
Assessing inter-observer reliability to support availability of osteometric measurement data from the Olivier collection

*Évaluation de la fiabilité interobservateurs pour soutenir la disponibilité des
données dans les mesures ostéométriques de la collection Olivier*

**Siam Knecht, Sacha Kacki, Yann Ardagna, Aline Thomas, Aurélie Fort,
Florent Détroit, Elle Liagre, Christophe Roman, Pascal Adalian et Sébastien
Villotte**



Édition électronique

URL : <https://journals.openedition.org/bmsap/17682>

DOI : 10.4000/15sl6

ISSN : 1777-5469

Éditeur

Société d'Anthropologie de Paris

Référence électronique

Siam Knecht, Sacha Kacki, Yann Ardagna, Aline Thomas, Aurélie Fort, Florent Détroit, Elle Liagre, Christophe Roman, Pascal Adalian et Sébastien Villotte, « Assessing inter-observer reliability to support availability of osteometric measurement data from the Olivier collection », *Bulletins et mémoires de la Société d'Anthropologie de Paris* [En ligne], 38 (1) | 2026, mis en ligne le 27 février 2026, consulté le 04 mars 2026. URL : <http://journals.openedition.org/bmsap/17682> ; DOI : <https://doi.org/10.4000/15sl6>



Le texte seul est utilisable sous licence CC BY-NC-ND 4.0. Les autres éléments (illustrations, fichiers annexes importés) sont susceptibles d'être soumis à des autorisations d'usage spécifiques.

Assessing inter-observer reliability to support availability of osteometric measurement data from the Olivier collection

Évaluation de la fiabilité interobservateurs pour soutenir la disponibilité des données dans les mesures ostéométriques de la collection Olivier

Siam Knecht ^{1*}, Sacha Kacki ^{2,3}, Yann Ardagna ¹, Aline Thomas ⁴, Aurélie Fort⁵, Florent Détroit ⁶, Elle Liagre ², Christophe Roman ⁷, Pascal Adalian ¹, Sébastien Villotte ^{4,8,9}

1 Aix Marseille Université, CNRS, EFS, ADES, Marseille, France

2 UMR 5199 PACEA, Université de Bordeaux, CNRS, MC, Pessac, France

3 Department of Archaeology, Durham University, Durham, United Kingdom

4 UMR 7206 Eco-anthropologie (EA), MNHN, CNRS, Université de Paris, musée de l'Homme Paris, France

5 Direction des collections, Muséum national d'histoire naturelle, musée de l'Homme, Paris, France

6 Département Homme & Environnement, MNHN, UMR 7194, CNRS, musée de l'Homme, Paris, France

7 LIS UMR 7020 CNRS, Aix Marseille Université, Marseille, France

8 Quaternary environments & Humans, OD Earth and History of life, Royal Belgian Institute of Natural Sciences, Brussels, Belgium

9 Unité de Recherches Art, Archéologie Patrimoine, Université de Liège, Belgium

* siam.knecht@univ-amu.fr

Reçu : 18 octobre 2025 ; accepté : 13 novembre 2026
Bulletins et Mémoires de la Société d'Anthropologie de Paris

Abstract – The field of anthropometry is crucial in developing methods for reconstructing the biological profile of skeletal remains. Ensuring the accuracy and reliability of osteometric data is therefore essential for methodological validation. This study assesses the repeatability and inter-observer reliability of skeletal measurements collected for the publication of the Olivier Collection. Four observers recorded 42 linear variables (11 cranial and 31 postcranial) on 40 randomly selected skeletons. Inter-observer variation was evaluated through Student's *t*-tests, Cohen's *d*, repeated-measures ANOVAs, intraclass correlation coefficients (ICCs), and both absolute and relative technical errors of measurement (TEM, rTEM). Additionally, a new hierarchical linear model method was used to partition measurement variance into components attributable to subjects, systematic and random measurement error or systematic effects associated with sex.

Significant systematic measurement effects ($p < 0.05$) were identified for 30 of 42 variables, particularly in cranial traits and femur/tibia diameters (explained variance 0.19–0.37). Systematic sex-related effects were detected for 11 variables, suggesting that observer variability may interact with sexual dimorphism. Nevertheless, random measurement error accounted for the largest proportion of total variance (typically 0.75–0.97), and ICCs above 0.8 confirmed high overall reliability.

Together, these results show that measurement variability in osteometry arises from both systematic and random components, with more marked observer-dependent effects in the case of complex or morphologically variable traits.

This study provides a comprehensive evaluation of inter-observer reliability and measurement error, establishing the methodological soundness of the Olivier Collection dataset and supporting its future use as a validated reference resource for osteometric and forensic research.

Keywords – Inter-observer reliability, osteometry, measurements, analyse of variance, technical error of measurement, Olivier collection

Résumé – Le domaine de l'anthropométrie joue un rôle crucial dans le développement de méthodes permettant de reconstituer le profil biologique des restes squelettiques. Il est donc essentiel de garantir la précision et la fiabilité des données ostéométriques pour la validation méthodologique. Cette étude évalue la répétabilité et la fiabilité inter-observateurs des mesures squelettiques recueillies pour la publication de la collection Olivier. Quatre observateurs ont enregistré 42 variables linéaires (11 crâniennes et 31 post-crâniennes) sur 40 squelettes sélectionnés au hasard. La variation inter-observateurs a été évaluée à l'aide de tests *t* de Student, du coefficient *d* de Cohen, d'ANOVA à mesures répétées, de coefficients de corrélation intraclass (ICC) et d'erreurs techniques de mesure (TEM, rTEM) absolues et relatives. De plus, une nouvelle méthode de modèle hiérarchique linéaire a été utilisée pour partitionner la variance des mesures en composantes attribuables aux sujets, aux erreurs de mesure systématiques et aléatoires ou aux effets systématiques associés au sexe.

Des effets de mesure systématiques significatifs ($p < 0,05$) ont été identifiés pour 30 des 42 variables, en particulier dans les traits crâniens et les diamètres fémoraux/tibiaux (variance expliquée 0,19-0,37). Des effets systématiques liés au sexe ont été détectés pour 11 variables, ce qui suggère que la variabilité des observateurs peut interagir avec le dimorphisme sexuel. Néanmoins, les erreurs de mesure aléatoires représentaient la plus grande partie de la variance totale (généralement 0,75-0,97), et les ICC supérieurs à 0,8 ont confirmé une forte fiabilité globale.

Ensemble, ces résultats montrent que la variabilité des mesures ostéométriques provient à la fois de composantes systématiques et aléatoires, avec des effets plus importants liés à l'observateur dans le cas de traits complexes ou morphologiquement variables. Cette étude fournit une évaluation complète de la fiabilité inter-observateurs et des erreurs de mesure, établissant la validité méthodologique de l'ensemble de données de la collection Olivier et soutenant son utilisation future comme ressource de référence validée pour la recherche ostéométrique et médico-légale.

Mots clés – Fiabilité inter-observateur, ostéométrie, mesures, analyse de la variance, erreur technique de mesure, collection Olivier

Introduction

Accurate estimation of biological profiles, including sex, age at death, stature and population affinity, is of paramount importance in the fields of bioarchaeology and forensic anthropology. The accuracy of a method is closely dependent on its reliability, so that the use of precise and validated techniques is essential in these fields (Fojas et al., 2018). In the field of anthropology, metric methods are frequently employed to estimate sex or stature (Constantinou and Nikita, 2022; Hartley et al., 2022; Yadav et al., 2023). For instance, in the absence of the pelvis, other postcranial skeletal measurements may serve as a reliable alternative for sex estimation (Spradley and Jantz, 2011; Krüger et al., 2017; Stock, 2020; Constantinou and Nikita, 2022; Knecht et al., 2023). These metric methods are based on measurements of various skeletal elements, including long bone length, bone head diameter and epiphyseal width, and thus rely on precise and standardized acquisition of quantifiable data. These data are indispensable to formulate discriminant formulae and determine cut-off points for the estimation of biological profiles. It is therefore of great importance to establish comprehensive skeletal databases. Such databases are often compiled collaboratively by multiple observers, sometimes across different laboratories or research teams, which increases the potential for measurement variability. Assessing observer error therefore becomes crucial to ensure that these datasets can be reliably used for research.

Both intra-observer and inter-observer errors can affect the accuracy of osteometric data. Studies have highlighted that osteometric measurements are subject to variability due to factors such as observer experience, ambiguity of

measurement definitions and differences in instrument handling or positioning (Adams and Byrd, 2002; Langley et al., 2018). These sources of error are well recognized in the field, but their assessment and quantification are nevertheless essential to guarantee the reliability of anthropometric data and support the development of robust methods for biological profile estimation (Langley et al., 2018).

Several previous studies, such as Langley et al. (2018), have addressed the issue of observer error in osteometric data collection, providing important methodological frameworks. However, no such validation exists for the Olivier Collection, a French skeletal reference collection of considerable scientific value. Before making the collection's measurements available for broader research use, it is essential to assess and document the inter-observer reliability of the data collected, following best practices in the field. Our goal is therefore not to propose novel statistical techniques, but rather to apply a rigorous error quantification protocol to a new dataset, thereby contributing a validated and reusable resource for future studies in forensic anthropology and bioarchaeology.

Although other collections have been extensively used for methodological validation and standards development (e.g., Spradley and Jantz, 2011; Langley et al., 2018; Constantinou and Nikita, 2022; Knecht et al., 2023), the Olivier Collection has only been applied in a few specific studies (e.g., Guyomarc'h and Bruzek, 2010; Guyomarc'h and Bruzek, 2011). It therefore remains underexploited as a resource for comparative and reference studies. This study therefore provides an essential contribution to its wider use as a validated reference for future osteometric work.

Consequently, the objective of this study was to quantify the inter-observer error in osteological measurements from the Olivier collection. The project, informally designated the "Barbus Project," was launched following a round-table discussion held during the SAP (Société d'Anthropologie de Paris) conference. By evaluating the reliability of measurements made by different observers, we can gain a more comprehensive understanding of potential sources of error and propose recommendations to improve the standardization of measurement protocols. This in turn will help to improve the reliability of metric methods employed in bioarchaeology and forensic anthropology.

Materials and methods

The Olivier Collection is a significant anthropological reference collection held at the National Museum of Natural History (MNHN, Paris). Established in the 1960s by Professor Georges Olivier, it comprises cranial and postcranial skeletal elements from 103 identified individuals (63 males and 40 females) who donated their bodies to the MNHN or were left unburied. The biological profile, including sex, age at death and stature, is known for almost all the individuals. Figures 1 and 2 illustrate the distribution of age at death and stature, respectively. The supplementary information provided includes a file containing both the original data

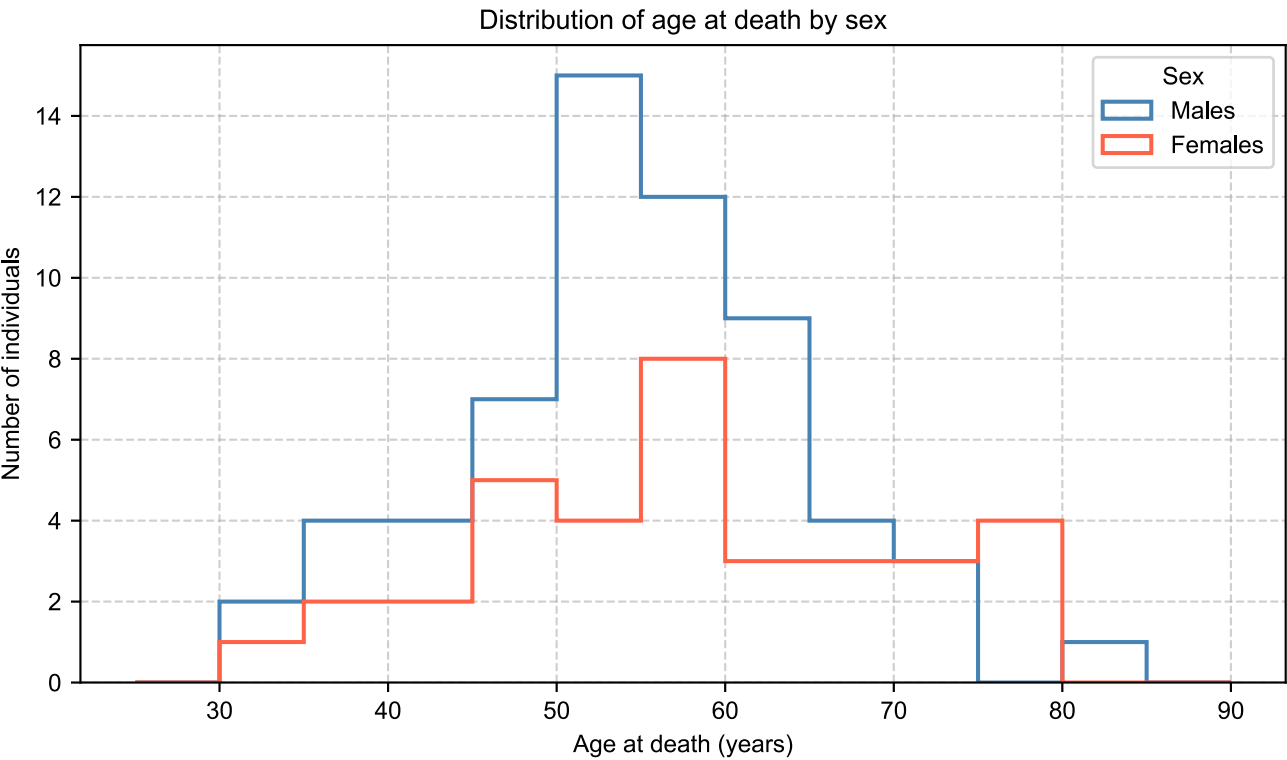


Figure 1. Age at death distribution in the Olivier collection (years) /
Répartition par âge au décès de la collection Olivier (années)

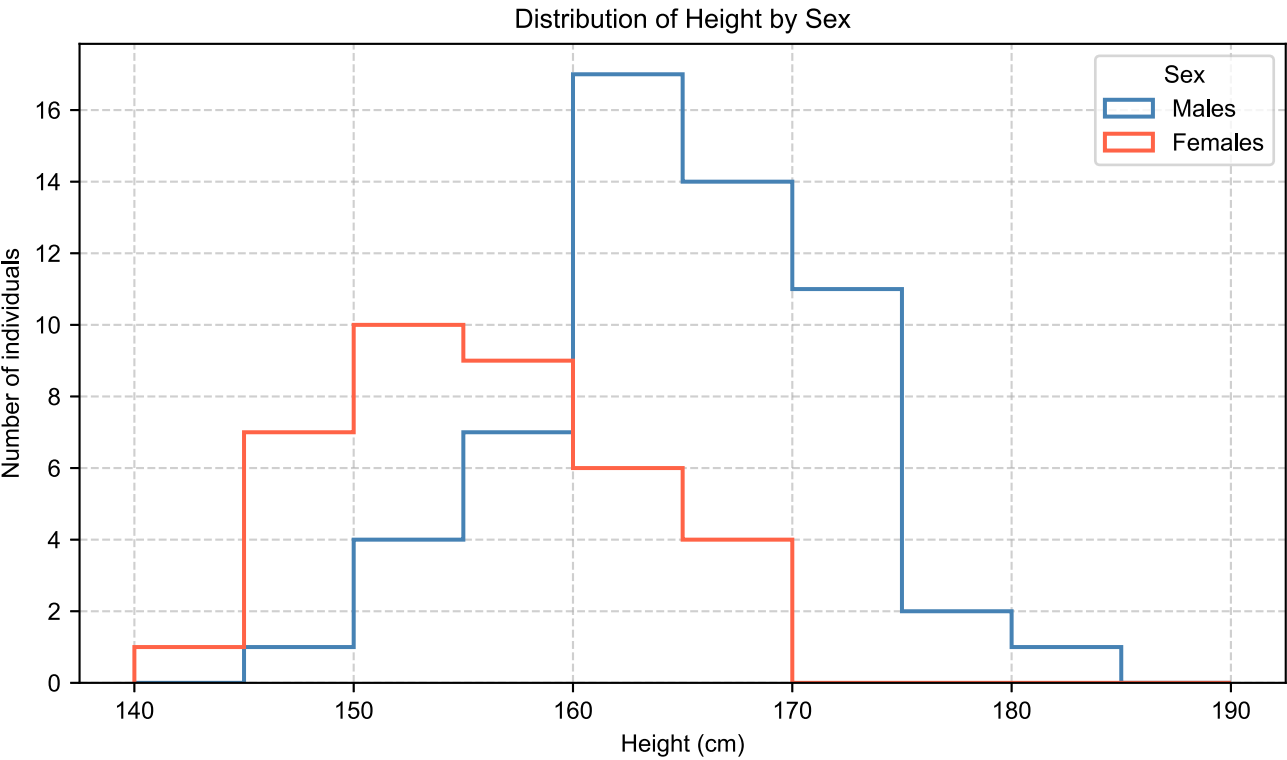


Figure 2. Stature distribution in the Olivier collection (cm) /
Distribution de la stature dans la collection Olivier (cm)

collected (SI1) and the cleaned data set (SI2). These supplementary resources provide researchers with access to both the original data and a refined dataset, thereby facilitating further analysis and ensuring transparency in the research process.

Measurements

The dataset employed in this study was collated by the authors of the paper. The dataset comprises 42 variables derived from nine bones: the cranium, mandible, clavicle,

humerus, radius, ulna, femur, tibia and talus. As detailed in table 1, the variables comprise lengths, breadths, circumferences and diameters. The linear measurements were taken in accordance with the Martin system (M-#) (Martin, 1928; Bräuer G., 1988). Of the 42 variables, 32 are bilateral, resulting in 74 measurements ($32 \times 2 + 10$ unilateral). For the bilateral variables, the left and right sides were pooled to increase the number of observations per variable and enhance statistical robustness. Since the objective was to assess measurement reproducibility rather than directional asymmetry, the data were treated by variable rather than by laterality.

Anatomical location	Bone	Code	Measurement	Tool
Upper limb	Clavicle	CLA_M1	Maximum length	Osteometric board
		CLA_M6	Circumference of the midshaft	Tape measure
	Humerus	HUM_M1	Maximum length	Osteometric board
		HUM_M4	Distal epicondylar breadth	Osteometric board
		HUM_M5	Maximum midshaft diameter	Digital callipers
		HUM_M6	Minimum midshaft diameter	Digital callipers
		HUM_M7	Least shaft circumference	Tape measure
		HUM_M9a	The transverse diameter of the head	Digital callipers
		HUM_M10a	Longitudinal diameter of the head	Digital callipers
	Radius	RAD_M1	Maximum length	Osteometric board
		RAD_M2	Physiological length	Cephalometer
		RAD_M3	Circumference of the shaft under tuberosity	Tape measure
Lower Limb	Ulna	ULN_M1	Maximum length	Osteometric board
		ULN_M2	Physiological length	Cephalometer
		ULN_M3	Minimum shaft circumference	Tape measure
	Femur	FEM_M1	Maximum length	Osteometric board
		FEM_M2	Physiological length	Osteometric board
		FEM_M21	Epicondylar breadth	Osteometric board
		FEM_M6	Anteroposterior diameter of the midshaft	Digital callipers
		FEM_M7	Mediolateral diameter of the midshaft	Digital callipers
		FEM_M8	Circumference of the midshaft	Tape measure
		FEM_M9	Transverse subtrochanteric diameter	Digital callipers
		FEM_M10	Subtrochanteric anterior-posterior diameter	Digital callipers
		FEM_M18	Vertical diameter of the head	Digital callipers
		FEM_M19	The transverse diameter of the head	Digital callipers
	Tibia	TIB_M1	Lateral condyle-medial malleolus length	Osteometric board
		TIB_M2	Physiological length	Large callipers
		TIB_M8a	Anteroposterior diameter at the nutrient foramen	Digital callipers
		TIB_M9a	Transverse diameter at the nutrient foramen	Digital callipers
		TIB_M10b	Minimum shaft circumference	Tape measure
	Talus	TAL_M1a	Maximum length	Osteometric board
Skull	Cranium	CRA_M1	Maximum cranial length	Cephalometer
		CRA_M8	Maximum cranial breadth	Cephalometer
		CRA_M9	Least frontal breadth	Cephalometer
		CRA_M10	Maximum frontal breadth	Cephalometer
		CRA_M12	Asterionic breadth	Cephalometer
		CRA_M17	Basion-bregma height	Cephalometer
		CRA_M20pb	Porion-bregma	Cephalometer
		CRA_M20pp	Porion-porion	Digital callipers
	Mandible	MAN_M68(1)	Maximum projective length of the mandible	Osteometric board
		MAN_M65	Bicondylar breadth	Osteometric board
		MAN_M66	Bigonial breadth	Digital callipers

Table 1. Measurements taken from the Olivier Collection, along with their definitions and the measurement tools used to collect them / Mesures réalisées sur la collection Olivier, accompagnées de leurs définitions et des outils de mesure utilisés pour les recueillir

Bones exhibiting evidence of trauma or abnormalities were included in the final dataset. These bones were intentionally retained in order to preserve the representativeness of the Olivier Collection and to reflect the natural variability encountered in skeletal assemblages. This decision prevents bias that could result from systematically excluding altered specimens, which form part of the biological and taphonomic reality of human osteological samples. Additional analyses excluding these individuals were also conducted and did not substantially affect the results.

The measurements were taken by the authors themselves using standardized equipment in order to ensure consistency. To assess inter-observer variability, four observers measured a randomly selected subsample of 40 individuals. All four observers have extensive experience in osteometric data collection, each with over 15 years of professional practice in anthropology and familiarity with standardized osteometric procedures. All are trained in the Martin (1928) measurement system and routinely apply these measurements in both research and field contexts, ensuring a high level of consistency and methodological rigour across observations. While intra-observer error was not formally quantified, the extensive experience and standardized training of the observers ensured high internal consistency. The set of measurements included in this study was selected because these variables are among the most commonly used to estimate sex, stature and population affinity (e.g., Spradley and Jantz, 2011; Langley et al., 2018; Constantinou and Nikita, 2022; Knecht et al., 2023). This selection was intended to ensure comparability with existing studies and methodological frameworks in biological and forensic anthropology. To further minimize potential differences related to instrument handling, and to avoid confirmation bias, all measurements were conducted independently and at different times by the four observers using the same set of standardized tools. The supplementary information includes a file providing both the raw inter-observer measurements (SI3) and their cleaned versions (SI4), to ensure transparency and facilitate further analysis by other researchers.

Statistical analysis

The statistical analyses were conducted using the R software (version 4.2.2). Following visualization of the data, an initial data preprocessing phase was undertaken with the objective of cleaning the dataset. In line with the approach outlined by Santos (2020), boxplots and the inter-quantile range (IQR) were employed to identify and eliminate extreme values that could be classified as outliers, for instance as a result of data entry errors. Furthermore, some variables exhibited inconsistencies in inversion between observers, particularly in the case of diaphysis diameters (HUM_M5/HUM_M6 and FEM_M9/FEM_M10), which were subsequently corrected manually.

The degree of agreement between the four observers was first explored using a repeated-measures analysis of variance (ANOVA). This test was used to identify systematic mean

differences between observers, in other words to determine whether any observer consistently produced higher or lower values for a given variable. To increase the number of observations in the analysis, left and right side measurements were treated as separate instances of the same variable (e.g., left femur length and right femur length were both included as independent observations), allowing us to better evaluate measurement reproducibility across sides without averaging or excluding data. When the Mauchly test indicated that the assumption of sphericity had not been met, a Greenhouse-Geisser correction was applied. When a significant main effect was detected in the repeated-measures ANOVA, post hoc pairwise comparisons were conducted to determine which observers' mean values differed significantly from each other. In other words, while the ANOVA assesses whether any overall difference exists among observers' means, post hoc analyses specify the particular pairs of observers whose mean measurements differ. These comparisons were adjusted using the Bonferroni correction to control for the inflation of Type I error due to multiple testing. The results of these post hoc analyses therefore provide detailed information on which pairs of observers exhibit statistically significant mean differences.

In instances where significant ANOVA differences were identified, absolute and relative technical errors of measurement (TEM) were calculated in order to assess the degree of error between the four observers. In accordance with the methodology proposed by Langley et al., (2018), the absolute technical error of measurement (TEM) is calculated as follows:

$$\sqrt{\frac{\sum_{i=1}^N \left[\sum_{j=1}^K M_{ij}(n)^2 - \frac{\left(\sum_{j=1}^K M_{ij}(n) \right)^2}{K} \right]}{N(K-1)}}$$

where N is the sample size, K is the number of observers, M is the measurement, and M(n) is the nth repetition of the measurement (Mony et al., 2016). The relative TEM (rTEM) was calculated by dividing the absolute TEM by the mean and multiplying the result by 100 (Langley et al., 2018). The relative TEM serves as a metric of accuracy, irrespective of the scale or sample size, thus facilitating direct comparison of measurements across different scales. In the field of osteometry, there is currently no consensus regarding the threshold value that should be used to determine the validity of a TEM. However, some studies, such as that conducted by Langley et al., have employed an arbitrary threshold of rTEM <1.5% for intra-examiner error and <2% for inter-examiner error (Langley et al., 2018). This approach was based on the findings of previous anthropometric studies on living individuals (Perini et al., 2005). These thresholds will be used as a reference point for the assessment of measurement validity in the study.

Inter-observer reliability was also assessed using the intraclass correlation coefficient (ICC). The ICC is a statistical measure that is commonly employed to quantify the degree of agreement between multiple observers who are measuring the same quantity (Koo and Li, 2016). The aim

was to assess the extent of the variability between the observers in relation to the total variability of the measurements. The Intraclass Correlation Coefficient (ICC) was calculated for each of the 42 variables in the Olivier dataset in order to assess the degree of agreement between the various observers who performed the osteological measurements. Unlike TEM or ANOVA, the ICC quantifies the proportion of total variance attributable to inter-individual differences, offering an additional dimension to assess the reproducibility of each variable. The ICC values were subsequently interpreted in accordance with the guidelines set forth by Koo and Li (2016). A value of less than 0.50 indicates poor agreement, a value between 0.50 and 0.75 indicates moderate agreement, a value between 0.75 and 0.90 indicates good agreement, and a value greater than 0.90 indicates excellent agreement. A high ICC value indicates a high degree of similarity between the measurements taken by different observers for a given variable.

In response to recent discussions regarding the limitations of conventional methods for assessing measurement error (Collyer and Adams, 2024), we also implemented a complementary analysis using the *measurement.error()* function from the RRPP package in R. Using a hierarchical linear model, this approach partitions the total variance into components attributable to differences among individuals, systematic differences between observers, group-related effects (here sex), and residual random measurement error. It can thus quantify the relative contribution of each source of error. In this study, the function was applied separately to each osteometric variable to assess the presence and magnitude of systematic inter-observer differences. By combining this method with classic ANOVA and TEM analyses, we aimed to obtain a more comprehensive evaluation of measurement reliability and to distinguish between systematic and random components of observer error.

In addition, a Student's *t*-test was performed to compare the means between the two sexes for each measurement and each observer. This approach was intended to assess whether each observer, independently, was able to detect consistent sex-based differences for each measurement. By comparing male and female means within each observer's dataset, we aimed to evaluate whether inter-observer variation affected the interpretation of biological differences, rather than directly comparing the observers to one another. Applying this test enabled us to determine whether the observed difference between the means is statistically significant, based on a *p*-value threshold of 0.05, and whether this result is consistent between observers. In addition to statistical significance, it is important to consider the size effect, which indicates the magnitude of differences between groups (here, males and females) (Tanner-Smith et al., 2018). Cohen's *d* is a measure of effect size that is commonly employed in Student's *t*-tests. The value was calculated by dividing the difference between the means by the standard deviation common to both groups. The *d* values were interpreted in accordance with the guidelines set forth by Cohen. A value of 0.2 represents a small effect, 0.5 a medium effect, and 0.8 a large effect (Gro   and M  ller, 2023). In instances where the Student's *t*-test did not yield statistically significant differences between observers, the application of Cohen's *d* was not deemed necessary.

Results

The descriptive statistics (Table S5 in Supplementary Information) show that for all measurements, the means and standard deviations of the four observations were relatively similar. Figure 3 shows that, visually, some variables are highly reproducible, whereas for others, there are significant differences between observers.

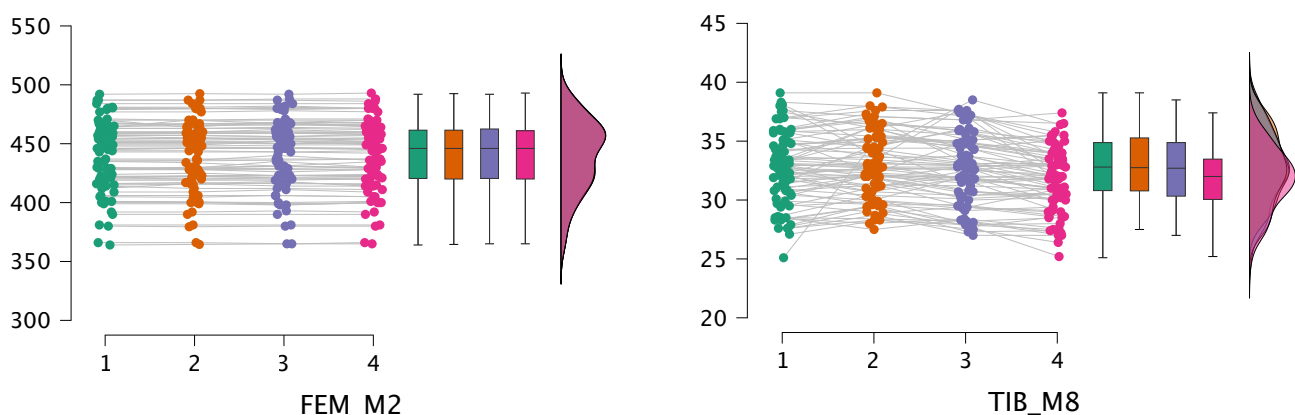


Figure 3. Example raincloud plot comparing a highly reproducible variable (left: FEM_M2; physiological length of femur) with a variable that shows differences between observers (right: TIB_M8a; Anteroposterior diameter at the nutrient foramen of the tibia) (mm) / *Exemple de graphique en nuage de pluie comparant une variable fortement reproductible (   gauche : FEM_M2 ; longueur physiologique du f  mur)    une variable pr  sentant des diff  rences entre les observateurs (   droite : TIB_M8a ; diam  tre ant  ropost  rieur au niveau du foramen nourricier du tibia) (mm)*

The results of the analysis of variance (ANOVA) are presented in table 2. These results show that 31 of the 42 measurements exhibited statistically significant differences between observers. With regard to the aforementioned variables, seven are lengths (RAD_M2, ULN_M1, ULN_M2, FEM_M1, TIB_M1, TIB_M2, TAL_M1), one is a breadth (HUM_M4), and nine are diameters (HUM_M5, HUM_M6, HUM_M9, HUM_M10, FEM_M6, FEM_M7). The remaining measurements are circumferences (CLA_M6, HUM_M7, RAD_M3, FEM_M8, TIB_M10) and skull measurements (CRA_M1, CRA_M8, CRA_M10, CRA_M12, CRA_M17, RA_M20pb, CRA_M20pp, MAN_M65,

MAN_M66). Upon grouping the measurements by type (diameters, lengths, breadths, circumferences), it was found that significant differences existed for all of these types of measurements. Subsequent analyses indicated that the majority of differences were observed between observers 3 and 4. In terms of frequency, observer 4 exhibited the greatest discrepancy with the other observers, followed by observer 3. Conversely, the ICC (table 2) between the measurements of the four observers were all above 0.8, indicating a strong correlation between observers, with the exception of the variable CRA_M12, which exhibited a moderate correlation coefficient (0.75).

Measurement	ANOVA (<i>p</i> -value)	Differences between post-hoc pairs observers	ICC
CLA_M1	0.172	Not significant	0.98
CLA_M6	0.014*	1-3 / 1-4	0.94
HUM_M1	0.506	Not significant	0.97
HUM_M4	0.009*	2-3 / 3-4	0.98
HUM_M5	<0.001*	1-3 / 2-3 / 3-4	0.95
HUM_M6	<0.001*	1-2 / 1-4 / 2-3 / 3-4	0.96
HUM_M7	0.006*	1-4 / 2-4 / 3-4	0.96
HUM_M9	<0.001*	1-4 / 2-4 / 3-4	0.98
HUM_M10	0.033*	1-4	0.89
RAD_M1	0.062	Not significant	0.99
RAD_M2	<0.001*	1-4 / 2-3 / 2-4 / 3-4	0.99
RAD_M3	0.05	Not significant	0.97
ULN_M1	<0.001*	1-3 / 2-3 / 3-4	0.97
ULN_M2	0.006*	1-2 / 1-4	0.99
ULN_M3	0.147	Not significant	0.94
FEM_M1	0.046*	3-4	0.99
FEM_M2	0.092	Not significant	1
FEM_M21	0.739	Not significant	1
FEM_M6	0.006*	2-3 / 3-4	0.98
FEM_M7	0.018*	1-4	0.96
FEM_M8	<0.001*	1-4 / 2-3 / 2-4 / 3-4	0.99
FEM_M9	<0.001*	2-3 / 2-4	0.93
FEM_M10	<0.001*	1-2 / 1-4 / 2-3 / 3-4	0.83
FEM_M18	0.543	Not significant	0.97
FEM_M19	0.193	Not significant	0.97
TIB_M1	0.008*	2-4	1
TIB_M2	0.008*	1-4 / 3-4	1
TIB_M8	<0.001*	1-4 / 2-4 / 3-4	0.83
TIB_M9	0.064	Not significant	0.93
TIB_M10	<0.001*	1-2 / 1-3 / 2-4 / 3-4	0.93
TAL_M1	0.003*	1-2 / 1-3	0.94
CRA_M1	0.001*	2-3 / 3-4	0.97
CRA_M8	<0.001*	1-3 / 2-3 / 3-4	0.86
CRA_M9	0.067	Not significant	0.93
CRA_M10	<0.001*	1-2 / 1-3 / 1-4 / 2-3 / 3-4	0.94
CRA_M12	0.024*	3-4	0.75
CRA_M17	<0.001*	1-3 / 2-3 / 3-4	0.97
CRA_M20pb	<0.001*	1-3 / 1-4	0.95
CRA_M20pp	<0.001*	1-4 / 2-4 / 3-4	0.86
MAN_M68	0.081	Not significant	0.87
MAN_M65	0.018*	2-4 / 3-4	0.93
MAN_M66	0.029*	2-4 / 3-4	0.82

Table 2. Comparison obtained between ANOVAs and ICC. * represents a significant difference between observers /
*Comparaison obtenue entre les ANOVA et les ICC. * représente une différence significative entre les observateurs*

Table 3 presents the technical error measurement (TEM) for each measurement exhibiting significant differences according to the ANOVA ($p < 0.05$). The relative technical error measurement (rTEM) values ranged from 0.4% to 3.7%. For these measurements, the majority demonstrated acceptable technical errors in accordance with the conventional criteria (rTEM $< 2\%$; Langley et al., 2018). Of the ten measurements with unacceptable errors, six diameters (HUM_M5, HUM_M6, HUM_M10, FEM_M9, FEM_M10, TIB_M8), one circumference (CLA_M6), and three skull measurements (CRA_M10, CRA_M12, MAN_M66) exhibited higher inter-observer variability, probably reflecting challenges to consistent identification of anatomical landmarks.

The results of the measurement error analysis conducted using the *measurement.error()* function from the RRPP package are summarized in table 4. This method partitioned the total variance of each variable into three main components: (1) variance among individuals (Subjects), (2) systematic measurement error (Systematic ME), (3) systematic measurement error associated with sex (Systematic ME: Sex), and (4) residual variance (Random ME).

Significant systematic measurement effects ($p < 0.05$) were observed for 30 of the 42 variables, indicating that a non-negligible portion of variability was attributable to consistent differences between observers. The highest values of Systematic ME were observed for cranial (CRA_M8, CRA_M10, CRA_M20pp) and diameter (FEM_M10, TIB_M8) measurements, with explained variances ranging from 0.19 to 0.37. Systematic effects associated with sex (Systematic ME: Sex) were significant for 11 variables, most notably for HUM_M7, RAD_M2, TIB_M8, TIB_M10, and several cranial traits (CRA_M8, CRA_M9, CRA_M10, CRA_M12, CRA_M20pp). These results suggest that, for certain measurements, observer-related differences interacted with sex-related variation in bone morphology. Random measurement error (Random ME) accounted for the majority of the total variance for most variables, typically ranging from 0.75 to 0.97, indicating that residual variability not explained by systematic observer effects remained the dominant source of dispersion.

Overall, the results of the hierarchical linear model confirm the presence of both systematic and random measurement components, with some variables showing stronger

Bone	Measurement	TEM relative	TEM absolute
Clavicle	CLA_M6	2.63%	0.95
Humerus	HUM_M4	1.47%	0.87
	HUM_M5	2.13%	0.47
	HUM_M6	2.28%	0.41
	HUM_M7	1.51%	0.93
	HUM_M9	1.39%	0.59
	HUM_M10	3.09%	1.42
Radius	RAD_M2	0.53%	1.17
Ulna	ULN_M1	0.62%	1.54
	ULN_M2	0.55%	1.20
Femur	FEM_M1	0.41%	1.2
	FEM_M6	1.95%	0.55
	FEM_M7	1.83%	0.50
	FEM_M8	1.02%	0.89
	FEM_M9	2.62%	0.82
	FEM_M10	3.33%	0.89
Tibia	TIB_M1	0.41%	1.45
	TIB_M2	0.40%	1.36
	TIB_M8	3.7%	1.20
	TIB_M10	1.62%	1.14
Talus	TAL_M1	1.96%	1.12
Cranium	CRA_M1	0.83%	1.49
	CRA_M8	1.48%	2.14
	CRA_M10	2.73%	2.72
	CRA_M12	3.34%	3.34
	CRA_M17	0.54%	0.71
	CRA_M20pb	1.1%	1.10
	CRA_M20pp	1.94%	2.32
Mandible	MAN_M65	0.86%	1.01
	MAN_M66	2.5%	2.28

Table 3. TEM values for inter-observer error. Relative TEM (%) is unitless, and absolute TEM is in millimetres / *Valeurs TEM pour l'erreur inter-observateurs. La valeur relative TEM (%) est sans unité, et la valeur absolue TEM est exprimée en millimètres (mm)*

Measurement	Systematic ME	Systematic ME: Sex	Random ME
CLA_M1	0.03	0.02	0.95
CLA_M6	0.09*	0.03	0.88
HUM_M1	0.01	0.02	0.97
HUM_M4	0.05*	0.01	0.94
HUM_M5	0.15*	0.01	0.84
HUM_M6	0.10*	0.01	0.89
HUM_M7	0.17*	0.08*	0.75
HUM_M9	0.15*	0.01	0.83
HUM_M10	0.05*	0.01	0.94
RAD_M1	0.05*	0.01	0.94
RAD_M2	0.15*	0.04*	0.81
RAD_M3	0.04*	0.04*	0.92
ULN_M1	0.09*	0.02	0.89
ULN_M2	0.06*	0.02	0.92
ULN_M3	0.02	0.05*	0.93
FEM_M1	0.04	0.01	0.95
FEM_M2	0.03	0.01	0.96
FEM_M21	0.01	0.01	0.98
FEM_M6	0.07*	0.01	0.92
FEM_M7	0.05*	0.01	0.94
FEM_M8	0.15*	0.02	0.83
FEM_M9	0.09*	0.03	0.88
FEM_M10	0.19*	0.03	0.77
FEM_M18	0.01	0.02	0.97
FEM_M19	0.02	0.02	0.96
TIB_M1	0.06*	0.03	0.91
TIB_M2	0.06*	0.02	0.92
TIB_M8	0.20*	0.06*	0.74
TIB_M9*	0.04*	0.05*	0.91
TIB_M10	0.15*	0.08*	0.78
TAL_M1	0.07*	0.02	0.91
CRA_M1	0.15*	0.02	0.93
CRA_M8	0.33*	0.04*	0.63
CRA_M9	0.08*	0.01*	0.92
CRA_M10	0.37*	0.05*	0.93
CRA_M12	0.10*	0.02*	0.87
CRA_M17	0.23*	0.04	0.73
CRA_M20pb	0.07*	0.01	0.92
CRA_M20pp	0.29*	0.05*	0.66
MAN_M68	0.05	0.06	0.89
MAN_M65	0.03	0.02	0.95
MAN_M66	0.12*	0.04	0.84

Table 4. Summary of results using the hierarchical linear model method. Values represent the proportion of total variance explained by each component. * indicates significant effects ($p < 0.05$). ME = Measurement Error / *Résumé des résultats de la méthode du modèle linéaire hiérarchique. Les valeurs représentent la proportion de la variance totale expliquée par chaque composante. * indique des effets significatifs ($p < 0,05$). EM = erreur de mesure*

observer-dependent effects, particularly in more complex or morphologically variable traits such as cranial and femoral measurements. These findings complement the ANOVA and TEM results, showing that while reproducibility was generally high, some morphological variables were more susceptible to subtle observer effects.

For the majority of postcranial measurements, the p -values derived from the Student's t -test were less than 0.05 for each observer, indicating that, despite inter-observer variability, all observers independently detected statistically

significant differences between sexes for the majority of variables, reflecting consistency in the biological signal interpretation. A closer examination of the data reveals that only a few p -values exceeded the 0.05 threshold, specifically for the CLA_M1 left, TIB_M2 right, and TIB_M8 left postcranial measurements. These non-significant results are present on a single side only, and the statistical tests conducted on each observer's dataset do not always yield the same significance outcome, indicating variability in the detection of sex-based differences for that side. With regard

to the skull measurements, only a small number of p -values exceeded the 0.05 threshold (CRA_M1, CRA_M8, CRA_M9 right, CRA_M10, CRA_M12, CRA_M17 right, CRA_M20pp left), thereby indicating that the observed differences are not statistically significant. For the measurements CRA_M1 left, CRA_M9 right, CRA_M10 right, CRA_M12 left, and CRA_M20pb right, there was considerable variation between observers in the results obtained. Some pairs of observers demonstrated significant differences between their results, whereas other pairs did not. Notwithstanding the differences in the Student's t -test results, the effect size coefficients provided by Cohen were found to be very similar for all measurements, with the exception of four skull measurements (CRA_M1 left, CRA_M9 right, CRA_M10 left, CRA_M12 left). This indicates that the differences in measurements between observers have a negligible impact on the conclusions of the comparison of metric differences between male and female subjects. Consequently, these measurement inconsistencies do not result in interpretative differences regarding the comparisons made between sexes.

Discussion

This study aimed to assess the degree of inter-observer agreement in linear osteometric measurements commonly used for estimating sex, stature and population affinity, and to evaluate the indirect impact of observer error on biological parameter estimations, using sex as an illustrative example. The variables analysed correspond to those routinely implemented in widely used reference frameworks, such as *Fordisc* (Jantz and Ousley, 2005) or *SexEst* (Constantinou and Nikita, 2022), thereby enhancing the broader applicability and comparability of our results in both forensic and anthropological contexts. By examining the reproducibility of measurements performed by four experienced observers, this research sought to identify which types of osteometric traits are most susceptible to inter-observer variability and to discuss the potential implications of such variability for biological profile reconstruction. In doing so, the study contributes to ongoing efforts to standardize osteometric methodologies and strengthen measurement reliability across institutions and practitioners.

A repeated-measures ANOVA was conducted to determine whether significant differences existed among the mean values obtained by the four observers. To increase the sample size, left and right elements were combined into a single variable for this analysis. A total of 31 out of the 42 measurements yielded p -values below 0.05, indicating significant inter-observer differences. Conversely, non-significant results should not be interpreted as an absence of signal; rather, they may reflect lower statistical power or smaller effect size. Subsequent analyses demonstrated that observers 3 and 4 were more prone to diverge in their assessments. For several measurements, the four observers exhibited significant differences in their results. These included the minimum midshaft diameter of the humerus (HUM_M6),

the physiological length of the radius (RAD_M2), maximum length of the ulna (ULN_M1), circumference of the femur midshaft (FEM_M8), subtrochanteric anterior-posterior diameter of the femur (FEM_M10), anteroposterior diameter of the tibia at the nutrient foramen (TIB_M8), minimum shaft circumference of the tibia (TIB_M10), maximum cranial breadth (CRA_M8), maximum frontal breadth (CRA_M10), basion-bregma height (CRA_M17) and porion-porion distance (CRA_M20pp). Such between-observer differences in measurements could be a reflection of various compounding factors, such as locating biologically variable and idiosyncratic anatomical landmarks, different interpretations of measurement definitions and variability in observer training.

Observers 1 and 2, who were trained at the same institution, showed higher agreement, whereas observers 3 and 4, trained in different institutions, diverged more frequently. This pattern supports the idea of a “school effect,” in which variations in training protocols and pedagogical approaches influence measurement practice. This observation highlights the need for greater standardization of osteometric training across institutions to minimize inter-observer variation and improve reproducibility in anthropological datasets.

In order to ascertain the significance of these differences, it was deemed pertinent to calculate the technical error of measurement (TEM), which enabled the quantification of the accuracy and reliability of the collected data. This step was essential to assess the magnitude of observer error and to identify any sources of variation linked to the measurement procedure. By quantifying this error, the study assessed how inter-observer variability could influence the detection of sexual dimorphism in osteometric analyses.

Despite the considerable differences observed in the ANOVA, the majority of measurements exhibited a relatively low technical error (less than 4%). To complete this analysis, an ICC was conducted to verify the precision of each observer across all measurements, with highly favourable results, as evidenced by a high correlation coefficient of more than 0.8 for all variables, except for one cranial measurement (asterionic breadth, CRA_M12), for which agreement is moderate (ICC=0.75). When focusing on the rTEM, ten measurements were found to exceed the recommended acceptable value of 2% (as proposed by Langley et al., 2018). These included six diameters: maximum midshaft diameter of the humerus (HUM_M5), minimum midshaft diameter of the humerus (HUM_M6), longitudinal diameter of the humeral head (HUM_M10), transverse subtrochanteric diameter of the femur (FEM_M9), subtrochanteric anterior-posterior diameter of the femur (FEM_M10) and anteroposterior diameter of the tibia at the nutrient foramen (TIB_M8); one circumference, the midshaft circumference of the clavicle (CLA_M6); and three skull measurements, maximum frontal breadth (CRA_M10), asterionic breadth (CRA_M12), and bigonial breadth of the mandible (MAN_M66).

However, the selection of an appropriate rTEM threshold is not universal and must be contextualized. Previous studies have proposed a range of acceptable values depending on measurement type and method. For instance, Perini et al. (2005) suggested a 2% threshold for conventional anthropometry, whereas Norton and Olds (1996) reported higher acceptable limits (up to 5-10%) in other anthropometric contexts, where different sources of variability are at play. In our study, rather than adopting a fixed universal criterion, we interpret rTEM values in relation to the specific characteristics of osteometric data, variability in states of preservation and differences in measurement instruments. Given the variability in the morphology and state of preservation of skeletal remains, we considered rTEM <4% acceptable, as this corresponds to <1 mm absolute error for most variables, and thus remains within the tolerance typically reported for manual osteometry. Moreover, our results compare favourably with those reported by Langley et al. (2018), who observed similar or higher values even for well-established measurements.

Importantly, complementary analyses using Student's *t*-tests and Cohen's *d* demonstrated that these levels of inter-observer variability did not alter the statistical outcomes of sex-based comparisons. The same measurements produced equivalent levels of significance and comparable effect sizes across observers, indicating that minor differences did not meaningfully affect the detection of sexual dimorphism in this dataset. This finding reinforces the view that moderate levels of technical error, within reasonable limits, will not necessarily compromise biological interpretations derived from osteometric data.

Complementary analyses using a hierarchical linear model approach confirm that measurement variability comprises both systematic and random components, with stronger observer-dependent effects for more complex or morphologically variable traits such as cranial measurements and diameters. These systematic effects most likely reflect the combined influence of tool-specific constraints, anatomical complexity and inter-observer differences in landmark interpretation.

Maximum lengths and widths were measured using an osteometric board, while circumferences were gauged with a tape measure, diameters with digital callipers, and cranial measurements taken with a cephalometer. The results show that the most frequent errors involve the use of digital callipers, which are accurate to one tenth of a millimetre. The least accurate measurements, including longitudinal diameter of the humeral head (HUM_M10), subtrochanteric anterior-posterior diameter of the femur (FEM_M10), anteroposterior diameter of the tibia at the nutrient foramen (TIB_M8) and asterionic breadth (CRA_M12) (rTEM >3%), were indeed predominantly long bone diameters. Our findings align with Langley et al. (2018), who support the idea that positionally dependent measurements present greater challenges than maxima or minima, primarily due to difficulties in maintaining consistent bone orientation. Additional factors, such as minor preservation

differences or irregular surface morphology, may further contribute to these discrepancies.

While the use of digital callipers offers a high degree of precision, it is essential to ensure the correct orientation of the bone prior to measurement, as these instruments are designed for the measurement of transverse or anteroposterior diameters. It is likely that this need for exact orientation contributes considerably to the observed inter-observer differences, given that each observer may orient the bone in a slightly different manner (Langley et al., 2018). To mitigate these differences, the Data Collection Procedures 2.0 recommend substituting these measurements with maximum and minimum midshaft diameters (Langley et al., 2016). This approach may prove an effective means of reducing the impact of bone orientation on measurement consistency.

It should be noted that differences may also arise from other sources, beyond the specific tool or technique employed. For example, the discrepancy between measurements of the midshaft circumference of the clavicle (CLA_M6) may have been due to an error in reading the measuring tape by observer 1. In the case of cranial measurements with higher rTEM values, the difficulties are likely to be attributable to the inherent difficulty of precisely locating certain landmarks. For example, the definition of the asterion is notably imprecise, which can result in considerable inter-observer variability. Similarly, measurements involving the gonion can vary depending on whether observers use the full width or meticulously position the calliper points at the 'middle' of the bone. These anatomical ambiguities underscore the necessity for more precise definitions and standardized protocols for cranial measurements. Furthermore, the difficulty of reproducing skull measurements may have been contingent on the definition employed. These observations emphasize the need for precise and standardized definitions of anatomical landmarks and measurement techniques, particularly for complex structures such as the skull. While the existing literature (e.g., Langley et al., 2018, the Data Collection Procedures 2.0), has addressed many of these issues, it is recommended that future work should build upon these established protocols to further refine definitions and measurement procedures for problematic variables.

It should be noted that our sample included individuals with bone pathologies. Although these conditions can complicate the identification of landmarks and the application of measurement definitions, the study did not specifically aim to evaluate their effect on measurement variability. A supplementary comparison was conducted, excluding bones with visible pathological alterations, and no notable difference in the results was observed. However, this should not be interpreted as evidence that pathological conditions do not influence measurement consistency, as the limited number of pathological specimens may have reduced the capacity to detect such effects. These observations underline the need for future research explicitly designed to quantify the impact of pathological variation on inter-observer reliability in osteometric studies.

In the light of these findings, it is obvious that analyses of inter-observer reliability have a pivotal role in the field of anthropometry. Such analyses facilitate the identification of problematic measurements and potential sources of error, thereby enabling the implementation of improvements in measurement procedures and definitions. The slight differences observed between observers serve to emphasize the importance of rigorous, standardized training for all observers. The interpretation of a definition can result in the introduction of systematic errors and differences in the acquisition of measurements.

While this study used sex differences as a reference point for assessing the potential impact of inter-observer measurement variability, we acknowledge that osteometric data are employed in a wider range of analyses, including stature estimation and the study of population affinity. Future research should evaluate how measurement error may influence these applications, particularly those relying on multivariate or population-level comparisons.

Finally, emerging digital technologies offer promising avenues for addressing some of these challenges. Three-dimensional surface scanning and CT-based modelling allow for the replication and archiving of skeletal morphology in virtual form, enabling repeated measurements without physical manipulation. Nevertheless, these digital models are not immune to observer error: they are still influenced by manual processing, software choices and device-specific parameters, all of which can introduce variability. Recent developments, however, are gradually reducing such sources of error. For example, Anderson et al. (2024) demonstrated how a laser-based system can achieve osteometric measurements comparable to those obtained with a traditional osteometric board, while substantially reducing manual handling errors. In addition, Simon et al. (2023) illustrated how semi- and fully automated systems can standardize measurement acquisition and minimize user-dependent variability. While these advances do not eliminate human influence entirely, they represent a critical step toward more reproducible, objective and scalable osteometric data collection. Consequently, the adoption of digital and automated approaches must still be accompanied by robust calibration protocols and transparent reporting standards to ensure comparability across studies.

Conclusion

The objective of this study was to evaluate the inter-observer reliability of anthropometric measurements on a sample of 40 individuals from the Olivier collection. Although some statistically significant differences were observed between observers, the majority of deviations were minor and within the acceptable thresholds established in the literature. Our findings highlight the necessity for methodological rigour and standardized procedures in anthropometry, emphasizing the importance of proper observer training, bone orientation, appropriate tool usage and reliability analysis. The demonstrated reproducibility

serves to validate the metric dataset collected from the Olivier collection, which in turn serves to illustrate the additional value of this data collection and its analysis in providing a reliable metric dataset for future studies. This study makes a contribution to the field by offering a validated anthropometric dataset that can be shared and used with confidence in future research. The original data collected and the cleaned data are both provided in the supplementary information, thus ensuring transparency and facilitating further analysis by other researchers. This comprehensive approach to data sharing enhances the reproducibility and utility of our findings for the broader anthropometric research community.

Acknowledgments

We would like to thank the National Museum of Natural History (MNHN) in Paris for facilitating access to the Olivier collection. Furthermore, we would like to express our gratitude to all the members of the BARBUS project.

References

- Adams BJ, Byrd JE (2002) Interobserver Variation of Selected Postcranial Skeletal Measurements. *Journal of Forensic Sciences* 47(6):1555-1560 [<https://doi.org/10.1520/JFS15550J>]
- Anderson E, Malis SW, Osterholtz AJ et al. (2024) Investigating the Reliability and Validity of the Portable Osteometric Device. *Advances in Archaeological Practice* 12(2):133-142 [<https://doi.org/10.1017/aap.2023.33>]
- Bräuer G. (1988) Osteometrie. In: Knussmann R. (Ed.), *Anthropologie: handbuch der vergleichenden Biologie des Menschen*. G. Fischer, Stuttgart, pp. 160-232.
- Collyer ML, Adams DC (2024) Interrogating Random and Systematic Measurement Error in Morphometric Data. *Evolutionary Biology* 51(1):179-207 [<https://doi.org/10.1007/s11692-024-09627-6>]
- Constantinou C, Nikita E (2022) SexEst: An open access web application for metric skeletal sex estimation. *International Journal of Osteoarchaeology* 32(4):832-844 [<https://doi.org/10.1002/oa.3109>]
- Fojas CL, Kim J, Minsky-Rowland JD et al. (2018) Testing inter-observer reliability of the Transition Analysis aging method on the William M. Bass forensic skeletal collection. *American Journal of Physical Anthropology* 165(1):183-193 [<https://doi.org/10.1002/ajpa.23342>]
- Groß J, Möller A (2023) A Note on Cohen's d From a Partitioned Linear Regression Model. *Journal of Statistical Theory and Practice* 17(2):22 [<https://doi.org/10.1007/s42519-023-00323-w>]
- Guyomarc'h P., Bruzek J (2010) Dimorphisme sexuel du crâne de sujets identifiés (collection Olivier, MNHN, Paris): évaluation par morphométrie géométrique. *Bulletins et mémoires de la Société d'anthropologie de Paris* 22(3-4):216-229 [<https://doi.org/10.1007/s13219-010-0019-6>]
- Guyomarc'h P, Bruzek J (2011) Accuracy and reliability in sex determination from skulls: A comparison of Fordisc® 3.0 and the discriminant function analysis. *Forensic Science In-*

- ternational 208(1-3):180.e1-180.e6 [<https://doi.org/10.1016/j.forsciint.2011.03.011>]
- Hartley S, Winburn AP, Dror IE (2022) Metric forensic anthropology decisions: Reliability and biasability of sectioning-point-based sex estimates. *Journal of Forensic Sciences* 67(1): 68-79 [<https://doi.org/10.1111/1556-4029.14931>]
- Jantz RL, Ousley SD (2005) *FORDISC 3.0: Personal computer forensic discriminant functions*. University of Tennessee, Knoxville
- Knecht S, Santos F, Ardagna Y et al. (2023) Sex estimation from long bones: a machine learning approach. *International Journal of Legal Medicine* [<https://doi.org/10.1007/s00414-023-03072-4>]
- Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15(2):155-163 [<https://doi.org/10.1016/j.jcm.2016.02.012>]
- Krüger GC, L'Abbé EN, Stull KE (2017) Sex estimation from the long bones of modern South Africans. *International Journal of Legal Medicine* 131(1):275-285 [<https://doi.org/10.1007/s00414-016-1488-z>]
- Langley NR, Meadows Jantz L, McNulty S et al. (2018) Error quantification of osteometric data in forensic anthropology. *Forensic Science International* 287:183-189 [<https://doi.org/10.1016/j.forsciint.2018.04.004>]
- Martin R (1928) *Lehrbuch der Anthropologie in systematischer Darstellung mit besonderer Berücksichtigung der Anthropologischen Methoden, für Studierende, Ärzte und Forschungsreisende, Zweiter Band: Kraniologie, Osteologie*. Gustav Fischer Ed, Jena, 3 vol., 1805 p.
- Mony P, Swaminathan S, Gajendran J et al. (2016) Quality assurance for accuracy of anthropometric measurements in clinical and epidemiological studies [Errare humanum est = to err is human]. *Indian Journal of Community Medicine* 41(2):98 [<https://doi.org/10.4103/0970-0218.173499>]
- Norton K., Olds T. (Eds.) (1996) *Anthropometrica: a textbook of body measurement for sports and health courses*. UNSW press.
- Perini TA, Oliveira GLD, Ornellas JDS et al. (2005) Cálculo do erro técnico de medição em antropometria. *Revista Brasileira de Medicina do Esporte* 11(1):81-85 [<https://doi.org/10.1590/S1517-86922005000100009>]
- R Core Team A., R Core Team (2022) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. 2012
- Santos F (2020) Modern methods for old data: An overview of some robust methods for outliers detection with applications in osteology. *Journal of Archaeological Science: Reports* 32: 102423 [<https://doi.org/10.1016/j.jasrep.2020.102423>]
- Simon S, Fischer B, Rinner A et al. (2023) Body height estimation from automated length measurements on standing long leg radiographs using artificial intelligence. *Scientific Reports* 13(1): 8504 [<https://doi.org/10.1038/s41598-023-34670-2>]
- Spradley MK, Jantz RL (2011) Sex estimation in forensic anthropology: skull versus postcranial elements. *Journal of Forensic Sciences* 56(2):289-296 [<https://doi.org/10.1111/j.1556-4029.2010.01635.x>]
- Stock MK (2020) Analyses of the postcranial skeleton for sex estimation. In: *Sex Estimation of the Human Skeleton*. Elsevier, pp. 113-130 [<https://doi.org/10.1016/B978-0-12-815767-1.00008-0>]
- Tanner-Smith EE, Durlak JA, Marx RA (2018) Empirically Based Mean Effect Size Distributions for Universal Prevention Programs Targeting School-Aged Youth: A Review of Meta-Analyses. *Prevention Science* 19(8):1091-1101 [<https://doi.org/10.1007/s11121-018-0942-1>]
- Yadav S, Nayak VC, Palimar V et al. (2023) An Estimation of Stature and Sex using Various Anthropometric Parameters in South Indian Population. *Journal of Forensic Science and Medicine* 9(2):130-136 [https://doi.org/10.4103/jfsm.jfsm_33_22]

Supplementary Information

Supplementary information 1 (SI1): Original data collected “OLIVIER.csv”

Supplementary information 2 (SI2): Data cleaned “OLIVIER_CLEANED.csv”

Supplementary information 3 (SI3): Barbus (inter-observer) original data “BARBUS_INTER.csv”

Supplementary information 4 (SI4): Barbus (inter-observer) cleaned data “BARBUS_INTER_CLEANED.csv”

Supplementary information 5 (SI5): Descriptive statistics of the individuals studied.