Accepted Manuscript

Title: Error measurement in craniometrics: The comparative performance of four popular assessment methods using 2000 simulated cranial length datasets (g-op)

Authors: Hayley S.M. Fancourt, Carl N. Stephan

PII: S0379-0738(18)30060-4
DOI: https://doi.org/10.1016/j.forsciint.2018.02.008
Reference: FSI 9165

To appear in: *FSI*

Please cite this article as: Hayley S.M.Fancourt, Carl N.Stephan, Error measurement in craniometrics: The comparative performance of four popular assessment methods using 2000 simulated cranial length datasets (g-op), Forensic Science International https://doi.org/10.1016/j.forsciint.2018.02.008

Error Measurement in Craniometrics: The Comparative Performance of Four Popular Assessment Methods Using 2000 Simulated Cranial Length Datasets (g-op)*

Hayley S. M. Fancourt [a], Carl N. Stephan [a]

[a] Laboratory for Human Craniofacial and Skeletal Identification (HuCS-ID Lab), School of Biomedical Sciences, The University of Queensland, Brisbane, Australia, 4072.

* Portions of this work have been presented at the 17th Biennial Meeting of the International Association of Craniofacial Identification July 2017, in Brisbane, Australia.

Running head: Error in Craniometrics

Corresponding author:

Hayley Fancourt

Laboratory for Human Craniofacial and Skeletal Identification

School of Biomedical Sciences, The University of Queensland, Brisbane, Australia, 4072

Email: hayley.fancourt@uq.net.au

Highlights

- Four error metrics were tested across 2000 simulations of repeat measurements (g-op)
- With large error, intraclass (ICC) and Pearson's correlation coefficients (r) were high
- t-tests resulted in insignificant P-values when errors were normally-distributed
- Technical error of measurement (TEM) intuitively increases with increasing error
- TEM is the preferred error metric in contrast to r, ICC and P-value result

Abstract

For measurements to be accurate and precise, measurement errors should be small. In the anthropometry and craniofacial identification literature, four methods are commonly used for assessing measurement error: Pearson's product moment correlation coefficient (r), intra-class correlation coefficients (ICC), statistical significance tests (often reported by P-values) and the technical error of measurement (TEM; also known as Dalberg's error/ratio). In this paper, the performance of all four of these statistics were evaluated using maximum cranial lengths (g-op) from Howells (n = 2524), by duplicating the dataset and mathematically adding known degrees of error to the second set. This was repeated under a broad array of trials (2000 total) each with slightly different amounts of error simulation to comprehensively assess the four error metrics in terms of descriptive power and utility, using the same data for each of the four error assessment methods. Data simulations included the addition of random and systematic errors of different sizes with absolute differences ranging from 1–50 mm (or in relative terms, 28 % of the original measurement). Two sample sizes (n = 25 and 2524 individuals) were explored and all analyses were conducted in R. P-values from Student's t tests only showed significant differences ($P < 0.05$) for the larger sample size when the error was systematic. Small samples, and/or any with random error, did not yield low or significant P-values ($P < 0.05$). When raw differences were < 4 mm for 95 % of the sample (n = 2524), the ICC and r were high (> 0.97) and remained so even after tripling the error, such that 95 % of the sample possessed raw differences up to 12 mm (r = 0.8). In contrast, the TEM was low initially (< 2 mm or r-TEM < 1 %), and then increased (< 4.5 mm and 2.5 % corresponding to TEM and r-TEM respectively). These data show that P-values, ICC and r values hold substantial limits for error description as they do not always flag error well. In contrast, TEM appears to covary with error more saliently and holds the advantage that changes are reported in the units of the

original measurement. For these reasons, TEM is recommended in favour to P-values, ICC and r.

Keywords: Forensic science; Error statistics; Measurement error; Intraobserver error; Inter-observer error; Anthropometry

Introduction

The term 'measurement error' is used to describe inaccuracies in measurement that may be manifested by differences between repeated measurements of the same object. These differences can communicate important information of the data accuracy (correctness) and precision (tightness of cluster addressing reliability/repeatability/reproducibility depending on the context). While the standardization of data collection procedures is vital to ensure correctness and consistency of measurement [1-4], measurement error is a fact of real world practice that is essentially impossible to eliminate [5]. Thereby, any measurement can be considered as the sum of the measurement's true value and the error component [6].

Before proceeding it is worth defining basic terms used to describe error in a dataset. Accuracy refers to how correct the data are. Precision refers to consistency, and is also called reliability. Note here that accurate data may not be precise and precise data may not be accurate. Repeatability is the similarity of repeat measurements taken by the same investigator, i.e., the intra-observer error [6]. If an investigator does not deviate from a particular measurement method, it is important to note that intra-observer error can still result from instrument error and/or a change in the investigator's measurement technique, i.e., the style in which a specific measurement instruction is executed [7]. Reproducibility is similar to repeatability, but refers

to measurements taken by different investigators and thereby describes the inter-observer error [6].

Prime sources of error in anthropometry and craniometrics not only include ability to accurately and reliably locate skeletal landmarks [1-4], but also the ability to correctly read and record measurements taken from instruments [7]. Here factors such as proper technique, specimen stands, and well-lit and comfortable laboratory conditions can be vital. In terms of measurements, it is important to note that some measurements/landmarks may be metrically or instrumentally determined, not anatomically defined [4, 8, 9].

As part of anthropometry research practice or forensic casework, it is crucial for the measurement error to be determined and explicitly documented to communicate the trustworthiness of the data [10-12]. The statistics used to describe the measurement error should thereby be carefully chosen so that they clearly communicate the amount of error present in an easy to appreciate manner. Error statistics that cryptically encode the data or encourage misleading interpretations should be avoided. Four commonly used statistics to report error in anthropometry and craniofacial identification are: Pearson's product-moment correlation coefficients (r), P-value results from statistical significance tests (commonly Student's t-test), intraclass correlation coefficients (ICC), and technical error of measurement (TEM) [11, 13-16].

Pearson's Product-Moment Correlation Coefficient

Pearson's product-moment correlation coefficient, *r*, is a measurement of the linear relationship between two variables [13, 17]. The calculation for r in a sample of *n* paired

observations is outlined below, where x_i and y_i are paired measurements indexed by i, and $\bar{x}$ and $\bar{y}$ are the sample mean [18]:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}}$$

The r ranges from -1.00 to 1.00, with an r value of 1.00 indicating a perfect positive linear relationship, -1.00 a perfect negative linear relationship, and 0.00 indicating no relationship [17]. The main shortcoming of r, lies in its assessment of the consistency between samples while disregarding agreement [10, 17, 19, 20]. For instance, r calculates whether two observers/instruments gave values that 'ranked' the sample in the same order (consistency), disregarding whether the values are of the same magnitude (agreement). This makes r unable to differentiate errors between samples where a repeated measurement is consistently higher or lower than its paired value [13, 17] (Fig. 1). This may be identified by scatterplots, however r itself cannot provide the differentiation. Further to this, the choice of dependent variable in calculating r is entirely arbitrary in a repeated measures design, and samples with a large data range are likely to give higher r values than data with a smaller range, even if the residuals about the regression line are similar [10] (Fig. 1).

In reporting r, there appears to be little agreement for an appropriate threshold to differentiate large from small r values, further complicating this metric's dimensionless nature. Atkinson and Nevill [13] suggest when $r > 0.8$ and statistically significant, reliability is high. Goto and Mascie-Taylor [21] state $r > 0.95$ reflect only small errors. While Harris and Smith [10], without specifying any numbers, indicate r is always high unless measurement error is large. While these reports are not necessarily conflicting, it is evident that there is no commonly

agreed interpretative scale for r with regards to measurement error. Also problematic for r is its sensitivity to outliers which may skew r values to provide misleading results.

Intraclass Correlation Coefficient

Intraclass correlation coefficient (ICC) demonstrates the strength of relationship between two variables and was initially developed to help overcome some of the limitations associated with r [16, 19]. To assess reliability, labelled *p*, we can determine the proportion of true variance to total variance within a sample (see below).

$$p = \frac{true\ variance}{true\ variance\ +\ error\ variance}$$

As the true variance of a sample is not known, the ICC estimates this proportion using results from analysis of variance (ANOVA) [19]. Therefore, from McGraw and Wong [22], ICC can be thought of as the proportion of variance attributable to the objects being measured (between-subject). If this proportion is high, the objects themselves can be seen to explain most of the variability in the sample, conversely indicating that error yields relatively little influence. Hence, ICC indicates reliability by splitting variance into between- and within-subject variability and identifying these relative to whole-sample variability [19, 23].

There are several different ICC types, each suitable for specific study designs and research questions (see Weir [24] for an excellent guideline). From Shrout and Fleiss [25] and McGraw and Wong [22], there are ten different ICC types, each differing in fundamental ways. Summarized by Weir [24], ICC type depends on whether:

a) One- or two-way ANOVA should be used;
b) Fixed or random-effects should be modelled (i.e., should the reliability score generalize to a larger group of observers);

c) Systematic error should be included (i.e., whether consistency or agreement should be identified); and

d) Mean or single values be taken as final measurements.

As an example, we can identify the reproducibility of measurements from the same sample taken by two randomly selected analysts. Defining a single measurement to be taken as true, and considering agreement between observers to be important, an ICC that can reasonably generalize reliability to any similarly-skilled analyst is required. This indicates ICC(2,1), given below where, obtained from a two-way ANOVA table, BMS is the between-subject mean square, EMS the mean square of the residual component of within-subject, JMS the mean square of the between-observer component of within-subject, n is the number of subjects and k the number of observers [16].

$$ICC(2,1) = \frac{BMS - EMS}{BMS + (k-1)EMS + \frac{k}{n}(JMS - EMS)}$$

The ICC(2,1) formula can be thought of as the proportion of true variance in the sample (that is, variance only due to differences between the objects measured) to total sample variance, which includes all possible sources of error (that is, both random and systematic errors, and error associated with the random sampling of observers) . Weir [24] highlights that the use of ICC(2,1) is interchangeable for test-retest and inter-observer study designs when assessing agreement (i.e., including systematic error), and thus the ICC(2,1) model will be used in this study.

By using the mean square of partitioned within-subject error (between-observer and residual [error] variance), ICC can identify systematic and random errors, overcoming shortcomings

seen with r. However, ICC is calculated relative to sample heterogeneity, meaning heterogeneous samples will give a higher ICC value simply due to the smaller level of error relative to total variability [13, 19]. Essentially, ICC normalizes measurement error relative to total sample heterogeneity, a weakness for an error metric when used to identify absolute error information [24]. For repeat measures, it is important to recognize that while a low ICC value indicates low reliability, the ICC value may reflect low heterogeneity and/or small sample size not just a large error value [26, 27].

Even further, both calculation and interpretation of ICC values depend on study design and ICC type [16, 27]. The dimensionless nature of ICC makes its interpretation subjective and difficult in similar ways to r [26]. This is reflected in the marked differences seen in reported values for interpreting reliability based on ICC (Fig. 2). For example, in a review of statistical methods used in medical instrument reliability tests, Zaki et al [28] identified that ICC is the most common statistic used, however only 28 % of studies report ICC type and confidence intervals. In this review, most studies were seen to use only one method to assess reliability, and inappropriate application of statistical methods was identified in 19 % of studies [28]. Koo & Li [16] and Müller and Büttner [29] further highlight that different forms of ICC yield different results, even when applied to the same dataset.

Caution should be exercised when interpreting ICC values from previously published studies, or making comparisons between them, and even more so when drawing conclusions in association with ICC results. As with r, ICC is unable to give direct information regarding raw differences between repeat measures, making comparison between studies and interpretation of true measurement error extremely difficult.

Significance Tests

P-values tell the probability of the data under the assumption that they are due to chance. Significance tests are thereby used to identify rare events, typically, by a P-value threshold set at 0.05 [33]. Student's t-tests are used in continuous, normally distributed datasets and are classified as paired in instances of repeat measurements. Paired t-tests compare measurement differences against the null hypothesis of zero (i.e., repeat measurements being the same). This is done by calculating a t-score and comparing against the t-distribution for n-1 degrees of freedom. Then, by assuming measurement differences are zero, a probability of the observed data is reported as a P-value. The t-score for a paired t-test is given by the formula below, where X_d is the average paired difference, S_d is the standard deviation of paired differences and n is the number of pairs [34]. Here, the numerator can be considered to represent signal, and the denominator noise within the sample.

$$t = \frac{\bar{X}_d}{(\frac{S_d}{\sqrt{n}})}$$

The limitations of using t-tests when assessing measurement error are twofold. Firstly, should error be randomly and normally distributed, with a mean equal to the true sample mean, negligible differences between means would exist and thus, t-tests would be unable to identify any errors. Secondly, t-tests tend to exaggerate small differences in large samples, as any difference manifests with low P-values, even extremely small uninformative differences [35-37].

Technical Error of Measurement

Given by the formula below, the technical error of measurement (TEM) is an average of paired measurement differences [10, 15, 38].

$$TEM = \sqrt{\frac{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}{2n}}$$

Here, x is the first and second measurement result for the *i*th subject, and n is the sample size.

TEM is reported in the same units as the original measurements, making interpretation straight forward [15, 38]. As small mean measurement error may have more practical consequences for overall smaller measurements than larger ones, the relative TEM (r-TEM) can also be calculated by dividing TEM by the mean measurement to convert to a percentage value [15]. It is worth noting here that TEM appears to be unaffected by measurement size [38].

Each study, dependent on its design and objective, will have different levels of acceptable error, and thus no common threshold of TEM, or any other metric, can be definitively outlined. However, Perini et al [15] have suggested that for beginner and skilled analysts, acceptable intra-observer r-TEM to be 1.5 and 1 % respectively, with inter-observer r-TEM as 2 and 1.5 %.

In the anthropological literature, and more specifically in the literature concerning skulls as relevant to craniofacial identification, there has been a wide range of error statistics employed and no consensus as to which is ideal. Additionally, there is a large disparity in both the reporting and analysis of measurement error in craniometrics, with some studies reporting only one statistic and others reporting upwards of four [14, 30, 39]. This study aims to explore the utility of all the above-mentioned error metrics and determine which one performs best across multiple simulated error scenarios and thereby should be the statistic of choice for future reporting.

Materials and Methods

Maximum cranial lengths (g-op) used in this study were drawn from Howells' craniometric dataset (mean = 179 mm, interval = 151 to 206 mm, n = 2524) [40]. While any cranial cord length could be used; we chose maximum length as it represents a familiar and basic craniometric measurement relevant to craniofacial identification. It is also one that all physical/biological anthropologists should be familiar.

To examine how the error metrics described above perform in a sample with known error, we duplicated Howells' data to create a second dataset to which known degrees of artificial error could be added. This was done by adjusting the raw data values in specific and purposeful ways, so that they differed to the starting 'ground truth' measurements. The standard error metrics (described above) were then calculated using the two samples, and under a large array of simulated error scenarios (including some with extreme error amounts), to see how the error metrics performed. As a large number of calculations were required across multiple error evaluation methods, including samples of two different sizes (n = 25 and n = 2524), R [41] was employed to automate routines. The smaller sample was specifically set to be on the smaller side for which the t-test was derived (n ≈ 30), and equivalent to small samples commonly used in craniofacial identification practice where constraints on sample accessibility restrict numbers. For example, samples used for calculating craniometrics or facial soft tissue depths in craniofacial identification are routinely in the vicinity of 25 individuals (see e.g., [42-45]).

Random Error

To add random error, a normally distributed set, with a mean of zero and pre-determined standard deviation, was created and added to a duplicate set of Howells' original measurements. These error simulations were run for standard deviations ranging in error at whole point integer values from 1 to 10 mm (representing 1.8 to 18 % of the mean measurement

respectively; Table 1). These simulations represented error randomly spread across all measurements (Fig. 3).

Systematic Error

Firstly, Howells' original measurements were duplicated and ranked in ascending order. This duplicate data was then split into ten different groups (henceforth referred to as quantiles), each containing roughly 10% of the sample (i.e., deciles). To add systematic error, three normally distributed sets were created and each respectively added to the 8^{th}, 9^{th} or 10^{th} quantile of a duplicate set of Howells' original measurements, so that error increased as the g-op measurement increased. These datasets were of increasing mean size and spread, where both the mean and standard deviation increased by whole point integer values between the three generated sets. The error dataset of the smaller mean and spread was added to the 8^{th} quantile, and the error dataset with the largest mean and spread added to the 10^{th} quantile, reflecting increasing error with increasing g-op measurement (Fig. 4). The addition of these three error datasets to the 8^{th}, 9^{th} and 10^{th} quantiles of a duplicate set of Howell's original measurements represented one systematic error simulation. These simulations were then run for average standard deviations of error (i.e., the average of the three added error datasets) from 1 to 10 mm, with the mean of the dataset added to each quantile remaining constant between the different simulations (Table 2).

Error Metrics

One-hundred simulations of each error type (random and systematic) and magnitude (ranging from 1 to 10 mm at every whole point integer value for standard deviation) were generated.

This resulted in 2000 error simulations conducted for each error statistic and since four error metrics were analyzed in each simulation (r, ICC, P-value [t-test], and TEM), a grand total of 8000 error calculations were conducted.

Howells' original data and the artificial repeat samples showed differences in means of < 1 mm, but included raw differences of up to 50 mm in some simulations, corresponding to 28% of mean measurement (Fig. 5), thereby providing an interval of error that more than adequately covered realistic error amounts.

For small sample tests, 25 measurements from Howells' dataset were randomly selected once, with error then added to this sample in the same process as previously described. Again, 100 artificial repeat samples for each error type and magnitude were generated (bootstrapped), and the mean of these 100 replicates taken as the final statistic of interest.

Results

For most error scenarios, many of the statistics explored in this study remained high, misleadingly indicating reliable results, when in fact error was large (Fig. 6).

Random Error

TEM

TEM displays a roughly linear trend for samples with random error. The value of TEM increased as random error increased (positive trend), ranging from 1 to 7 mm. Up to 4 mm of

error randomly added to 95 % of the sample (2 mm standard deviation of error), gave TEM < 2 mm. After tripling this error (up to 12 mm differences in 95 % of sample), TEM increased to > 4 mm. Negligible differences were seen between TEM results from small and large samples.

r

Following a mostly linear trend, r values decreased as added error increased in random error simulations (negative trend). For up to 2 mm error added to 95 % of the sample (1 mm standard deviation of error), r values lie close to 1. With fourfold more error (4 mm standard deviation of error) r decreased to slightly below 0.9, and was approximately 0.5–0.6 for simulations with the largest added error (10 mm standard deviation of error). Simulations from the large sample gave consistently higher r values than those from the smaller sample under the same conditions, with this difference becoming larger as error increased.

ICC

ICC values were approximately 1 for the lowest added error simulations (1 mm standard deviation of error). When added error was tripled, ICC remained above 0.9, and at up to 12 mm added error in 95 % of the sample (6 mm standard deviation of error), ICC values were approximately 0.8 in the large sample and 0.7 in the small sample. As with r, ICC values were consistently higher in the large sample under the same added error conditions; with the largest added error (10 mm standard deviation of error) resulting in an ICC of approximately 0.6 and 0.45 for the large and small sample respectively.

P-values from paired t-tests

All random error simulations resulted in P-values near 0.5, with negligible differences seen between sample sizes.

Systematic Error

TEM

Similar to random error, TEM displays a positive linear trend for simulations of systematic error with negligible differences between sample sizes. Ranging from 1 to 5 mm, TEM increased as the average standard deviation of systematic error increased, although this increase was less pronounced than that seen under random error. Under systematic error, an average standard deviation of 2 mm gave TEM < 2 mm. Tripling this error (6 mm average standard deviation of error), TEM increased to < 4 mm. With the largest systematic error added (10 mm average standard deviation), TEM was approximately 5 mm.

r

Under simulations of systematic error, r values decreased as added error increased, however this decrease was less pronounced than that seen for random error. In systematic error simulations with an average standard deviation of 1 mm, r values were close to 1. Increasing the average standard deviation of error six-fold (i.e., 6 mm) resulted in r values above 0.9. With the largest added error (10 mm average standard deviation of error), r values were approximately 0.8 and above 0.7 for the large and small sample respectively. Simulations from the large sample again gave consistently higher r values than those from the smaller sample under the same conditions, with this difference becoming larger as error increased.

ICC

ICC values decreased as error increased in simulations of systematic error, however this decline was less pronounced than that seen under random error. ICC values were close to 1 for systematic error with an average standard deviation of 1 mm, and remained above 0.9 when error was increased fourfold (4 mm average standard deviation of error). In large samples, ICC values remained higher than respective small sample ICC values, with this difference increasing as error increases. For the largest added systematic error (10 mm average standard deviation), the large and small sample showed ICC values of approximately 0.75 and 0.65 respectively.

P-values from paired t-tests

Systematic error in large samples always resulted in P-values < 0.05, while only 1 mm average standard deviation of systematic error gave a P-value < 0.05 in the smaller sample. P-values for the smaller sample then increased as error increased, finishing around 0.5 for the largest average standard deviation of systematic error (10 mm).

Discussion

TEM

Under the simulations in this study, the only error metric that showed intuitive trends across all error scenarios and displayed negligible differences between sample sizes and error types, was TEM. Only error of up to 2 and 4 mm added to 67 % and 95 % of the sample respectively (1.1 % and 2.2 % of mean measurement), resulted in TEM < 2 mm. Although TEM cannot indicate the directionality or frequency of error, it gives a clear indication of the error magnitude because it is in the same units as the measurement itself, and the statistic moves in the same direction as the error (i.e., TEM increases as error increases). This latter aspect does not apply to all error metrics examined here.

For small error magnitudes, TEM was almost proportional to the error magnitude; that is, standard deviation of error of 1 and 2 mm gave an approximate TEM of 1 and 2 respectively. While TEM values are lesser for scenarios of larger error, this proportionality in instances of lower error, as expected given the calculation, is extremely useful for direct interpretation of raw differences. Further, by simply reporting the most likely raw difference between two repeat measurements for a given investigation, TEM remains unaffected by differences in size and/or spread between different samples, making it suited to direct comparison between studies. Overall, TEM is easy to interpret, directly related to raw error magnitude in a repeat sample, intuitively increases as error increases, is unaffected by differences in sample size, and can easily be compared between studies. From this, TEM overcomes many of the limitations seen in the other error metrics investigated in this study as described below.

ICC and r

While ICC overcomes some of the shortcomings of r in capturing systematic error, both values remained high when error was substantial (> 3 mm standard deviation of error). An ICC > 0.8 represented data with up to 10 mm of error added to 95% of the sample (standard deviation 5 mm). From Jamaiyah et al [30], Koo and Li [16] and Rosner [32], this indicates that these samples have almost perfect, good/excellent, or perfect reliability respectively. In this study, 10 mm of error corresponds to 5.6 % of the mean measurement - almost certainly too high to be considered a reliable value in most scientific or forensic contexts of anthropometry.

There is also clear disparity between ICC values from samples of different sizes, indicating error magnitude may be concealed in larger samples and thus may require more stringent criteria than in smaller samples. Further to this, as with most craniometrics studies, the data

used was highly heterogeneous, which may have contributed to ICC values remaining high despite substantial added error.

With ICC and r values decreasing as raw error increases, the interpretation of these statistics becomes somewhat counterintuitive, and gives no directly relatable information indicating the size of raw error in a repeat sample (for example, for a large sample with 1 mm standard deviation of random error ICC is approximately 1, but at 3 times this error ICC decreases to only 0.95). Given ICC and r values are difficult to interpret and fundamentally rely on the variance of the sample from which they are calculated, these metrics should be cautiously approached when making comparisons between studies.

P-values from paired t-tests

As a statistic that is widely known to be frequently misinterpreted [37, 46, 47], the trends seen in P-values from the paired t-tests in this study are particularly interesting. As artificially added error was normally distributed with a mean of zero for random error, the mean of differences between repeat measurements was effectively zero. Because of this, no matter how varied the differences between repeat measurements, signal in the dataset (t-statistic numerator) would still be considered zero as the t-statistic formula only employs a mean paired difference. In every case, this leads to a t-statistic of zero, and thus a statistically insignificant P-value (> 0.05) indicating no difference between datasets, despite practically significant raw differences of up to 50 mm in some repeat simulations. From this, it is even more apparent that paired t-tests are fundamentally unable to identify instances of random error, at any magnitude, when the means of paired differences are zero. This is particularly concerning when calculating intra-observer error, as, by design, error from one observer who uses the same measurement

method and tools is likely to be normally distributed with a similar mean between repeat sets and raises red flags for the use of t-tests for error assessment.

For systematic error however, the means of repeat datasets change slightly due to the deviation of added error from a mean of zero in the largest 30 % of data. In the large sample, the signal to noise ratio of the t-statistic shows a significant difference between repeat datasets at all error magnitudes, likely due to the low denominator (noise) value given when n is large. This leads to a large t-value, resulting in a low P. In the smaller sample however, only systematic error with an average standard deviation of 1 mm resulted in $P < 0.05$. This is likely because less noise (low average standard deviation of error) permits small differences between dataset means (means of added error datasets: 1, 2, 3 in three uppermost quantiles) to yield large t-values. While $P < 0.05$ for 1 mm average standard deviation of error shows statistical significance, here little practical significance may exist, particularly relative to the larger error scenarios, and the result is at risk of being wrongly interpreted. In application, higher error magnitudes commonly have a larger impact on study outcomes, and thus the lack of significant P-values in the smaller sample for large average standard deviations of systematic error is not ideal.

It should be noted here that t-tests are likely to be more effective when addressing error that involves constant differences between measurements (e.g., where a repeat measurement is always 20 mm above a previous measurement); however, the true presence of such an effect would likely be indeterminable from paired t-tests if undertaken with a large sample size. With this, paired t-tests cannot, by design, identify random error (with a mean of zero) of any magnitude and are likely to show statistical significance in all large samples. These are

shortcomings of this error metric, and such limitations should be well understood by any practitioner wishing to apply or interpret P-values as an indicator of error embedded in samples.

General

This study investigated the performance of a several commonly used error metrics in instances of known measurement error using, as an example, maximal cranial length. While the rational for the use of maximum cranial length measurements was based on popularity of cranial length as a cornerstone craniometric, in practice any linear dimension of similar size could have been used and comparable results would have been obtained. A pronounced range of error was added to the data in a controlled manner to highlight the trends of these error metrics under both practically small and impractically large errors (e.g., up to 50 mm in some cases). While some of these values may be unrealistic in practice, it is important to consider these trends, to appreciate the error metric performance across the full range of error magnitudes.

There are many other statistics, different to those investigated in this study, that can be employed to assess reliability. For example, Cohen's kappa can be used to identify repeatability and reproducibility by identifying the percent of measurement agreement, accounting for chance [48-50]. Kappa is most commonly used to identify inter-observer reliability, but it is worth noting that this metric has limitations due to its dimensionless nature and subjective interpretation much the same as those seen with r and ICC [48, 49]. Further to this, kappa assesses percent agreement by identifying the number of times observers did not agree on a given measurement, but gives no weight to the amount by which they disagree [48-50]. Weighted kappa has been developed to help overcome this issue, however both kappa and weighted kappa are more suited to qualitative/ordinal data than continuous data [49, 50]—thus the reason why Kappa was not evaluated in this study of continuous data for cranial lengths.

Note here that Fleiss and Cohen [50] highlight that in some situations, particularly those identifying systematic error such as ICC(2,1), weighted kappa and intraclass correlation coefficient are equivalent.

All the error metrics explored herein aim to conveniently define one value indicative of reliability. While such an overarching value is important, the frequency and direction of error are additional important factors to consider [12]. Almost all error metrics explored in this study have an average calculation in their formula, and while helpful for estimating error size and to some extent frequency, such calculations may be sensitive to outliers. Therefore, it is always an important step to visualize error within a repeat sample by plotting as per approaches of exploratory data analysis [51]. Bland-Altman plots are common for such visualization [52-54], and are highly recommended as a helpful addition to error assessment.

It is also worth noting here, that with regards to r values, Spearman [55] developed a post-hoc correction that attempts to account for attenuation as a result of the measurement error. This results in strengthening (upwards revision) of all raw r-values [56] and means that the r-value becomes an estimate (with associated confidence intervals) [57]. Whether this approach should be employed is a matter of some controversy [56-58] as point estimates of r-values become unreliable with small samples and large errors such that the disattenuated values may not offer much improvement or, even worse, may be misleading. In the end, there is ultimately no substitute for high-quality error free data and should defective sets be encountered, then investigators should seriously consider whether they should be disregarded.

From the common statistics and their associated limitations highlighted in this study, practitioners should be aware that high r values, ICCs or non-significant P-values do not

necessarily mean 'tight' data with little error. For direct error information and comparability between studies, a single TEM, or r-TEM, value is very useful. In some circumstances, a combination of measurement error values may be useful or necessary, e.g., correlations between repeated measurements will be required if calculating disattenuated r-values, however, TEM should be reported as a standard statistic in any analysis given its more intuitive covariation with error and documentation of the error in the units of the measurements subject to investigation.

Conclusion

In the simulations run here, and compared to other r, ICC and paired t-tests evaluated, the TEM is the most generically useful error indicator under differing sample sizes and error types. The t-tests are unable to identify most normally distributed raw error, while both ICC and r remain high when raw error is substantial. In addition, ICC, r, and P-values from t-tests are both difficult to interpret and difficult to compare between different investigations. Comparatively, TEM intuitively increases as error increases, is easily understood and interpreted and can be directly compared between studies for measurements collected in the same units (e.g., mm for craniometrics). For tests conducted herein, TEM was the most robust and therefore useful statistic for assessing measurement error in craniometrics. We suggest it should be reported as a standard for error assessment in craniofacial identification, including facial soft tissue thickness values, and if not reported in favour of P-values, ICC and r, it should at least accompany these other statistics.

Acknowledgements

The authors would like to acknowledge Maciej Henneberg for initial supervisory training on anthropometry methods.

References

[1] J. Caple, C.N. Stephan, A standardized nomenclature for craniofacial and facial anthropometry, Int J Legal Med 130 (2016) 863–879.

[2] P.H. Moore-Jansen, S.D. Ousley, R.L. Jantz, Data collection procedures for forensic skeletal material, 3 ed., University of Tennessee, Knoxville, TN, 1994.

[3] A. Hrdlička, The anthropometric committee of the American Association of Physical Anthropologists, Am J Phys Anthropol 21 (1936) 287–300.

[4] G. Olivier, Practical Anthropology, Charles C Thomas, Springfield, IL, 1969.

[5] K. Krishan, T. Kanchan, Measurement error in anthropometric studies and its significance in forensic casework, Ann Med Health Sci Res 6 (2016) 62-63.

[6] S.B. Vardeman, J.M. Jobe, Statistics and measurement, Statistical methods for quality assurance: basics, measurement, control, capability, and improvement, Springer, New York, NY, 2016, pp. 33-105.

[7] J.A. Gavan, The consistency of anthropometric measurements, Am J Phys Anthropol 8 (1950) 417-426.

[8] H.H. Wilder, A laboratory manual of anthropometry, P. Blakiston's Son & Co., Philadelphia, 1920.

[9] R. Martin, Lehrbuch der Anthropologie in systematischer Darstellung: mit besonderer Berücksichtigung der anthropologischen Methoden für Studierende, Ärzte und Forschungsreisende. Gustav Fischer, Jena, Gustav Fischer, Jena, 1928.

[10] E.F. Harris, R.N. Smith, Accounting for measurement error: a critical but often overlooked process, Arch Oral Biol 54 Suppl 1 (2009) S107-17.

[11] L. Corron, F. Marchal, S. Condemi, K. Chaumoitre, P. Adalian, Evaluating the consistency, repeatability, and reproducibility of osteometric data on dry bone surfaces,

scanned dry bone surfaces, and scanned bone surfaces obtained from living individuals, B Mem Soc Anthro Par 29 (2017) 33-53.

[12] C.J. Utermohle, S.L. Zegura, G.M. Heathcote, Multiple observers, humidity, and choice of precision statistics: Factors influencing craniometric data quality, Am J Phys Anthropol 61 (1983) 85-95.

[13] G. Atkinson, A.M. Nevill, Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine, Sports Med 26 (1998) 217-238.

[14] S. Stomfai, W. Ahrens, K. Bammann, E. Kovacs, S. Marild, N. Michels, L.A. Moreno, H. Pohlabeln, A. Siani, M. Tornaritis, T. Veidebaum, D. Molnar, I. Consortium, Intra- and inter-observer reliability in anthropometric measurements in children, Int J Obes (Lond) 35 Suppl 1 (2011) S45-51.

[15] T.A. Perini, G.L. Oliveira, J.S. Ornellas, F.P. Oliveira, Technical error of measurement in anthropometry, Rev Bras Med Esporte 11 (2005) 86-90.

[16] T.K. Koo, M.Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J Chiropr Med 15 (2016) 155-63.

[17] M.-T. Puth, M. Neuhäuser, G.D. Ruxton, Effective use of Pearson's product–moment correlation coefficient, Animal Behaviour 93 (2014) 183-189.

[18] C. Clapham, J. Nicholson, Pearson's product moment correlation coefficient, The Concise Oxford Dictionary of Mathematics, Oxford University Press, 2009.

[19] A. Bruton, J.H. Conway, S.T. Holgate, Reliability: What is it and how is it measured?, Physiotherapy (2000).

[20] J.M. Bland, Statistics notes: Measurement error and correlation coefficients, BMJ 313 (2017) 41.

[21] R. Goto, C.G.N. Mascie-Taylor, Precision of measurement as a component of human variation, J Phys Anthropol 26 (2007) 253-256.

[22] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients, Psychol Methods 1 (1996) 30-46.

[23] H.Y. Kim, Statistical notes for clinical researchers: evaluation of measurement error 1: using intraclass correlation coefficients, Restor Dent Endod, 2013, pp. 98-102.

[24] J.P. Weir, Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM, J Strength Cond Res 19 (2005) 231-240.

[25] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol Bull 86 (1979) 420-428.

[26] J.W. Bartlett, C. Frost, Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables, Ultrasound Obstet Gynecol 31 (2008) 466-475.

[27] K.M. Lee, J. Lee, C.Y. Chung, S. Ahn, K.H. Sung, T.W. Kim, H.J. Lee, M.S. Park, Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research, Clin Orthop Surg 4 (2012) 149-55.

[28] R. Zaki, A. Bulgiba, N. Nordin, N.A. Ismail, A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables, Iran J Basic Med Sci 16 (2013) 803-807.

[29] R. Müller, P. Büttner, A critical discussion of intraclass correlation coefficients, Stat Med 13 (1994) 2465-2476.

[30] H. Jamaiyah, A. Geeta, M.N. Safiza, G.L. Khor, N.F. Wong, C.C. Kee, R. Rahmah, A.Z. Ahmad, S. Suzana, W.S. Chen, M. Rajaah, B. Adam, Reliability, technical error of measurement and validity of length and weight measurement for children under two years old in Malaysia, Med J Malaysia 65 (2010) 131-137.

[31] S. Chinn, Repeatability and method comparison, Thorax 46 (1991) 454-456.

[32] B. Rosner, Fundamentals of biostatistics, 7 ed., Brooks/Cole Cengage Learning, Boston, MA, 2011.

[33] R.A. Fisher, Statistical methods for research workers, 5 ed., Oliver and Boyd, Edinburgh, UK, 1934.

[34] S. Daya, Paired t-test, Evidence-based Obstet Gynecol 5 (2003) 105-106.

[35] C.N. Stephan, L. Munn, J. Caple, Facial soft tissue thicknesses: noise, signal, and p, Forensic Sci Int 257 (2015) 114-122.

[36] J. Cohen, Things I have learned (so far), American Psychologist 45 (1990) 1304-1312.

[37] R.L. Wasserstein, N.A. Nicole A. Lazar, The ASA's statement on p-values: context, process, and purpose, The American Statistician 70 (2016) 129-133.

[38] P.L. Jamison, R.E. Ward, Measurement size, precision, and reliability in craniofacial anthropometry - bigger is better, Am J Phys Anthropol 90 (1993) 495-500.

[39] A.H. Richard, C.L. Parks, K.L. Monson, Accuracy of standard craniometric measurements using multiple data formats, Forensic Sci Int 242 (2014) 177-85.

[40] B.M. Auerbach, Howells' craniometric data set. <https://web.utk.edu/~auerbach/HOWL.htm>, 2014 (accessed August 23.2017).

[41] R.C. Team, R: A language and environment for statistical computing., R Foundation for Statistical Computing, Vienna, Austria., 2017.

[42] E. Simpson, M. Henneberg, Variation in soft-tissue thicknesses on the human face and their relation to craniometric dimensions, Am J Phys Anthropol 118 (2002) 121-33.

[43] M. Domaracki, C.N. Stephan, Facial soft tissue thicknesses in Australian adult cadavers, J Forensic Sci 51 (2006) 5-10.

[44] C.N. Stephan, E.K. Simpson, Facial soft tissue depths in craniofacial identification (part I): an analytical review of the published adult data, J Forensic Sci 53 (2008) 1257-1272.

[45] S. De Greef, P. Claes, D. Vandermeulen, W. Mollemans, P. Suetens, G. Willems, Large-scale in-vivo Caucasian facial soft tissue thickness database for craniofacial reconstruction, Forensic Sci Int 159 Suppl 1 (2006) S126-46.

[46] M.J. Lew, Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P, Br J Pharmacol 166 (2012) 1559-1567.

[47] D.J. Biau, B.M. Jolles, R. Porcher, P value and the theory of hypothesis testing: an explanation for new researchers, Clin Orthop Relat Res 468 (2010) 885-892.

[48] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem Med 22 (2012) 276-82.

[49] J.N. Mandrekar, Measures of interrater agreement, J Thorac Oncol 6 (2011) 6-7.

[50] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educ Psychol Meas 33 (1973) 613-19.

[51] J.W. Tukey, Exploratory data analysis, Addison-Wesley, Reading, 1977.

[52] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307-10.

[53] J.M. Bland, D.G. Altman, Measuring agreement in method comparison studies, Stat Methods Med Res 8 (1999) 135-60.

[54] D.G. Altman, J.M. Bland, Measurement in medicine: the analysis of method comparison studies, The Statistician 32 (1983) 307-17.

[55] C. Spearman, The proof and measurement of assocation between two things, Am J Psychol 15 (1904) 72-101.

[56] P.M. Muchinsky, The correction for attenuation, Educ Psychol Meas 56 (1996) 63-75.

[57] E.P. Charles, The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets, Psychol Methods 10 (2005) 206-226.

[58] D.W. Zimmerman, R.H. Williams, Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions, Appli Psychol Meas 21 (1997) 253-270.

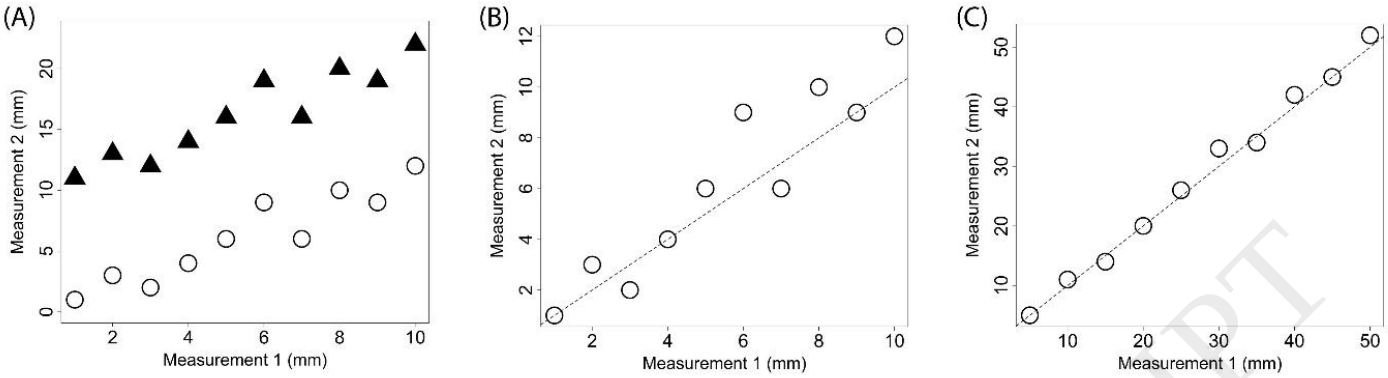

Figure 1: Some limitations of r for error measurement: (A) r cannot differentiate differences of error magnitude between sets of repeated measurements – both datasets illustrated in (A) hold the same r value (0.941), even though the data illustrated by closed triangles for Measurement 2 are vastly different to that of Measurement 1. (B/C) Illustrate that closeness of data to the regression line (small residuals) is not the only information influencing r. (B) the same data represented by circles in (A) with the slope y = x plotted (dotted line) has r = 0.941; and (C) new data with the same raw differences about the line y = x (shown as a dotted line) as Graph B, over a larger range of measurements; in (C) r = 0.997. To force (C) to hold the same r value for data as (B), the residuals from the regression line would need to be magnified by a factor of 5.

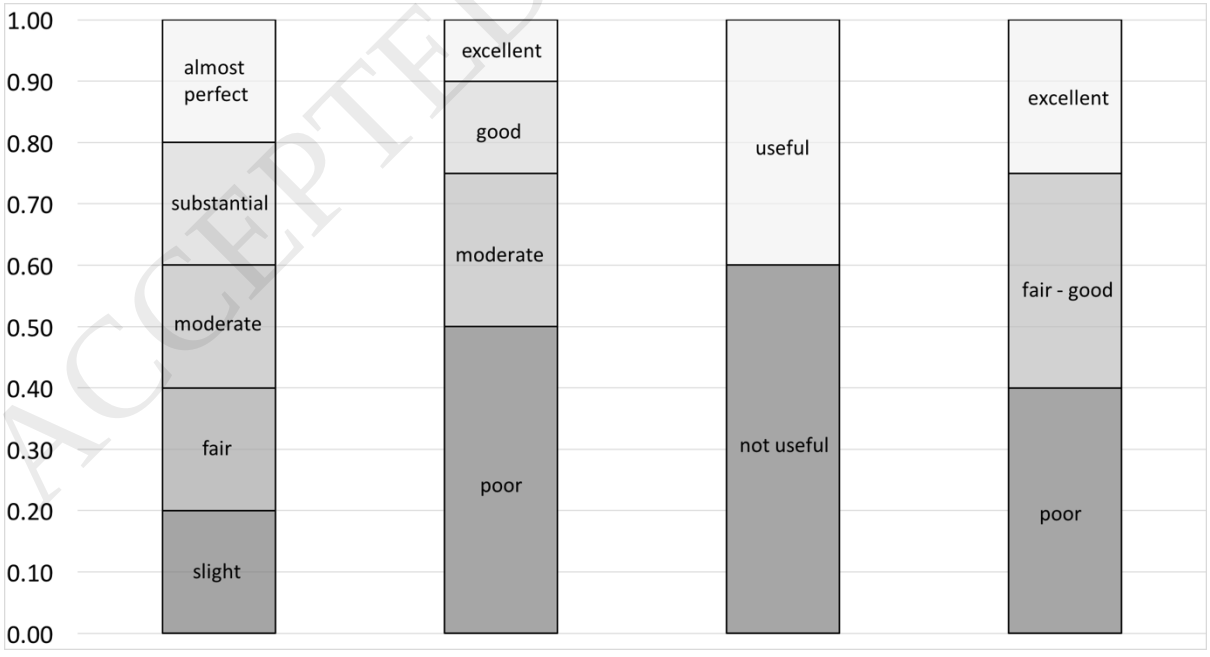

Figure 2: Reported ICC thresholds and their associated reliability interpretations. From left to right, ICC interpretations are from Jamaiyah et al [30], Koo & Li [16], Chinn [31] and Rosner [32].

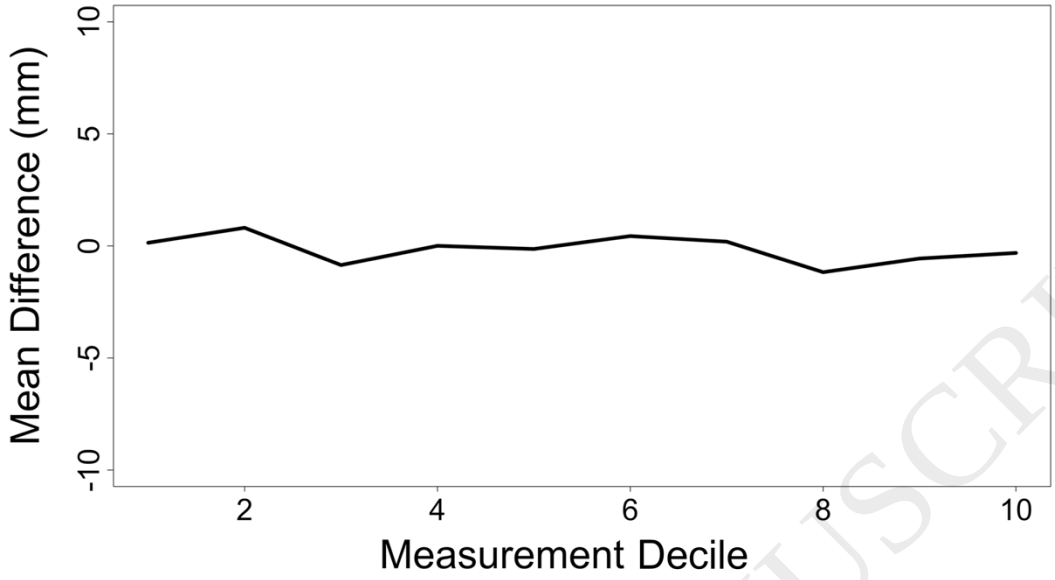

Figure 3: Mean differences in each decile of Howells' original measurements for the random error set. Differences represent raw differences between Howells' dataset and a duplicated set with random error added using a standard deviation of 10 mm. The results of one simulation are shown.

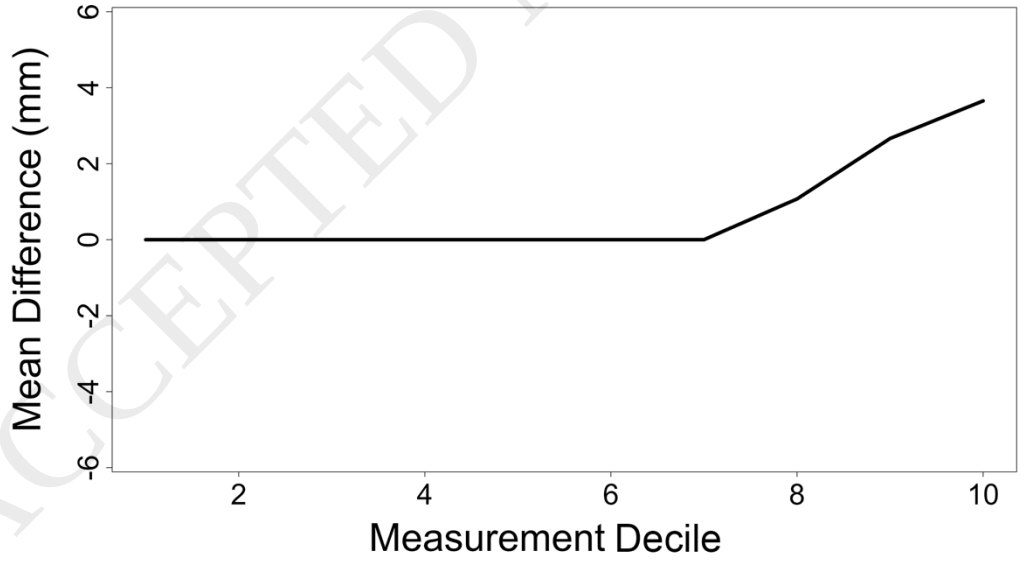

Figure 4: Mean differences in each decile of Howells' original measurements for the systematic error set. Differences represent raw differences between Howells' dataset and a duplicated set with systematic error added using an average standard deviation of 10 mm. The results of one simulation are shown.

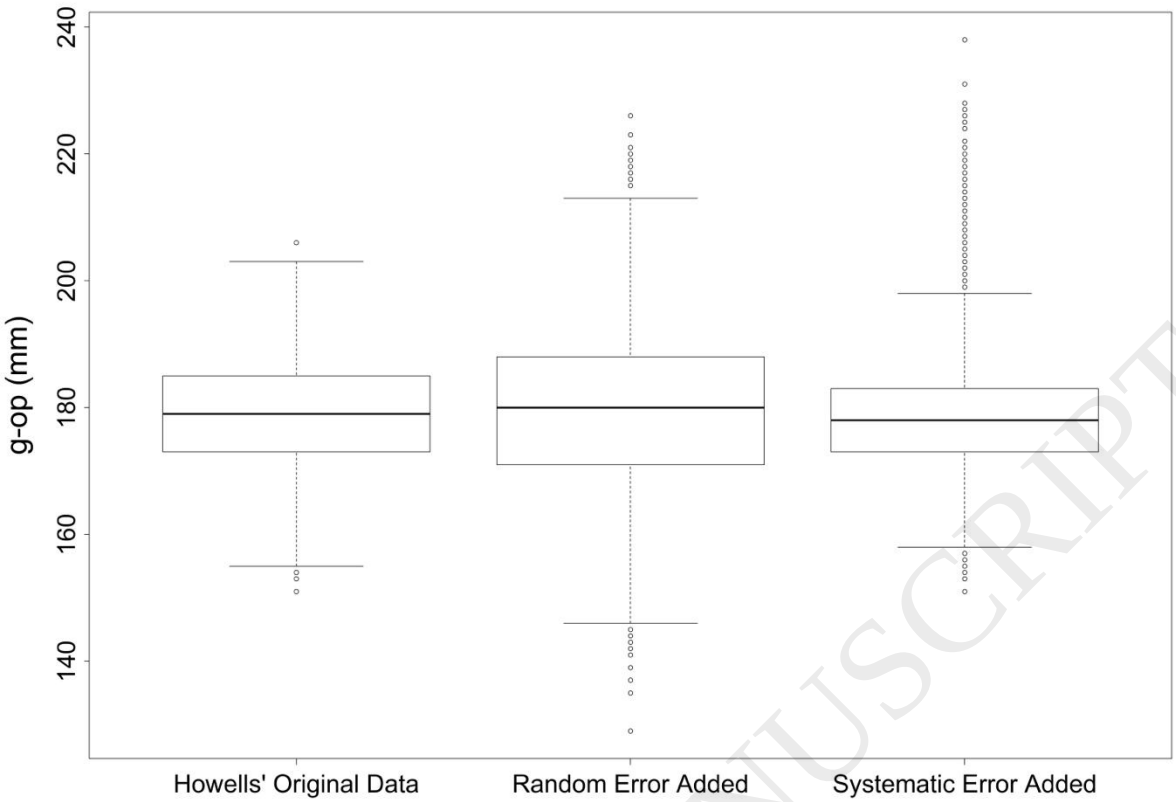

Figure 5: Boxplot of true measures and repeat samples with added error. Random error refers to normally distributed (mean: 0, standard deviation: 10mm) error added to true measures, while systematic error refers to normally distributed (mean: 1, 2, 3; standard deviation: 9, 10, 11, respectively) error added to the three upper quantiles of true measurements. Boxplot represents one simulation of added error, note here that for the study such protocols were run 100 times to generate 100 error simulations for each error type.

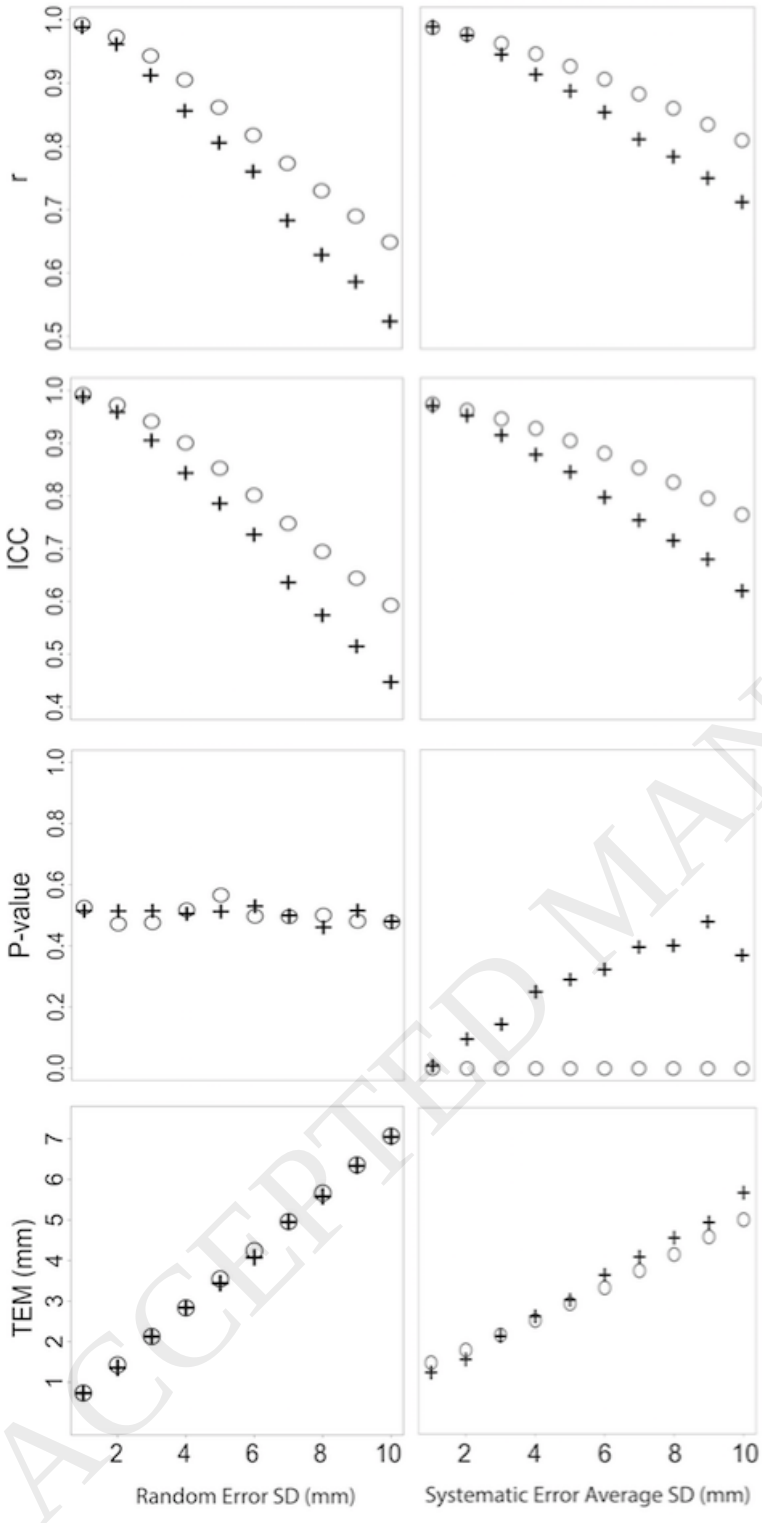

Figure 6: Performance of error metrics under random and systematic error. Unfilled circles (○) indicate the large sample (N=2524) and pluses (+) indicate the smaller sample (n=25). Columns indicate the application of random or systematic error, with units indicating the standard deviation (SD) of random error (left), or the average standard deviation of the three datasets added for systematic error (right). From top to bottom, performance of the following error metrics is shown: Pearson's product-moment correlation coefficient (r), intraclass correlation coefficient (ICC), P-value from Student's t-test, and technical error of measurement (TEM).

Table 1: Table of specifications for added random error. These error datasets were each added to a duplicate set of Howells' original g-op measurements to give an artificial repeat sample.

Error Type	**Sample Size**	**Added Error Values**		
		Mean (mm)	**Standard Deviation (mm)**	**Number of simulations**
Random	25	0	1	100
			2	100
			3	100
			4	100
			5	100
			6	100
			7	100
			8	100
			9	100
			10	100
Random	2524	0	1	100
			2	100
			3	100
			4	100
			5	100
			6	100
			7	100
			8	100
			9	100
			10	100

Table 2: Table of specifications for added error datasets generated to simulate systematic error. Each error dataset generated was added to a respective quantile of a duplicate set of Howells' original g-op measurements to represent increasing error with an increase in measurement value. The entire duplicate set of measurements (7 quantiles with no added error, 3 quantiles with error of increasing mean size and spread added to each) represents an artificial repeat sample, with error metrics then run in each of these simulations.

Error Type	**Sample Size**	**Error Adjustment Details**			
		Error added to Quantile	**Mean (mm)**	**Standard Deviation (mm)**	**Number of simulations**
Systematic	25	7	1	0	100
		8	2	1	
		9	3	2	
		7	1	1	100
		8	2	2	
		9	3	3	
		7	1	2	100
		8	2	3	
		9	3	4	
		7	1	3	100
		8	2	4	
		9	3	5	
		7	1	4	100
		8	2	5	
		9	3	6	
		7	1	5	100
		8	2	6	
		9	3	7	
		7	1	6	100
		8	2	7	
		9	3	8	
		7	1	7	100
		8	2	8	
		9	3	9	
		7	1	8	100
		8	2	9	
		9	3	10	
		7	1	9	100
		8	2	10	

		9	3	11	
Systematic	2524	7	1	0	100
		8	2	1	
		9	3	2	
		7	1	1	100
		8	2	2	
		9	3	3	
		7	1	2	100
		8	2	3	
		9	3	4	
		7	1	3	100
		8	2	4	
		9	3	5	
		7	1	4	100
		8	2	5	
		9	3	6	
		7	1	5	100
		8	2	6	
		9	3	7	
		7	1	6	100
		8	2	7	
		9	3	8	
		7	1	7	100
		8	2	8	
		9	3	9	
		7	1	8	100
		8	2	9	
		9	3	10	
		7	1	9	100
		8	2	10	
		9	3	11	