

Observer repeatability and inter-rater reproducibility of DXA-derived TW3-RUS skeletal age assessment in elite youth football players: a complementary framework for evaluating measurement quality

original paper

© Wrocław University of Health and Sport Sciences

DOI: <https://doi.org/10.5114/hm/231285>

MIROSLAV GROBAR¹ , MARTIN MUSALEK² , PETR STASTNY¹ , JINDRICH VAMPOLA¹ ,
JAKUB KOKSTEJN¹ 

¹ Department of Sport Games, Faculty of Physical Education and Sport, Charles University, Prague, Czech Republic² Department of Social Science Foundation in Kinanthropology, Faculty of Physical Education and Sport, Charles University, Prague, Czech Republic

ABSTRACT

Purpose. The purpose of the study was to evaluate observer repeatability and inter-rater reproducibility of skeletal age assessment using the Tanner–Whitehouse 3 radius, ulna, and short bones (TW3-RUS) method on dual-energy X-ray absorptiometry (DXA) images, and to apply a complementary framework for evaluating observer-dependent measurement quality.

Methods. Seventy-one elite male youth football players from the Czech Republic were assessed at two time points separated by one year. Skeletal age was estimated from DXA hand scans by an expert and a trained novice rater using the TW3-RUS method. Both observers repeated all assessments after two weeks. Measurement quality was evaluated using the intraclass correlation coefficient $ICC(2,1)$, standard error of measurement (SEM), Bland–Altman analysis with proportional bias testing, bootstrap confidence intervals, and a predefined ± 0.5 -year decision threshold.

Results. Observer repeatability and inter-rater reproducibility were excellent at both annual assessments ($ICC = 0.965–0.999$). The expert maintained consistently low measurement error ($SEM < 1$ month), whereas the trained novice reduced SEM from 3.7 to 2.5 months after one year of practical experience. Bland–Altman analyses demonstrated only minor systematic differences and no meaningful proportional bias in the overall sample, whereas only a minority of observations exceeded the predefined decision threshold.

Conclusions. DXA-derived TW3-RUS skeletal age assessment provides excellent repeatability and reproducibility under standardised conditions. Observer experience primarily improves measurement precision rather than ranking consistency. The principal contribution of this study is a complementary validation framework integrating repeatability, reproducibility, precision, agreement, and decision-oriented analyses, providing a methodological basis for evaluating observer-dependent assessment methods in sport science.

Key words: growth and development, observer variation, absorptiometry, adolescent, soccer

Introduction

Interindividual variability in biological maturation is a major source of developmental heterogeneity in studies involving growing children and adolescents [1]. During childhood, and particularly during adolescence, individuals of the same chronological age may differ markedly in skeletal maturation, growth tempo, body size, physical performance, and physiological development [2, 3]. These differences become especially pronounced around peak height velocity (PHV), when mat-

uration progresses asynchronously across individuals [1, 4]. Consequently, chronological age alone inadequately characterises developmental status and may confound developmental, clinical, and sport-related research [3, 5]. Biological age has, therefore, become an essential methodological variable for distinguishing maturation-related variability from chronological-age effects and for improving the interpretation of growth, physical performance, fitness assessment, physical activity interventions, health, and developmental outcomes [1, 3, 5–7].

Correspondence address: Jakub Kokstejn, Department of Sport Games, Faculty of Physical Education and Sport, Charles University, Jose Martího 269/31, 16252 Prague 6, Czech Republic, e-mail: kokstejn@ftvs.cuni.cz; <https://orcid.org/0000-0002-8046-5830>

Received: May 29, 2026

Accepted for publication: August 18, 2026

Citation: Grobar M, Musalek M, Stastny P, Vampola J, Kokstejn J. Observer repeatability and inter-rater reproducibility of DXA-derived TW3-RUS skeletal age assessment in elite youth football players: a complementary framework for evaluating measurement quality. Hum Mov. 2026;27(4):33–46; doi: <https://doi.org/10.5114/hm/231285>.

Despite the widespread use of skeletal age in developmental, clinical, and sport-related research, the methodological quality of repeated skeletal age assessment remains incompletely understood. Previous studies have generally evaluated inter-rater reproducibility without first establishing observer repeatability and have relied predominantly on intraclass correlation coefficients (*ICC*), thus providing only a partial characterisation of measurement quality. As a result, it remains unclear whether reported reproducibility reflects true between-observer agreement or variability introduced by repeated assessments within individual observers.

Among the available indicators of biological maturity, skeletal age is considered the reference standard because it directly reflects skeletal ossification and maturation [8–10]. Several standardised methods have been developed, including the Greulich–Pyle, Fels, Tanner–Whitehouse 3 (TW3), and Thiemann–Nitz methods [11–13]. Of these, the TW3 method is widely used in both clinical and research settings because of its standardised scoring system, broad international acceptance, and ability to predict adult stature [12, 14, 15].

Hand-wrist radiographs remain the reference imaging modality for TW3 assessment [16]. However, repeated radiographic examinations are less desirable in longitudinal studies because of cumulative ionising radiation exposure, particularly in healthy paediatric populations [17]. Dual-energy X-ray absorptiometry (DXA) provides hand images suitable for TW3 assessment while substantially reducing radiation exposure [18, 19]. Because DXA is widely used to assess body composition and bone health in clinical, research, and sport-science settings, skeletal age can be evaluated within the same DXA-based assessment framework without an additional radiographic examination, thereby increasing research efficiency and clinical feasibility [20, 21]. Validation studies have consistently demonstrated excellent agreement between DXA-derived hand scans and conventional radiographs ($ICC \approx 0.97\text{--}0.99$), establishing DXA as an accurate and practical method for repeated skeletal age assessment [16, 20, 22].

As DXA-derived hand scans are increasingly used in longitudinal studies, methodological attention has shifted from image validity to measurement quality. Observer repeatability and inter-rater reproducibility represent distinct sources of measurement variability and should be evaluated sequentially, because between-observer differences cannot be interpreted independently of within-observer variability [23–26]. This issue becomes particularly relevant during adolescence, when rapid skeletal maturation is accompanied by increas-

ingly subtle morphological changes in the hand and wrist. Transitional ossification stages may be more difficult to classify consistently, increasing observer variability and amplifying differences between experienced and novice raters. Nevertheless, longitudinal evidence simultaneously evaluating observer repeatability and subsequent inter-rater reproducibility across adolescence remains scarce.

The present study addresses this methodological gap by establishing observer repeatability before evaluating inter-rater reproducibility and by integrating complementary measures of relative reproducibility, measurement precision, agreement, and practical decision-making within a single analytical framework. Accordingly, the aim of this study was to comprehensively evaluate observer repeatability and inter-rater reproducibility of DXA-derived TW3-RUS skeletal age assessment performed by an expert and a trained novice rater in adolescent football players. Observer repeatability was first established as the methodological prerequisite for interpreting inter-rater reproducibility. Relative reproducibility (*ICC*), measurement precision (*SEM*), agreement (Bland–Altman analysis), bootstrap-based comparisons, and decision-threshold analysis were subsequently integrated to characterise complementary dimensions of measurement quality across two annual assessments spanning a period of rapid biological maturation.

Material and methods

Study design

This longitudinal study evaluated observer repeatability and inter-rater reproducibility of DXA-derived TW3-RUS skeletal age assessment across two annual assessments separated by one year (Figure S1 in Supplementary materials). Measurements were performed in the same month (February) of 2023 and 2024 to minimise seasonal variation in growth-related variables. All assessments were conducted under standardised laboratory conditions using identical equipment, acquisition protocols, image-display settings, and environmental conditions.

Raters and scoring procedure

Two raters independently evaluated all DXA-derived hand-wrist images in a randomised order using the TW3-RUS method. The expert rater (RE) had prior experience in assessing 1,036 conventional hand-wrist radiographs and 647 DXA-derived hand scans using

the TW3 method. The second rater (novice rater, RN) was a PhD student with no previous practical experience in TW3 assessment prior to the training. Before formal data collection, the novice completed a structured two-month TW3-RUS training programme supervised by the expert rater. Training comprised atlas familiarisation, supervised joint scoring, two independent scoring rounds of 30 DXA hand scans separated by a four-week washout period, and structured feedback following each round. This standardised training protocol was designed to ensure consistent application of the TW3-RUS scoring criteria before the formal assessment. Accordingly, RN is referred to throughout the manuscript as the trained novice rater. At each annual assessment, both raters independently evaluated skeletal age from all DXA hand scans under blinded conditions. Raters were blinded to participants' chronological age, anthropometric characteristics, competitive age category, training-related information, and each other's evaluations. To assess observer repeatability, all scans were re-evaluated independently by both raters after a two-week interval, under identical anonymised scoring conditions. The same protocol was applied at both annual assessments.

Subjects

The study included 71 elite male football players from a professional academy representing the under-12 to under-15 (U12–U15) age categories. Descriptive characteristics at the first annual assessment (T1) are presented in Table 1. Eligible participants had completed at least three years of organised football training, attended both annual assessments, and provided written informed consent from a parent or legal guardian. Players were excluded if they missed any assessment or reported using dietary supplements known to influence biological maturation. Body height and body mass were measured according to standardised International Society for the Advancement of Kinanthropometry

(ISAK) procedures using calibrated equipment (Seca 217 stadiometer; Seca 899 scale; Seca, Hamburg, Germany) with a precision of 0.1 cm and 0.1 kg, respectively.

DXA image acquisition

DXA-derived hand-wrist images were acquired from the non-dominant hand using a Horizon-A densitometer (Hologic Inc., Marlborough, MA, USA) operated by a single trained examiner. Scans were performed in single-energy acquisition mode following a standardised protocol described previously [22, 27]. Participants were scanned in a seated position with the non-dominant hand placed flat on the scanning table. Scan parameters were 100 kV, 0.188 mA, an acquisition time of approximately 69 s, and a maximum entrance surface dose of 1.7 mGy. Image-display settings were standardised across all assessments (brightness 2210; contrast 685; Supplementary Figure S2). DXA acquisition parameters, image-display settings, and subsequent image processing were kept identical throughout the study to minimise measurement variability [28].

TW3-RUS skeletal age assessment

Skeletal age was assessed from DXA-derived hand-wrist images using the Tanner–Whitehouse 3 Radius–Ulna–Short bones (TW3-RUS) method [14]. The method evaluates the developmental stage of the 13 RUS bones, with each bone being assigned a standardised maturity stage and a corresponding numerical score. Skeletal maturity stages were entered into the RUS Bone Age software (version 1.0; AlphaBridge Computers, 1999), which is distributed with the TW3 manual by Tanner *et al.* [14]. The software automatically calculated the total RUS score and converted it into skeletal age according to the TW3 reference standards.

All skeletal age assessments were performed using identical software settings throughout the study.

Statistical analysis

All statistical analyses were performed in Python (version 3.11) using the Pandas, NumPy, SciPy, StatsModels, and Pingouin libraries. An *a priori* power analysis was conducted in G*Power 3.1 [29]. Based on $ICC(2,1)$ with $\rho_0 = 0.85$, $\rho_1 = 0.95$, $\alpha = 0.05$, and $1-\beta = 0.90$, a minimum sample of 28 participants was required for the reproducibility analyses. The lower threshold of $\rho_0 = 0.85$ was selected *a priori* as a conservative minimum acceptable reliability for applied skeletal age assessment, whereas $\rho_1 = 0.95$ was chosen to represent excel-

Table 1. Descriptive anthropometric characteristics of the study sample at the first annual assessment (T1)

Category	<i>n</i>	Age (years) mean $\pm$ SD	Height (cm) mean $\pm$ SD	Weight (kg) mean $\pm$ SD
U12	21	11.86 $\pm$ 0.27	150.13 $\pm$ 4.28	39.33 $\pm$ 3.41
U13	21	12.89 $\pm$ 0.27	162.43 $\pm$ 9.16	48.88 $\pm$ 7.41
U14	14	13.89 $\pm$ 0.47	169.08 $\pm$ 8.94	56.54 $\pm$ 10.18
U15	15	14.64 $\pm$ 0.46	175.13 $\pm$ 5.20	63.88 $\pm$ 6.50

U12–U15 – under-12 to under-15 age categories

Age refers to chronological age at T1.

lent reliability expected under standardised TW3-RUS scoring conditions, consistent with established *ICC* interpretation guidelines [30]. For the regression analyses, an *F*-test assuming one predictor, $\alpha = 0.05$, $1-\beta = 0.80$, and an effect size of $f^2 = 0.10$ indicated a required sample of 67 participants. The final sample of 71 players exceeded both requirements.

Analyses followed a predefined hierarchical framework in which observer repeatability was established before inter-rater reproducibility. Observer repeatability was evaluated from repeated within-rater assessments performed two weeks apart under identical blinded conditions. Inter-rater reproducibility was subsequently evaluated using the corresponding assessments of the expert and trained novice raters. Relative reproducibility was quantified using the intraclass correlation coefficient [*ICC*(2,1), two-way random-effects model with absolute agreement] [30, 31]. Absolute measurement precision was quantified using the standard error of measurement (*SEM*) [32]. Agreement was evaluated using Bland–Altman analysis, including estimation of the mean bias, 95% limits of agreement (LoA), and proportional bias [33, 34]. *ICC*, *SEM*, and Bland–Altman analyses were interpreted as complementary measures describing relative reproducibility, measurement precision, and agreement, respectively. To quantify uncertainty, 95% confidence intervals for *ICC* and *SEM* were estimated using non-parametric bootstrap resampling (5,000 iterations) [35]. All analyses were performed for the overall sample, and separately by annual assessment and chronological age category. Bootstrap resampling was additionally used to compare observer repeatability between the expert and trained novice raters.

To examine whether inter-rater disagreement varied with age, linear regression and Pearson correlation analyses were performed using absolute inter-rater differences ($|RN - RE|$) as the dependent variable. Finally, a predefined, literature-informed threshold of ± 0.5 years [36, 37] was applied, because discrepancies of this magnitude may shift an individual across a skeletal-age

boundary and affect the interpretation of skeletal maturity. Cases exceeding this threshold were examined individually. Statistical significance was set at $p < 0.05$.

Results

Results are presented according to the predefined analytical framework. Observer repeatability was evaluated first, because inter-rater reproducibility can be interpreted only when repeated assessments within each observer are demonstrably stable. Subsequent analyses examined inter-rater reproducibility and agreement, the influence of chronological age, and the practical significance of observer disagreement using the predefined decision threshold.

Observer repeatability as the methodological basis for inter-rater reproducibility

Observer repeatability was first established independently for each rater as the methodological prerequisite for evaluating inter-rater reproducibility. Complementary analyses of relative repeatability (*ICC*), measurement precision (*SEM*), and agreement (Bland–Altman analysis) converged on the same conclusion, providing consistent evidence of high measurement quality. Observer repeatability results are summarised in Tables 2 and 3, with bootstrap comparisons between raters presented in Table S1 of Supplementary materials.

Overall observer repeatability

Observer repeatability was excellent for both raters at both annual assessments (Tables 2 and 3), confirming that repeated skeletal age evaluations were internally stable regardless of observer experience. However, the pattern of repeatability differed between raters. Whereas the expert demonstrated stable measurement performance across both years, the trained novice exhibited a clear improvement in measurement precision while maintaining excellent relative repeatability. Rela-

Table 2. Observer repeatability of the trained novice rater at T1 and T2

Group	<i>n</i>	T1 <i>ICC</i> (95% <i>CI</i>)	T1 <i>SEM</i> (years)	T2 <i>ICC</i> (95% <i>CI</i>)	T2 <i>SEM</i> (years)
Overall	71	0.965 (0.946–0.978)	0.310	0.982 (0.962–0.993)	0.209
U12/U13	21	0.966 (0.872–0.982)	0.272	0.982 (0.949–0.993)	0.210
U13/U14	21	0.904 (0.796–0.955)	0.337	0.911 (0.672–0.987)	0.288
U14/U15	14	0.960 (0.870–0.984)	0.268	0.995 (0.980–0.998)	0.080
U15/U16	15	0.960 (0.859–0.992)	0.342	0.992 (0.954–0.999)	0.138

U12–U16 – under-12 to under-16 age categories, T1 – first annual assessment, T2 – second annual assessment
Paired group labels show the age category at T1 followed by that at T2.

Table 3. Observer repeatability of the expert rater at T1 and T2

Group	<i>n</i>	T1 <i>ICC</i> (95% <i>CI</i>)	T1 <i>SEM</i> (years)	T2 <i>ICC</i> (95% <i>CI</i>)	T2 <i>SEM</i> (years)
Overall	71	0.999 (0.998–0.999)	0.057	0.999 (0.998–0.999)	0.057
U12/U13	21	0.993 (0.977–0.998)	0.065	0.998 (0.994–0.999)	0.068
U13/U14	21	0.996 (0.990–0.999)	0.065	0.998 (0.994–0.999)	0.055
U14/U15	14	0.999 (0.996–1.000)	0.041	0.999 (0.995–0.999)	0.046
U15/U16	15	0.999 (0.997–1.000)	0.041	0.999 (0.996–1.000)	0.051

U12–U16 – under-12 to under-16 age categories, T1 – first annual assessment, T2 – second annual assessment
Paired group labels show the age category at T1 followed by that at T2.

tive repeatability remained consistently high throughout the study. For the trained novice, *ICC* increased from 0.965 (95% *CI*: 0.946–0.978) at the first annual assessment (T1) to 0.982 (95% *CI*: 0.962–0.993) at the second assessment (T2), while the expert maintained near-perfect repeatability at both assessments (*ICC* $\approx$ 0.999). These findings indicate that repeated assessments produced highly consistent rankings of skeletal maturity irrespective of observer experience. The most pronounced longitudinal change occurred in absolute measurement precision. The trained novice reduced the *SEM* from 0.310 years (3.7 months) at T1 to 0.209 years (2.5 months) at T2, corresponding to an approximately 33% reduction in expected measurement error. In contrast, the expert maintained an *SEM* of approximately 0.057 years (0.7 months) at both assessments, indicating consistently high precision of repeated measurements.

Taken together, these findings show that observer experience primarily influenced measurement precision rather than ranking consistency. The expert achieved maximal precision from the outset, whereas the trained novice significantly improved the precision level after one year of practical experience. This pattern indicates that observer training predominantly improves the precision of skeletal age estimation, rather than the consistency with which individuals are ranked according to skeletal maturity.

These findings demonstrate that excellent ranking consistency can coexist with meaningful differences in measurement precision, highlighting the complementary value of *ICC* and *SEM* for evaluating skeletal age assessment.

Influence of observer experience

Age-specific analyses confirmed that observer repeatability remained high in all chronological age categories (Tables 2 and 3). The expert maintained uniformly low measurement error across all groups, whereas

the trained novice demonstrated the greatest gains in measurement precision, in the older age categories, during the second assessment. Bootstrap comparisons reinforced this pattern (Table S1 in Supplementary materials). Overall measurement precision differed by 3.04 months between raters at T1 but only by 1.82 months at T2, indicating a substantial reduction in the precision gap following one year of practical experience. In contrast, differences in relative repeatability remained consistently small ($\Delta ICC = 0.034$ at T1; $\Delta ICC = 0.017$ at T2), indicating that observer experience had little influence on the relative ordering of skeletal maturity.

Agreement analysis

Agreement analyses independently corroborated the repeatability findings (Figures S3–S4 in Supplementary Materials). Mean differences remained close to zero at both annual assessments, indicating no systematic tendency for either observer to overestimate or underestimate skeletal age during repeated evaluations. The limits of agreement provided greater discrimination between observers than mean bias alone. At T1, the trained novice's repeated assessments varied by approximately –8 to +12 months, compared with approximately ± 2 months for the expert. At T2, the limits of agreement became visibly narrower, demonstrating improved repeatability following one year of practical scoring experience. No proportional bias was detected for either rater in the overall sample at either annual assessment. An isolated age-specific effect was observed for the expert in the U14 category at T2, without evidence of a consistent age-related pattern across analyses.

Integrated summary

Taken together, all complementary analyses converged on the same conclusion. Observer repeatability

was sufficiently robust to support subsequent evaluation of inter-rater reproducibility. Observer experience primarily influenced the precision of skeletal age estimation rather than the consistency with which individuals were ranked according to skeletal maturity. Consequently, any disagreement observed between raters can be interpreted as reflecting genuine between-observer differences rather than instability of repeated measurements within individual observers.

Inter-rater reproducibility between observers

Inter-rater reproducibility was evaluated after establishing robust observer repeatability for both raters.

Complementary analyses of relative reproducibility (*ICC*), measurement precision (*SEM*), and agreement consistently demonstrated high reproducibility of skeletal age assessment between observers (Tables 4 and 5; Figures 1 and 2).

Overall inter-rater reproducibility

Inter-rater reproducibility remained high at both annual assessments (Table 4), indicating that skeletal age estimates were highly comparable despite substantial differences in observer experience. Observers ranked skeletal maturity almost identically at both assessments, with *ICC*(2,1) values of 0.978 (95% *CI*: 0.969–

Table 4. Inter-rater reproducibility and agreement between the trained novice rater and the expert rater

Assessment	<i>ICC</i> (2,1)	95% <i>CI</i>	<i>SEM</i> (years)	<i>SEM</i> (months)	Bias RN – RE (years)	95% LoA (years)
T1	0.978	0.969–0.985	0.241	2.89	0.152	–0.454 to 0.758
T2	0.977	0.965–0.984	0.243	2.92	0.208	–0.338 to 0.753

LoA – limits of agreement, RN – trained novice rater, RE – expert rater

Bias was calculated as RN – RE, positive values indicate higher skeletal age estimates by the trained novice rater.

Table 5. Age-specific inter-rater reproducibility stratified by chronological age category

Group	<i>n</i>	T1 <i>ICC</i> (2,1)	T1 95% <i>CI</i>	T1 <i>SEM</i> (years)	T2 <i>ICC</i> (2,1)	T2 95% <i>CI</i>	T2 <i>SEM</i> (years)
U12/U13	21	0.904	0.785–0.945	0.214	0.921	0.831–0.961	0.273
U13/U14	21	0.955	0.899–0.978	0.216	0.950	0.859–0.980	0.214
U14/U15	14	0.958	0.904–0.976	0.265	0.945	0.857–0.969	0.247
U15/U16	15	0.949	0.850–0.978	0.270	0.951	0.870–0.974	0.217

U12–U16 – under-12 to under-16 age categories, T1 – first annual assessment, T2 – second annual assessment

Paired group labels show the age category at T1 followed by that at T2.

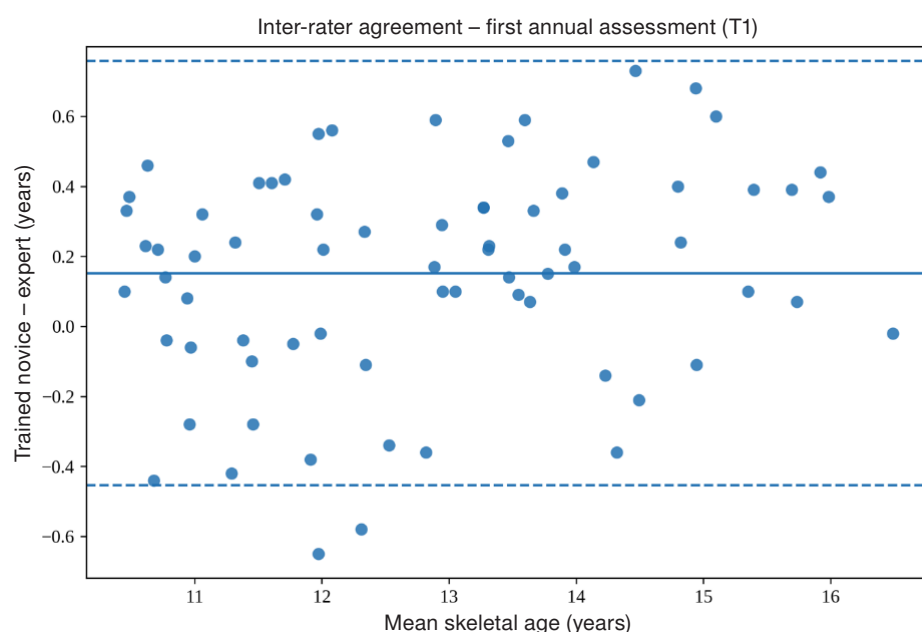


Figure 1. Bland–Altman plot of inter-rater differences at T1

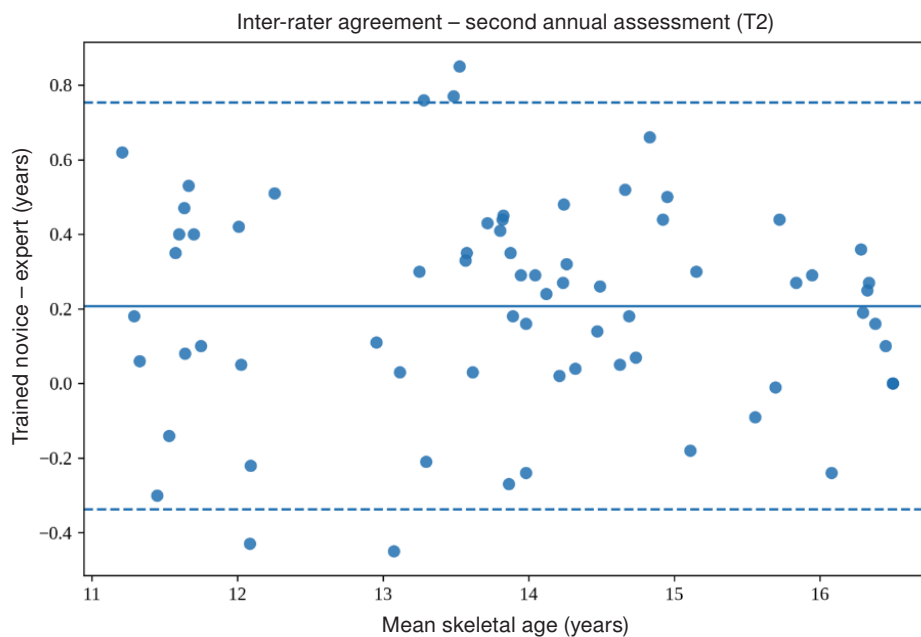


Figure 2. Bland–Altman plot of inter-rater differences at T2

0.985) at T1 and 0.977 (95% *CI*: 0.965–0.984) at T2. Between-observer measurement precision was descriptively similar at T1 and T2, with *SEM* values of 0.241 years (2.9 months) and 0.243 years (2.9 months), respectively. To summarise, the *ICC* and *SEM* estimates indicated similar inter-rater measurement performance across the two annual assessments.

Age-specific inter-rater reproducibility

Inter-rater reproducibility remained consistently high across all age categories (Table 5). Observers maintained highly consistent rankings of skeletal maturity throughout childhood and adolescence, with *ICC* values ranging from 0.904 to 0.958 across both annual assessments. No age category showed consistently poorer inter-rater reproducibility (*SEM*: 0.214–0.270 years at T1; 0.214–0.273 years at T2). Overall, chronological age did not meaningfully influence inter-rater reproducibility, indicating that agreement between observers was maintained across the full age range studied.

Agreement analysis

Agreement analyses further characterised the nature of inter-rater differences rather than their magnitude alone (Figures 1 and 2). The trained novice assigned slightly higher skeletal age estimates than the expert, with mean differences of 0.152 years at T1 and 0.208 years at T2, indicating a small but consistent positive bias. Observers differed systematically by only about two months on average. The 95% limits of agree-

ment ranged from –0.454 to 0.758 years at T1 and from –0.338 to 0.753 years at T2, demonstrating comparable between-observer variability at both annual assessments. Proportional bias was not detected in the overall sample at either assessment (T1: $B = 0.043$, $p = 0.057$; T2: $B = -0.007$, $p = 0.746$). Significant proportional bias was observed only in the youngest age group at T1 and the oldest age group at T2. In summary, the agreement analyses demonstrated that between-observer differences were small, largely unbiased, and independent of skeletal maturity across the study population.

Influence of chronological age

Chronological age did not explain between-observer disagreement (Table 6). Linear regression showed no significant association between chronological age and absolute inter-rater discrepancies at either annual assessment (T1: $B = 0.037$, $R^2 = 0.051$, $p = 0.058$; T2: $B =$

Table 6. Linear regression and Pearson correlation between chronological age and absolute inter-rater discrepancies

Assessment	<i>n</i>	<i>B</i> (<i>SE</i>)	R^2	<i>p</i>	Pearson <i>r</i>	<i>p</i>
T1	71	0.037 (0.019)	0.051	0.058	0.226	0.058
T2	71	–0.009 (0.021)	0.003	0.672	–0.051	0.672

T1 – first annual assessment

T2 – second annual assessment

Absolute inter-rater discrepancy was calculated as $|RN - RE|$.

-0.009 , $R^2 = 0.003$, $p = 0.672$). Pearson correlation analyses led to the same conclusion (T1: $r = 0.226$, $p = 0.058$; T2: $r = -0.051$, $p = 0.672$). These findings indicate that disagreement between observers was not systematically related to participant age.

Decision-threshold analysis

Only a small proportion of assessments exceeded the predefined ± 0.5 -year decision threshold (Table 7). Ten of 71 observations (14.1%) exceeded the threshold at T1, compared with eight (11.3%) at T2. Discrepancies ranged from 0.53 to 0.73 years at T1 and from 0.51 to 0.85 years at T2. At T2, all threshold exceedances were associated with higher skeletal age estimates assigned by the trained novice. Overall, disagreements large enough to influence biological-age classification were uncommon and became slightly less frequent after one year of additional observer experience.

Table 7. Cases exceeding the predefined ± 0.5 -year decision threshold

Measure	T1	T2
Cases exceeding threshold, n (%)	10 (14.1)	8 (11.3)
Direction of exceedances (positive / negative)	8 / 2	8 / 0
Range of discrepancy, years	0.53–0.73	0.51–0.85
Most affected category	U13, U14, U15 ($n = 3$ each)	U13 ($n = 5$)

U13–U15 – under-13 to under-15 age categories,
T1 – first annual assessment, T2 – second annual assessment

Positive discrepancies indicate higher skeletal age estimates by the trained novice rater.

Taken together, all complementary analyses converged on the same conclusion. Inter-rater reproducibility remained robust despite substantial differences in observer experience, with both observers producing highly comparable skeletal age estimates. Observer experience primarily influenced the precision of absolute skeletal age estimates, rather than the consistency with which individuals were ranked according to skeletal maturity.

Discussion

The present study demonstrates that DXA-derived TW3-RUS skeletal age assessment provides robust observer repeatability and inter-rater reproducibility

under standardised conditions. More importantly, complementary analyses showed that observer experience influenced primarily the precision of skeletal age estimation, rather than the consistency with which individuals were ranked according to skeletal maturity. These findings extend current evidence by demonstrating the value of integrating relative and absolute indices of measurement quality within a single analytical framework.

The principal methodological contribution of the present study is the use of a hierarchical framework in which observer repeatability was established before inter-rater reproducibility was interpreted. This sequence is fundamental because agreement between observers cannot be meaningfully evaluated unless repeated measurements within each observer are first shown to be internally stable. Although this principle is widely recognised in measurement science, previous studies of skeletal age assessment have typically focused on inter-rater reliability alone, thereby providing limited information about whether observed between-rater differences reflect true disagreement or instability of repeated measurements within individual raters.

The present findings further demonstrate that no single statistical index adequately characterises the quality of skeletal age assessment. Excellent *ICC* values were observed for both observer repeatability and inter-rater reproducibility, indicating highly consistent ranking of skeletal maturity. However, complementary analyses showed that substantial differences in measurement precision persisted despite similarly high *ICC* values, particularly between the expert and the trained novice rater during the first assessment. Likewise, Bland–Altman analyses identified differences in the magnitude and distribution of measurement error that could not be inferred from correlation coefficients alone, while the decision-threshold analysis translated these statistical findings into their potential practical consequences for biological-age classification.

These findings support a complementary approach to evaluating skeletal age assessment, in which observer repeatability, inter-rater reproducibility, absolute measurement precision, agreement, and decision-oriented analyses are interpreted jointly rather than as independent outcomes. Such an integrated framework provides a more comprehensive characterisation of measurement quality than reliance on *ICC* alone and may be applicable to other observer-dependent methods used for biological maturity assessment.

The present findings provide a more nuanced interpretation of observer experience than is commonly reported in skeletal age research. Structured training

and accumulated scoring experience did not materially alter the relative ranking of skeletal maturity, which remained consistently excellent for both raters throughout the study. Instead, they mainly improved the precision of absolute skeletal age estimation. Following one year of practical experience, the trained novice reduced measurement error by approximately 33%, whereas the expert's measurement error remained consistently low across both annual assessments. These findings suggest that observer experience primarily reduces the magnitude of scoring error, rather than altering the overall reproducibility of skeletal maturity classification.

This interpretation is consistent with the principles of the TW3 method, which relies on standardised morphological criteria and reference illustrations to support consistent stage allocation across observers [14]. At the same time, previous studies have shown that structured training, expert supervision, and repeated exposure to skeletal age assessment are essential for achieving high-quality scoring, particularly when evaluating subtle transitional stages of ossification [38, 39]. Similarly, studies using DXA-derived hand scans have reported excellent agreement between experienced observers, supporting the robustness of the method under standardised conditions [20, 22]. The present study extends this evidence by demonstrating that improvements associated with observer experience primarily reflect greater absolute measurement precision, rather than improved ranking consistency. This distinction has practical implications because observers with different levels of experience may classify individuals similarly in relative terms, yet differ in the expected measurement-error magnitudes, particularly when skeletal age estimates are used for longitudinal monitoring or for individual clinical and sport-related decision-making.

The present findings have several practical implications for the longitudinal assessment of skeletal maturity in research, clinical practice, and youth sport. First, the absence of a significant relationship between chronological age and inter-rater disagreement suggests that, when standardised procedures are applied, DXA-derived TW3-RUS assessment can be used consistently across the adolescent age range without introducing systematic age-related measurement error. This finding supports previous recommendations that biological maturity should complement, rather than be inferred from, chronological age when monitoring growth and development in young athletes [1, 5]. Second, although overall inter-rater agreement was excellent, the decision-threshold analysis demonstrated that a small proportion of assessments exceeded the predefined

± 0.5 -year threshold. These cases were concentrated around transitional stages of skeletal maturation, where small scoring differences may influence biological-age classification and, consequently, decisions regarding longitudinal monitoring, maturity grouping, or interpretation of developmental trajectories. Similar challenges have previously been reported for late-stage ossification of the radius and ulna, which are among the most demanding indicators within the TW3-RUS system [14, 15, 40]. Therefore, practical quality assurance should extend beyond initial observer training and include periodic calibration against an experienced reference observer, particularly when skeletal age assessment is performed by multiple raters or over prolonged longitudinal studies. Targeted calibration sessions using borderline and transitional cases, together with regular review of radius and ulna staging, are likely to improve agreement and reduce clinically or practically meaningful classification differences. These recommendations complement previous reports emphasising the importance of standardised image acquisition, structured TW3 training, and continuous quality control for maintaining high-quality skeletal age assessment [20, 22, 27, 38, 39].

Future research in sport science should prioritise the development of a unified validation framework for observer-dependent assessment methods, including biological maturity assessment, movement test batteries, and questionnaire-based instruments. Such a framework should integrate indices of relative and absolute reliability, systematic and proportional bias, observer experience, and clinically or practically meaningful thresholds to facilitate transparent interpretation and comparison across methods and populations.

Limitations

Several limitations should be acknowledged. First, the study included only two observers representing different levels of experience; therefore, the findings should not be generalised to larger observer panels without further investigation. Second, the sample consisted exclusively of elite male youth football players, limiting direct generalisation to females, non-athletic populations, or younger children. Third, the study evaluated DXA-derived TW3-RUS skeletal age assessment and does not necessarily extend to conventional radiographic imaging or alternative skeletal age methods. Finally, the analyses focused on overall measurement performance and did not quantify observer disagreement at the level of individual TW3-RUS indicators or specific bones. Such analyses may provide additional

insight into the anatomical sources of observer disagreement and help to further optimise training procedures.

Conclusions

DXA-derived TW3-RUS skeletal age assessment provides excellent observer repeatability and inter-rater reproducibility under standardised conditions. Observer experience primarily improves measurement precision rather than ranking consistency. Most importantly, this study demonstrates that observer-dependent assessment methods should not be validated solely on the basis of ICC but rather through a complementary framework that integrates repeatability, reproducibility, precision, agreement, and decision-oriented analyses. This framework provides a stronger methodological basis for skeletal age assessment and for other observer-dependent methods in sport science.

Detailed description of the use of AI

During manuscript preparation, ChatGPT (OpenAI) was used to assist with English-language editing, stylistic refinement, and improvement of textual clarity. All AI-assisted outputs were critically reviewed, verified, and revised by the authors. The authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Data availability statement

The anonymised dataset and Python scripts used to generate the main statistical analyses are available in a public repository: <https://zenodo.org/records/21788814>. Raw DXA images are not publicly available because they may contain potentially identifiable health-related information.

Ethical approval

The research related to human use complied with all the relevant national regulations and institutional policies, followed the tenets of the Declaration of Helsinki, and was approved by the Ethics Committee of Charles University, Faculty of Physical Education and Sport (approval No.: 253/2022).

Informed consent

Informed consent was obtained from all individuals included in this study. All participants provided written informed consent prior to participation, explicitly agreeing to take part in the study and authorising the use of their anonymised data for scientific publication purposes.

Conflict of interest

The authors state no conflict of interest.

Disclosure statement

No author has any financial interest or received any financial benefit from this research.

Funding

This work was supported by the Charles University Research Centre program No. UNCE24/SSH/012 and Cooperation programs (Sport Sciences – Biomedical & Rehabilitation Medicine; Sport & Social).

References

- [1] Cumming SP, Lloyd RS, Oliver JL, Eisenmann JC, Malina RM. Bio-banding in sport: applications to competition, talent identification, and strength and conditioning of youth athletes. *Strength Cond J*. 2017;39(2):34–47; doi: 10.1519/SSC.00000000000000281.
- [2] Lloyd RS, Cronin JB, Faigenbaum AD, Haff GG, Howard R, Kraemer WJ, Micheli LJ, Myer GD, Oliver JL. National Strength and Conditioning Association position statement on long-term athletic development. *J Strength Cond Res*. 2016;30(6): 1491–509; doi: 10.1519/JSC.00000000000001387.
- [3] Pichardo AW, Oliver JL, Harrison CB, Maulder PS, Lloyd RS. Integrating models of long-term athletic development to maximize the physical development of youth. *Int J Sports Sci Coach*. 2018;13(6): 1189–99; doi: 10.1177/1747954118785503.
- [4] Moran J, Sandercock GRH, Ramírez-Campillo R, Meylan C, Collison J, Parry DA. A meta-analysis of maturation-related variation in adolescent boy athletes' adaptations to short-term resistance training. *J Sports Sci*. 2017;35(11):1041–51; doi: 10.1080/02640414.2016.1209306.
- [5] Malina RM, Cumming SP, Rogol AD, Coelho-e-Silva MJ, Figueiredo AJ, Konarski JM, Kozieł SM. Bio-banding in youth sports: background, concept, and application. *Sports Med*. 2019;49(11):1671–85; doi: 10.1007/s40279-019-01166-x.
- [6] Parr J, Winwood K, Hodson-Tole E, Deconinck FJA, Parry L, Hill JP, Malina RM, Cumming SP. Predicting the timing of the peak of the pubertal growth spurt in elite youth soccer players: evaluation of methods. *Ann Hum Biol*. 2020;47(4):400–8; doi: 10.1080/03014460.2020.1782989.
- [7] Bchini S, Dergaa I, Dhahbi W, Ceylan Hİ, Stefanica V, Selmi T, Moussaoui D, Hammami N. Integrating biological maturity into fitness assessment and physical activity interventions in

- children and adolescents: a narrative review. *Sports*. 2026;14(5):196; doi: 10.3390/sports14050196.
- [8] Malina RM, Coelho e Silva MJ, Figueiredo AJ. Growth and maturity status of youth soccer players. In: Williams AM (ed.), *Science and Soccer. Developing Elite Performers*. 3rd ed. Abingdon: Routledge; 2013; pp. 307–32; doi: 10.4324/9780203131862-27.
- [9] Towlson C, Salter J, Ade JD, Enright K, Harper LD, Page RM, Malone JJ. Maturity-associated considerations for training load, injury risk, and physical performance in youth soccer: one size does not fit all. *J Sport Health Sci*. 2021;10(4):403–12; doi: 10.1016/j.jshs.2020.09.003.
- [10] Towlson C, Watson DJ, Cumming S, Salter J, Toner J. Soccer academy practitioners' perceptions and application of bio-banding. *PLOS ONE*. 2023; 18(1):e0274079; doi: 10.1371/journal.pone.0274079.
- [11] Schmeling A, Baumann U, Schmidt S, Wernecke KD, Reisinger W. Reference data for the Thiemann-Nitz method of assessing skeletal age for the purpose of forensic age estimation. *Int J Legal Med*. 2006;120(1):1–4; doi: 10.1007/s00414-005-0002-9.
- [12] Cavallo F, Mohn A, Chiarelli F, Giannini C. Evaluation of bone age in children: a mini-review. *Front Pediatr*. 2021;9:580314; doi: 10.3389/fped.2021.580314.
- [13] Prokop-Piotrkowska M, Marszałek-Dziuba K, Moszczyńska E, Szalecki M, Jurkiewicz E. Traditional and new methods of bone age assessment – an overview. *J Clin Res Pediatr Endocrinol*. 2021; 13(3):251–262; doi: 10.4274/jcrpe.galenos.2020.2020.0091.
- [14] Tanner JM, Healy MJR, Goldstein H, Cameron N. *Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 method)*. 3rd ed. London: Saunders; 2001.
- [15] Malina RM, Coelho-e-Silva MJ, Figueiredo AJ, Philippaerts RM, Hirose N, Peña Reyes ME, Gilli G, Benso A, Vaeyens R, Deprez D, Guglielmo LF, Buranarugsa R. Tanner-Whitehouse skeletal ages in male youth soccer players: TW2 or TW3? *Sports Med*. 2018;48(4):991–1008; doi: 10.1007/s40279-017-0799-7.
- [16] Cruz-Priego GA, Guagnelli MA, Miranda-Lora AL, López-González D, Clark P. Bone age reading by DXA images should not replace bone age reading by X-ray images. *J Clin Densitom*. 2022;25(4):456–63; doi: 10.1016/j.jocd.2022.08.004.
- [17] Hind K, Slater G, Oldroyd B, Lees M, Thurlow S, Barlow M. Interpretation of dual-energy X-ray absorptiometry-derived body composition change in athletes: a review and recommendations for best practice. *J Clin Densitom*. 2018;21(3):429–43; doi: 10.1016/j.jocd.2018.01.002.
- [18] Dallman J, Herda A, Cleary CJ, Morey T, Diedrich A, Vopat BG, Vopat LM. A brief review of the literature for published dual-energy X-ray absorptiometry protocols for athletes. *Sports Health*. 2024;16(5):735–43; doi: 10.1177/19417381231208204.
- [19] Hoyer-Kuhn H, Knoop K, Semler O, Kuhr K, Hellmich M, Schoenau E, Koerber F. Comparison of DXA scans and conventional X-rays for spine morphometry and bone age determination in children. *J Clin Densitom*. 2016;19(2):208–15; doi: 10.1016/j.jocd.2015.04.006.
- [20] Romann M, Fuchslocher J. Assessment of skeletal age on the basis of DXA-derived hand scans in elite youth soccer. *Res Sports Med*. 2016;24(3):185–96; doi: 10.1080/15438627.2016.1191490.
- [21] Guss CE, McAllister A, Gordon CM. DXA in children and adolescents. *J Clin Densitom*. 2021; 24(1):28–35; doi: 10.1016/j.jocd.2020.01.006.
- [22] Heppel DHM, Taal HR, Ernst GDS, van den Akker ELT, Lequin MMH, Hokken-Koelega ACS, Geelhoed JJM, Jaddoe VWV. Bone age assessment by dual-energy X-ray absorptiometry in children: an alternative for X-ray? *Br J Radiol*. 2012;85(1010):114–20; doi: 10.1259/bjr/23858213.
- [23] Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol*. 2008;31(4):466–75; doi: 10.1002/uog.5256.
- [24] Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96–106; doi: 10.1016/j.jclinepi.2010.03.002.
- [25] ISO 5725-1:2023. Accuracy (trueness and precision) of Measurement Methods and Results. Part 1: General principles and definitions. Geneva: International Organization for Standardization; 2023.
- [26] de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol*. 2006;59(10):1033–9; doi: 10.1016/j.jclinepi.2005.10.015.
- [27] Keylock L, Cameron N. Reproducibility of bone age assessment from DXA hand scans: expert versus novice. *Ann Hum Biol*. 2021;48(4):343–5; doi: 10.1080/03014460.2021.1956586.

- [28] Dhahbi W, Chamari K. The algorithmic athlete: a call to standardize assessment of sensor technologies and artificial intelligence. *Int J Sports Physiol Perform.* 2026;21(4):505–6; doi: 10.1123/ijsp.2025-0547.
- [29] Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17(1):101–10; doi: 10.1002/(SICI)1097-0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E.
- [30] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155–63; doi: 10.1016/j.jcm.2016.02.012.
- [31] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979; 86(2):420–428; doi: 10.1037/0033-2909.86.2.420.
- [32] Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation – a discussion and demonstration of basic features. *PLOS ONE.* 2019;14(7): e0219854; doi: 10.1371/journal.pone.0219854.
- [33] Mansournia MA, Waters R, Nazemipour M, Bland M, Altman DG. Bland–Altman methods for comparing methods of measurement and response to criticisms. *Glob Epidemiol.* 2021;3:100045; doi: 10.1016/j.gloepi.2020.100045.
- [34] Giavarina D. Understanding Bland–Altman analysis. *Biochem Med.* 2015;25(2):141–51; doi: 10.11613/BM.2015.015.
- [35] Efron B, Tibshirani RJ. An introduction to the Bootstrap. New York: Chapman and Hall; 1994; doi: 10.1201/9780429246593.
- [36] Ostojic SM, Castagna C, Calleja-González J, Jukic I, Idrizovic K, Stojanovic M. The biological age of 14-year-old boys and success in adult soccer: do early maturers predominate in the top-level game?. *Res Sports Med.* 2014;22(4):398–407; doi: 10.1080/15438627.2014.944303.
- [37] Monasterio X, Bidaurreazaga-Letona I, Larruskain J, Lekue JA, Diaz-Beitia G, Santisteban JM, Martin-Garetxana I, Gil SM. Relative skeletal maturity status affects injury burden in U14 elite academy football players. *Scand J Med Sci Sports.* 2022;32(9):1400–9; doi: 10.1111/sms.14204.
- [38] Silva SP, Freitas DL, Beunen GP, Maia JAR. The relevance of training in the TW3 method for the evaluation of biological maturation. *Rev Bras Cineantropom Desempenho Hum.* 2010;12(5):352–8; doi: 10.5007/1980-0037.2010v12n5p352.
- [39] Martín Pérez SE, Martín Pérez IM, Molina Suárez R, Vega González JM, García Hernández AM. The validation of the Tanner–Whitehouse 3 method for radiological bone assessments in a pediatric

population from the Canary Islands. *Osteology.* 2025;5(1):6; doi: 10.3390/osteology5010006.

- [40] Maratova K, Zemkova D, Sedlak P, Pavlikova M, Amaratunga SA, Krasnicanova H, Soucek O, Sumnik Z. A comprehensive validation study of the latest version of BoneXpert on a large cohort of Caucasian children and adolescents. *Front Endocrinol.* 2023;14:1130580; doi: 10.3389/fendo.2023.1130580.

Supplementary materials

Observer repeatability and inter-rater reproducibility of DXA-derived TW3-RUS skeletal age assessment in elite youth football players: a complementary framework for evaluating measurement quality.

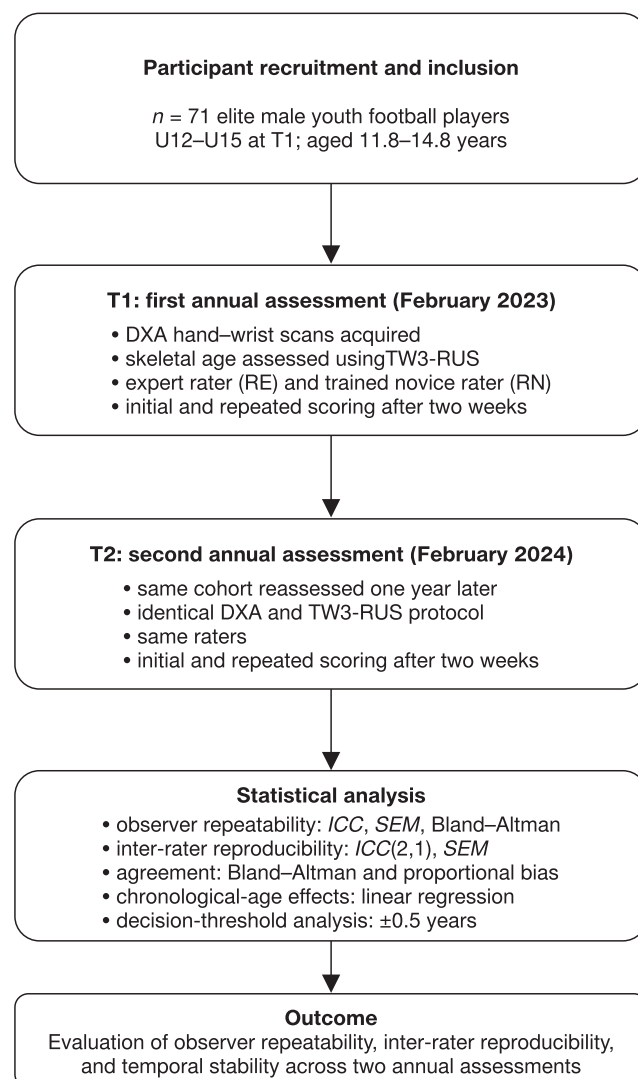


Figure S1. Research design of the two annual DXA-derived TW3-RUS skeletal age assessments



Figure S2. DXA hand scan with standardised brightness (2210) and contrast (685) (TW3)

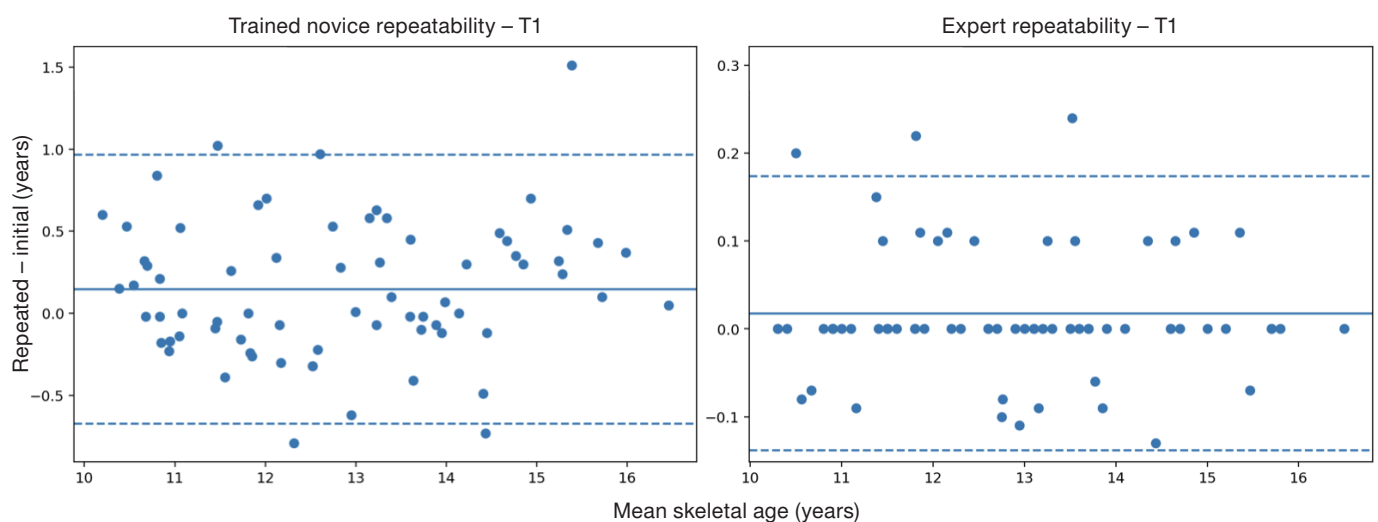


Figure S3. Bland-Altman plots for observer repeatability during the first annual assessment (T1)

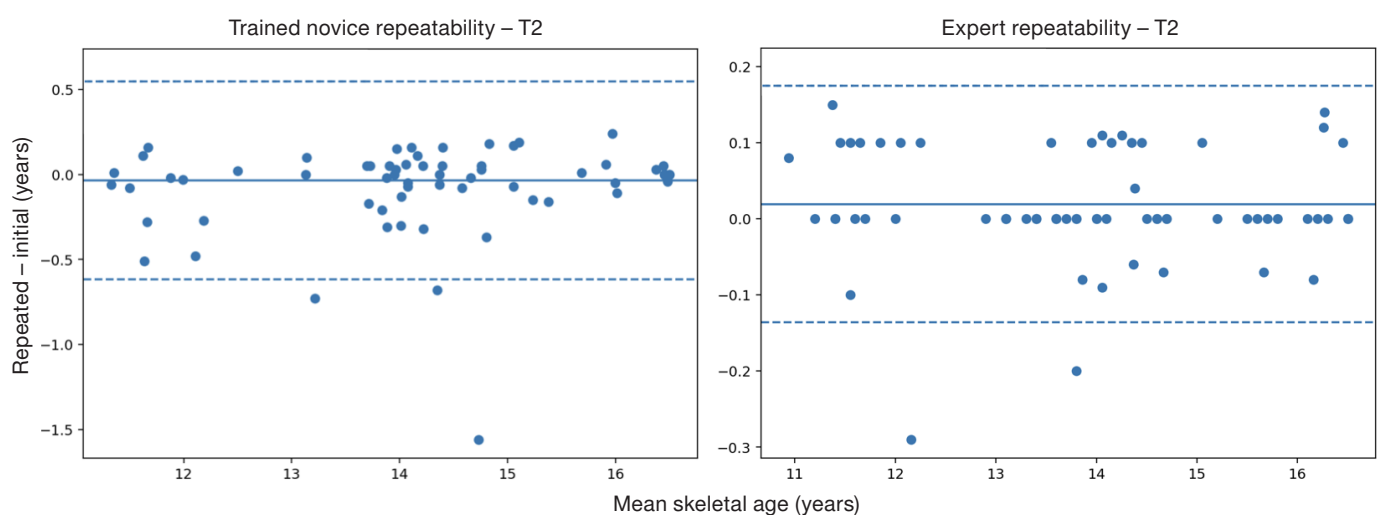


Figure S4. Bland-Altman plots for observer repeatability during the second annual assessment (T2)

Table S1. Bootstrap comparison of observer repeatability between the expert rater and the trained novice rater

Group	T1 ΔICC (95% <i>CI</i>)	T1 ΔSEM (months)	T2 ΔICC (95% <i>CI</i>)	T2 ΔSEM (months)
Overall	+0.034 (0.021–0.052)	–3.038	+0.017 (0.006–0.037)	–1.823
U12/U13	+0.027 (0.007–0.115)	–2.484	+0.016 (0.006–0.047)	–1.709
U13/U14	+0.092 (0.042–0.188)	–3.260	+0.086 (0.010–0.315)	–2.795
U14/U15	+0.039 (0.015–0.132)	–2.720	+0.003 (0.001–0.016)	–0.408
U15/U16	+0.039 (0.006–0.142)	–3.613	+0.007 (0.000–0.041)	–1.047

U12–U16 – under-12 to under-16 age categories, T1 – first annual assessment, T2 – second annual assessment

Paired group labels show the age category at T1 followed by that at T2.

Δ values were calculated as expert rater minus trained novice rater.

Negative ΔSEM values indicate lower measurement error for the expert rater.