's d (small effect size), can vary depending on the study $SD_{baseline}$ = standard deviation at baseline
Standard error of measurement (SEM)—ICC based	$SEM = \frac{SD_{Diff}}{\sqrt{2}}$ $SEM_{ICC} = SD_{baseline} \times \sqrt{1 - ICC}$ $SEM_{agreement} = \sqrt{\sigma_{rater}^2 + \sigma_W^2}$ $SEM_{consistency} = \sqrt{\sigma_W^2}$ SD_{Diff} = standard deviation of measurement differences σ_{rater}^2 = between-rater variance σ_W^2 = within subject variance ICC = intra class correlation coefficient at baseline $SD_{baseline}$ = standard deviation at baseline
Minimal detectable change (MDC)	$MDC = z \times SEM \times \sqrt{2}$ z = z-score for confidence level (e.g. 1.96 for 95% CI) SEM = Standard error of measurement
Typical error (TE)	$TE = \frac{SD_{dif}}{\sqrt{2}}$ SD_{dif} = standard deviation of the difference between to repeated measurements for the same subjects

to assess device objectivity/validity and within one device between days (Interday reliability). For instance, Ranganathan et al. (Ranganathan, Pramesh, & Aggarwal, 2017) validated haemoglobin measurement devices by using limits of agreement (LoA). The LoAs refer to the range in which 95% of the test–retest differences are plotted against the test–retest mean. 5% of the mean differences are outside these limits. This quantification of measurement errors faces practitioners and researchers with problems of interpretability. In one example, the authors concluded from a BA analysis LoAs analysis that the testing was reliable as "All the measurement points of the device were within the limits of agreement" (Brandl et al., 2024), which can only be attributed to $n = 16$. Secondly, the authors highlighted their thoracolumbar fascia deformation determination as being highly reliable since "Bland–Altman plots for intra- and interrater reliability showed that almost all points were within the limits of agreement", making their method appropriate to detect lower back pain (justifying clinical relevance via BA analysis). Thirdly, Koch & Wilke (2022) wrote "that the points were scattered in an unbiased pattern, with most of them falling within the limits of agreement". These examples illustrate that BA analyses are often inappropriately discussed. Any exclusively qualitative evaluation without clear measurement error quantification calls for improvements. One possibility to quantify the random scattering is provided via the mean absolute error (MAE) (calculation formula, **Table 1**).

The MAE provides an absolute error quantification by calculating the difference between the test–retest. This metric is objective and does not depend on data organization, meaning the MAE does not change when reorganizing test–retest data order to best-second best organisation (Warneke et al., 2025a). A further advantage is that it is reported in the measurement unit of the monitored parameter, which eases interpretation. It can be seen as an expression of the mean random scattering, while the paired sampled t-test between the test–retest results within a reliability analysis quantifies the systematic error (e.g. to account for warm-up, cool-down, or systematically differing measurement techniques between assessors) (Atkinson & Nevill, 2000; Barnhart et al., 2007). As for the BA analysis, we are confronted with the limitation of reporting the reference mean value (e.g. maximal strength was determined with 100 kg with a MAE of 5 kg). If the reference value (100 kg) is not provided, the reader cannot interpret whether a 5 kg LoA/MAE/SEM is to be classified as high or not, as 5 kg error

around 1000 kg seems negligible, while a 5 kg measurement error around a mean of 10 kg indicates a precision problem. To counteract, the mean absolute percentage error (MAPE) can be additionally provided (Willmott & Matsuura, 2005; Kim & Kim, 2016) (**Table 1**) as a dimensionless quantification of the mean absolute error. A 5% measurement error automatically refers to a reference value (5 kg measurement error around a mean of 1000 kg was 0.5%, while the 10 kg with 5 kg MAE is provided as 50% MAPE). Therefore, the MAPE could quantify the mean random scattering (a measure for precision), which should be interpreted in combination with the LoAs of the BA analysis and the systematic error, as a high MAPE can occur with or without a meaningful systematic bias. Taken together, a minimal variance (excellent ICC) without a systematic error and a small random error would indicate accurate testing (Warneke et al., 2025a). However, the MAPE in its current form can lead to invalid conclusions as percentage values always need a reference value, therefore, the question arises: percentage of what?

What the MAPE should adopt from the Bland–Altman analysis

In our previous review (Warneke et al., 2025a) we highlighted that excellent ICCs (Koo & Li, 2016) (as the most common metrics to quantify reliability) can be associated with meaningful random error. Moreover, ICCs increased when data were re-organized from test–retest to best-second best (Warneke et al., 2025a). Therefore, exclusively focusing on the ICC (especially in small sample sizes) can result in misleading results that bias subsequent conclusions. BA plots were also vulnerable for data re-organization, indicating improved precision when data were presented in best-to second best arrangement. The only parameter that remained robust for such problems was the MAE, favouring its implementation in measurement error analyses. The MAE can also be expressed as the MAPE. This percentage expression of measurement errors provides an easily interpretable metric, but dependent on the reference value.

What if there is a systematic error (impacting accuracy (Barnhart et al., 2007)), but we want to quantify the absolute error in percent? This problem can be reviewed in **Table 2**, showing (simulated) heart rate screening test–retest values with a mean of 64.07 ± 10.50 beats per minute (BPM) in the first test, and 64.20 ± 13.74 BPM in the second test. The MAE of 10.8 BPM indicates a precision rather than an accuracy problem as measurement errors can be attributed to random errors and no significant systematic error was found via the paired sample t-test for retest–test difference which is 0.13 BPM with $p = 0.92$. Depending on taking the test- or the retest as reference, the corresponding MAPE is 17.15% or 17.00%. If we are now interested in maximal heart rate and reorganize data (best-second best) as done in Warneke et al. (Warneke et al., 2025a), we produce a systematic error in our reliability statistics with $p < 0.001$, as the descriptives change to 58.73 ± 11.00 BPM and 69.53 ± 10.88 BPM, respectively. The MAE stayed constant with 10.8 BPM, so the question arises of which reference value should be chosen in this case. The problem is obvious as the MAPE depends on the reference value: referring to the higher heart rate as the baseline results in a difference which is 15.64% (MAPE). Choosing the lower value, the MAE referred to a smaller baseline; therefore, the percentage error increases to 19.30% when referring to the second best. How can this problem be solved?

For both the MAE and the MAPE, the reference value is important and both formulas refer to one value as the reference. However, when considering MAPE in future analyses as a percentage quantification of the random scattering graphically illustrated via the BA analysis (Carstensen & Ekstrom, 2011), the problem of the reference value is solved by improving two limitations: a) the reference problem of the MAPE by not referring to x_i but to the mean of the test–retest difference and b) the interpretation limitation of BA analysis. Therefore, we suggest a quantification by adjusting the MAPE as follows:

$$\mathrm{MAPE}_{adjusted} = \frac{1}{n} * \sum_{i=1}^{n} \left| \frac{x_i - y_i}{\frac{(x_i + y_i)}{2}} \right| * 100$$

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

This results in an adjusted MAPE of 16.87%, independent of whether there was a systematic error or not. As the BA plot illustrates the mean difference in dependency on the mean, this MAPE calculation leads to a percentage expression of the mean of the plotted measurement errors in the BA analysis. Therefore, in combination with the maximal percentage error (MPE; to account for outliers) it provides an objective measure for reliability, interpretable without knowledge of the reference mean (see, for only one example, meta-analysis from Grgic et al. (Grgic et al., 2020)). In contrast, although the MAE stays constant even in the presence of large systematic errors in the data organization, the question about the reference value is linked with the MAPE problem. In our example, the heart rate MAE of 10.8 BPM must be interpreted differently in case the mean heart rate is 58 BPM or 69 BPM. To solve this problem, the mean of both values should be considered, and the calculation formula of the MAE does not need to be changed for that.

Precision versus accuracy—what is reliability referring to?

While applying the BA analysis provides a combination of a quantitative and qualitative evaluation of the systematic and random error, it is not clear if both are required to assess reliability. It is currently unclear if reliability refers only to precision, or if accuracy is required. Accurate data collection would require hitting one spot repeatedly and assume that spot is the right ("true") one (whatever this might be). Equating reliability with accurate data collection would therefore cause a confusion of different metrics,

Table 2 Data Simulation for heart rate measurement to illustrate the problem of data reorganization (following Warneke et al. 2025)

Participant	HR T1	HR T2	MAE	MAPE T2–T1	MAPE T1–T2	Lower HR	Higher HR	MAE	MAPE T2–T1	MAPE T1–T2	MAPE adjusted
1	62	54	8	14.81	12.9	54	62	8	12.9	14.81	13.79
2	61	76	15	19.74	24.59	61	76	15	19.74	24.59	21.90
3	55	45	10	22.22	18.18	45	55	10	18.18	22.22	20.00
4	64	55	9	16.36	14.06	55	64	9	14.06	16.36	15.13
5	52	64	12	18.75	23.08	52	64	12	18.75	23.08	20.69
6	67	71	4	5.63	5.97	67	71	4	5.63	5.97	5.80
7	81	66	15	22.73	18.52	66	81	15	18.52	22.73	20.41
8	70	55	15	27.27	21.43	55	70	15	21.43	27.27	24.00
9	47	49	2	4.08	4.26	47	49	2	4.08	4.26	4.17
10	61	51	10	19.61	16.39	51	61	10	16.39	19.61	17.86
11	66	75	9	12	13.64	66	75	9	12	13.64	12.77
12	78	84	6	7.14	7.69	78	84	6	7.14	7.69	7.41
13	65	53	12	22.64	18.46	53	65	12	18.46	22.64	20.34
14	62	50	12	24	19.35	50	62	12	19.35	24	21.43
15	68	82	14	17.07	20.59	68	92	24	26.09	35.29	18.67
16	64	51	13	25.49	20.31	51	64	13	20.31	25.49	22.61
17	48	55	7	12.73	14.58	48	55	7	12.73	14.58	13.59
18	59	51	8	15.69	13.56	51	59	8	13.56	15.69	14.55
19	45	67	22	32.84	48.89	45	67	22	32.84	48.89	39.29
20	80	69	11	15.94	13.75	69	80	11	13.75	15.94	14.77
21	74	64	10	15.63	13.51	64	74	10	13.51	15.63	14.49
22	56	45	11	24.44	19.64	45	56	11	19.64	24.44	21.78
23	79	83	4	4.82	5.06	79	83	4	4.82	5.06	4.94
24	64	72	8	11.11	12.5	64	72	8	11.11	12.5	11.76
25	80	89	9	10.11	11.25	80	89	9	10.11	11.25	10.65
26	61	69	8	11.59	13.11	61	69	8	11.59	13.11	12.31
27	80	88	8	9.09	10	80	88	8	9.09	10	9.52
28	67	51	16	31.37	23.88	51	67	16	23.88	31.37	27.12
29	58	67	9	13.43	15.52	58	67	9	13.43	15.52	14.40
30	48	65	17	26.15	35.42	48	65	17	26.15	35.42	30.09
Mean	64.07	63.87	10.47	17.15	17	58.73	69.53	10.8	15.64	19.3	16.87
SD	10.5	13.16	–	–	–	10.79	10.7	–	–	–	–
MAX Error	–	–	–	32.84	48.89	–	–	–	32.84	48.89	39.29
Correlation	–	–	–	0.55	–	–	–	–	0.9	–	–

HR heart rate, *MAE* mean absolute error, *MAPE* mean absolute error percentage error, *SD* standard deviation

leading us to the suggestion to refer to precision when talking about reliability, while the actual requirement of data collection with clinical and practical meaning was accuracy. This differentiation is essential, as inaccurate measurements in medical and applied settings can result in misdiagnoses or inappropriate interventions (Parker et al., 2023). While precision ensures consistency, only accuracy guarantees that decisions based on the data truly reflect the underlying physiological or biomechanical reality. Therefore, a combination of the BA analysis, an investigation of the systematic error (inference quantification) is required, while the MAPE serves as a valid supplementation as a quantification of the mean random error.

Limitations

One limitation of the current approaches is that only two tests (timepoints, methods) can be plotted against each other. When having more than two test time points, advanced approaches are required to include e.g. three or four columns (Ekstrøm & Carstensen, 2024). However, this discussion goes beyond the scope of this work, which is to improve the MAPE formula. Furthermore, although BA analyses provide relevant information on agreement between two methods, the random error estimation via LoAs must be viewed as an estimation that was improved by determining the exact limits (Carkeet, 2015; Carkeet & Goh, 2018). Another limitation that can be seen in the formula of the MAE is that it suggests one value (test 1 or test 2) indicates that

one was the reference value. This is not always the case.

Conclusion

In addition to focusing on relative reliability metrics such as the intraclass correlation coefficient (ICC), there is a need to account for systematic and random errors. The mean absolute error (MAE) and mean absolute percentage error (MAPE) provide easily interpretable metrics, representing the actual absolute measure in relationship to its mean. While the Bland–Altman (BA) analysis provides a promising approach, quantification and interpretation are often performed inappropriately in the literature, calling for additional approaches that help the reader understand what was shown by the BA analysis. The MAPE adjustment provided here can serve as a dimensionless and easily interpretable quantification of precision without the limitations of commonly implemented metrics. Overall, current reliability measurements focus excessively on precision, and greater efforts should be implemented to assess accuracy.

Corresponding address

Dr. Konstantin Warneke
Institute of Sustainability Psychology, Leuphana University Lüneburg
Lüneburg, Germany
Konstantin.Warneke@uni-jena.de

Acknowledgements. N/A.

Funding. N/A.

Author Contribution. KW developed the idea for the manuscript, provided the first draft, simulated the data and performed the statistical calculations. SDS and JA provided critical feedback contributed with references and revised the first draft. SW supervised the project.

Funding. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and material. N/A.

Declarations

Conflict of interest. K. Warneke, S.D. Siegel, J. Afonso and S. Wallot declare that they have no competing interests.

Consent for publication N/A.

Ethics approval and consent to participate N/A.

For this article no studies with human participants or animals were performed by any of the authors. All studies mentioned were in accordance with the ethical standards indicated in each case.

References

Atkinson, G., & Nevill, A. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*, 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*, 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. *J Biopharm Stat*, *17*, 529–569. https://doi.org/10.1080/10543400701376480.

Berchtold, A. (2016). Test-retest: agreement or reliability? *Method Innov*. https://doi.org/10.1177/2059799116672875.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res*, *8*, 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*, *i*, 307–310.

Bonaventura, J. M., Sharpe, K., Knight, E., Fuller, K. L., Tanner, R. K., & Gore, C. J. (2015). Reliability and accuracy of six hand-held blood lactate analysers. *J Sports Sci Med*, *14*, 203–214.

Brandl, A., Acikalin, E., Bartsch, K., Wilke, J., & Schleip, R. (2024). Reliability and validity of an app-assisted tissue compliance meter in measuring tissue stiffness on a phantom model. *PeerJ*, *12*, e17122. https://doi.org/10.7717/peerj.17122.

Carkeet, A. (2015). Exact parametric confidence intervals for Bland-Altman limits of agreement. *Optometry and Vision Science*, *92*, e71–e80. https://doi.org/10.1097/OPX.0000000000000513.

Carkeet, A., & Goh, Y. T. (2018). Confidence and coverage for Bland-Altman limits of agreement and their approximate confidence intervals. *Stat Methods Med Res*, *27*, 1559–1574. https://doi.org/10.1177/0962280216665419.

Carstensen, B., & Ekstrom, C. T. (2011). Statistical analaysis of method comparison studies. http://BendixCarstensen.com/MethComp/Ancona.2011

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd edn.).

Doğan, N. Ö. (2018). *Bland-Altman analysis: a paradigm to understand correlation and agreement*. https://doi.org/10.1016/j.tjem.2018.09.001.

Ekstrøm, C. T., & Carstensen, B. (2024). Statistical models for assessing agreement for quantitative data with heterogeneous random raters and replicate measurements. *Int J Biostat*, *20*, 455–466. https://doi.org/10.1515/ijb-2023-0037.

Food and Drug Administration (FDA) Guidance for industry: bioanalytical method validation. http://ww.fda.gov/cder/guidance/index.htm.2001

Gerke, O. (2020). Reporting standards for a Bland-Altman agreement analysis: a review of methodological reviews. *Diagnostics*, *10*, 334. https://doi.org/10.3390/diagnostics10050334.

Grgic, J., Lazinica, B., Schoenfeld, B. J., & Pedisic, Z. (2020). Test-retest reliability of the one-repetition maximum (1 RM) strength assessment: a systematic review. *Sports Med Open*, *6*, 31. https://doi.org/10.1186/s40798-020-00260-z.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*, 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x.

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

Hopkins, W. G. (2004). How to interpret changes in an athletic performance test. *Sportscience*, *8*, 1–7.

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *Int J Forecast*, *32*, 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. *J Chiropr Med*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., et al. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*, *64*, 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.

Lamb, K. (1998). Test-retest reliability in quantitative physical education research: a commentary. *Eur Phy Educ Rev*, *4*, 145–152. https://doi.org/10.1177/1356336X9800400205.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. https://www.jstor.org/stable/2532051

Marocolo, M., Simim, M. A. M., Bernardino, A., Monteiro, I. R., Patterson, S. D., & da Mota, G. R. (2019). Ischemic preconditioning and exercise performance: shedding light through smallest worthwhile change. *Eur J Appl Physiol*, *119*, 2123–2149. https://doi.org/10.1007/s00421-019-04214-6.

Mokkink, L. B., Boers, M., van der Vleuten, C. P. M., Bouter, L. M., Alonso, J., Patrick, D. L., et al. (2020). COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol*, *20*, 293. https://doi.org/10.1186/s12874-020-01179-5.

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

Ozturk, O. G., Coskun, A., Yucel, D., Yazici, C., Serteser, M., Senes, M., et al. (2021). Utilization of biological variation data in the interpretation of laboratory test results—survey about clinicians' opinion and knowledge. *Biochem Med*, *31*, 93–102. https://doi.org/10.11613/BM.2021.010705.

Palta, M., Chen, H.-Y., Kaplan, R. M., Feeny, D., Cherepanov, D., & Fryback, D. G. (2011). Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Medical Decision Making*, *31*, 260–269. https://doi.org/10.1177/0272989X10380925.

Parker, H., Hunt, E. T., Brazendale, K., von Klinggraeff, L., Jones, A., Burkart, S., et al. (2023). Accuracy and precision of opportunistic measures of body composition from the Tanita DC-430U. *Childhood Obesity*, *19*, 470–478. https://doi.org/10.1089/chi.2022.0084.

Patel, K., El-Khoury, J. M., Aarsand, A. K., Badrick, T., Jones, G. R. D., Sikaris, K., et al. (2021). Current utility and reliability of biological variability. *Clin Chem*, *67*, 1050–1055. https://doi.org/10.1093/clinchem/hvab055.

Pazira, P., Rostami Haji-Abadi, M., Zolaktaf, V., Sabahi, M., & Pazira, T. (2016). The better way to determine the validity, reliability, objectivity and accuracy of measuring devices. *Work.*, *54*, 495–505. https://doi.org/10.3233/WOR-162310.

Pearson, K. V. I. I. I. (1901). Mathematical contributions to the theory evolution.—IX. On principle of homotyposis and its relation, the variability of the individual, and to that of the race. Part I.—Homotyposis in the vegetable Kingdom. *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*, *197*, 285–379. https://doi.org/10.1098/rsta.1901.0020.

Ploutz-Snyder, L. L., & Giamis, E. L. (2001). Orientation and familiarization to 1 RM strength testing in old and young women. *J Strength Cond Res*, *15*, 519–523.

Powden, C. J., Hoch, J. M., & Hoch, M. C. (2015). Reliability and minimal detectable change of the weight-bearing lunge test: a systematic review. *Man Ther.*, *20*, 524–532. https://doi.org/10.1016/j.math.2015.01.004.

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: measures of agreement. *Perspect Clin Res*, *8*, 187. https://doi.org/10.4103/picr.PICR_123_17.

Reimers, S., & Maylor, E. A. (2005). Gender effects on reaction time variability and trial-to-trial performance: reply to deary and der. *Aging, Neuropsychology, and Cognition*, *13*, 479–489. https://doi.org/10.1080/138255890969375.

Schmidt, H., Bashkuev, M., Weerts, J., Graichen, F., Altenscheidt, J., Maier, C., et al. (2018). How do we stand? Variations during repeated standing phases of asymptomatic subjects and low back pain patients. *J Biomech*, *70*, 67–76. https://doi.org/10.1016/j.jbiomech.2017.06.016.

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. *Phys Ther*. https://doi.org/10.1093/ptj/pzac068.

Srinivasan, P., Westover, M. B., & Bianchi, M. T. (2012). Propagation of uncertainty in Bayesian diagnostic test interpretation. *South Med J.*, *105*, 452–459. https://doi.org/10.1097/SMJ.0b013e3182621a2c.

de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *J Clin Epidemiol*, *59*, 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine*. Cambridge: University Press. https://doi.org/10.1017/CBO9780511996214.

Vetter, T. R., & Schober, P. (2018). Agreement analysis: what he said, she said versus you said. *Anesth Analg*, *126*, 2123–2128. https://doi.org/10.1213/ANE.0000000000002924.

Warneke, K., Keiner, M., Lohman, L. H., Brinkmann, A., Hein, A., Schiemann, S., et al. (2022). Critical evaluation of commonly used methods to determine the concordance between sonography and magnetic resonance imaging: a comparative study. *Front Imaging*. https://doi.org/10.3389/fimag.2022.1039721.

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

Warneke, K., Siegel, S. D., Drabow, J., Lohmann, L. H., Jochum, D., Freitas, S. R., et al. (2025b). Examiner experience moderates reliability of human lower extremity muscle ultrasound measurement—a double blinded measurement error study. *Ultrasound J*, *17*, 20. https://doi.org/10.1186/s13089-025-00424-6.

Warneke, K., Meder, J., Plöschberger, G., Oraze, M., Zechner, M., Jochum, D., et al. (2025c). Can measurement errors explain variance in the relationship between stiffness and range of motion? A blinded reliability and objectivity study. *Eur J Appl Physiol.*, , .

Warneke, K., Oraze, M., Plöschberger, G., Herbsleb, M., Afonso, J., & Wallot, S. (2025d). When testing becomes learning—underscoring the relevance of habituation to improve internal validity of common neurocognitive tests. *European Journal of Neuroscience*, *61*, e70117.

Weir, J. P. (2005). Quantifying test-retest reliability using the Intraclass correlation coefficient and the SEM. *The Journal of Strength and Conditioning Research*, *19*, 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., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*, *30*, 79–82. https://doi.org/10.3354/cr030079.

Zöger, M., Nimmerichter, A., Baca, A., & Wirth, K. (2025). Reproducibility of peak force for isometric and isokinetic multi-joint leg extension exercise. *BMC Sports Sci Med Rehabil*, *17*, 14. https://doi.org/10.1186/s13102-025-01061-1.

Koch V, Wilke J. Reliability of a New Indentometer Device for Measuring Myofascial Tissue Stiffness. J Clin Med. 2022 Sep 1;11(17):5194. doi: 10.3390/jcm11175194. PMID: 36079124; PMCID: PMC9457058.