Published in final edited form as:
BJOG. 2013 September ; 120(0 2): 48–v. doi:10.1111/1471-0528.12127.

Anthropometric standardisation and quality control protocols for the construction of new, international, fetal and newborn growth standards: the INTERGROWTH-21st Project

L Cheikh Ismail[a], **HE Knight**[a], **EO Ohuma**[a], **L Hoch**[a], and **WC Chumlea**[b] **for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st)**

[a]Nuffield Department of Obstetrics & Gynaecology, and Oxford Maternal & Perinatal Health Institute, Green Templeton College, University of Oxford, Oxford, UK

[b]Lifespan Health Research Center, Departments of Community Health and Pediatrics, Boonshoft School of Medicine, Wright State University, Dayton, OH, USA

Abstract

The primary aim of the INTERGROWTH-21st Project is to construct new, prescriptive standards describing optimal fetal and preterm postnatal growth. The anthropometric measurements include the head circumference, recumbent length and weight of the infants, and the stature and weight of the parents. In such a large, international, multicentre project, it is critical that all study sites follow standardised protocols to ensure maximal validity of the growth and nutrition indicators used. This paper describes in detail the anthropometric training, standardisation and quality control procedures used to collect data for these new standards. The initial standardisation session was in Nairobi, Kenya, using newborns, which was followed by similar sessions in the eight participating study sites in Brazil, China, India, Italy, Kenya, Oman, UK and USA. The intraobserver and inter-observer technical error of measurement values for head circumference range from 0.3 to 0.4 cm, and for recumbent length from 0.3 to 0.5 cm. These standardisation protocols implemented at each study site worldwide ensure that the anthropometric data collected are of the highest quality to construct international growth standards.

Keywords

Anthropometry; fetal growth; INTERGROWTH-21st; newborn measurements; quality control protocols; standardisation

Correspondence: Dr L Cheikh Ismail, Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Women's Centre, Level 3, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK. leila.cheikhismail@obs-gyn.ox.ac.uk.

Disclosure of interests
None.

Contribution to authorship
L Cheikh Ismail and W Chumlea wrote the manuscript and all the authors read and approved the final version.

Details of ethics approval
The INTERGROWTH-21st Project was approved by the Oxfordshire Research Ethics Committee 'C' (reference:08/H0606/139) and the research ethics committees of the individual participating institutions and corresponding health authorities where the Project was implemented.

Introduction

The International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) is a large-scale, population-based, multicentre project involving health institutions from eight geographically diverse countries, which aims to assess fetal, newborn and preterm growth under optimal conditions, in a manner similar to that adopted by the World Health Organization (WHO) Multicentre Growth Reference Study (MGRS).[1] The INTERGROWTH-21st Project has three major components, which were designed to create: (1) longitudinally derived, prescriptive, international, fetal growth standards using both clinical and ultrasound measures; (2) preterm, postnatal growth standards for those infants born at $\geq 26^{+0}$ but $<37^{+0}$ weeks of gestation in the longitudinal cohort; and (3) birthweight-for-gestational-age standards derived from all newborns delivering at the study sites over an approximately 12 month period.[2]

To achieve uniformity in such a large, multicentre project, it was important for all study sites to follow the same standardised measurement procedures to ensure maximal validity of the resulting standards as indicators of growth and nutrition. Hence, the same standardised anthropometric protocols are being used in all three components of the INTERGROWTH-21st Project—the Fetal Growth Longitudinal Study (FGLS), the Preterm Postnatal Follow-up Study (PPFS) and the Newborn Cross-Sectional Study (NCSS).[3] This paper describes in detail the training, standardisation and quality control procedures that are being used to collect the data to construct the new standards.

Training and standardisation of anthropometry personnel

An important goal of training and standardisation is to ensure that measurers develop, refine and maintain their techniques so that they take accurate and precise measurements that yield repeatable and reproducible values. Standardisation within and between study sites is a crucial component that ensures the data are comparable across measurers and sites over the course of the project. Before data collection, a 5 day training and standardisation session was held in Nairobi, Kenya, for the lead anthropometrists from each study site. Several international anthropometry experts from the MGRS team facilitated the training. The INTERGROWTH-21st Project Coordinating Unit prepared all training materials based on the published MGRS anthropometry protocols.[4]

During training, these lead anthropometrists familiarised themselves extensively with the measuring equipment and techniques, which are described in detail elsewhere in this supplement.[3] They watched the MGRS anthropometry training video (www.who.int/childgrowth/training/en/), which highlights key measurement techniques and calibration procedures. A general overview of the study protocols and measurement techniques was presented followed by a formal standardisation training session to assess measurer accuracy and precision. Measurers are considered *accurate* if, on average, they record measurement values (interobserver as compared with the expert) that are consistently close to the true or an expert's value. A *precise* measurer is one who, when re-measuring the same child, records values (intraobserver) that are close to each other and not widely dispersed. This is independent of whether the average is close to the true value or not and so independent of

accuracy. To assess accuracy and precision a test–retest study is conducted in which an expert, (considered to be the 'gold standard'), and an anthropometrist both measure a group of newborns twice. Bias is calculated as the average difference between the values obtained by the expert and anthropometrist. Inter-observer precision is calculated by comparing each observer's replicate measurements of the same participants. The most common parameter describing a lack of precision is the technical error of measurement (TEM), which is the square root of the sum of the squared differences between duplicate measurements, divided by twice the number of participants measured.[5] In the INTERGROWTH-21st Project, an acceptable TEM for a measurer should be no more than twice that of the expert.

During the initial standardisation session in Nairobi, Kenya, a group of 20 newborns (34–40 weeks of gestation at birth and 1–5 days old) were measured. The anthropometrists were standardised against W.C.C., who served as the 'expert' for both MGRS and the INTERGROWTH-21st Project. The session consisted of two phases: first, the expert and each anthropometrist measured the head circumference (HC) and length of each baby and recorded their results independently. The same 20 newborns were then remeasured by the expert and each measurer and the inter-observer and intraobserver variabilities were calculated.

The results were analysed using the same Microsoft Excel spreadsheet used in MGRS, preprogrammed with standard formulae for calculating the relevant statistics, including the TEM. The sign test for precision assesses 'the measurement effect' where an observer's retest measurement can be systematically higher or lower than his/her own first measurement. [6] For accuracy, both the sign test and the *F* test are useful. The *F* test indicates whether accuracy is significantly different between an observer and all observers together, whereas the sign test checks whether poor accuracy results from systematic or sporadic bias.[7] For instance, the average bias can be low and non-significant when a single large deviation overwhelms small but systematic differences; in this case, the sign test indicates bias not the *F* test.

To illustrate the observers' performances, Tables 1 and 2 present the HC and length data respectively for the initial standardisation session in Nairobi, Kenya. The expert anthropometrist (W.C.C.) had the lowest mean TEM in each group, indicating a high level of precision. For HC measurements, with the exception of Observer 4 in Group 2, each observer displayed a good level of precision with TEM values within the acceptable limit of twice the expert's mean TEM value. There was no evidence of systematic bias for any of the anthropometrists' second measurements, compared with their first (sign test $P > 0.05$). There was no systematic tendency either to overestimate (positive bias) or underestimate (negative bias) any of the anthropometrists' measurements compared with the expert or the group mean for each measurement (Table 1).

For length measurements, there was a high level of precision between anthropometrists in both groups as none of their TEMs exceeded twice that of the lead anthropometrist's value. Similarly, there was no systematic tendency either to overestimate or underestimate any of the anthropometrists' measurements compared with the expert (with the exception of

Observers 1 and 5 in Group 1) or the group mean (with the exception of Observer 1 in Group 1) (Table 2).

Site training and standardisation sessions

Before commencing data collection at the study sites, each local lead anthropometrist conducted a similar training and standardisation session at their site with their anthropometry team. These sessions again involved measuring 20 newborns; they were carried out under the supervision of a visiting expert member of the INTERGROWTH-21st Anthropometry Group. The 'gold standard' was the mean measurement of all observers in these sessions. The results were analysed by the lead anthropometrists using the MGRS spreadsheet, which allowed them to identify any group member deviating from the recommended techniques. These standardisation sessions are repeated every 3 months at each study site to: (1) ensure anthropometrists are following the recommended techniques; (2) monitor their reliability (precision and accuracy); and (3) take corrective measures if required. These sessions over 1–2 days involve the measurement and re-measurement of ten newborns and the results are sent to the INTERGROWTH-21st Anthropometry Group soon after completion.

The TEMs for all anthropometrists' first and second standardisation sessions at each study site were plotted to monitor overall performance for HC and length, as illustrated in Figures 1A and 2A, respectively. Overall, the figures depict an improvement in precision in the second, compared with the first, standardisation session. This trend is consistent across all study sites for HC and length measurements. Plots of bias (each anthropometrist compared with the overall mean of the study site group) similarly showed increased accuracy in the second, compared with the first, standardisation session. There was no systematic bias for either HC or length measurements across study sites (Figures 1B and 2B).

At the half-way point for data collection at each study site's involvement in the INTERGROWTH-21st Project, the expert anthropometrist visits to supervise a local standardisation session. This exercise ensures that all sites are adhering to the protocols and rigorously following the standardisation procedures. Further training is conducted at each site, if deemed necessary.

Quality control activities

The on-going quality control measures require anthropometrists at each study site to take and record all measurements independently, and compare their values with the maximum allowable differences, as documented elsewhere in this supplement.[3] The anthropometrists check the forms visually after each measurement session to ensure that appropriate remeasurements are performed when necessary. They also check the completed forms for missing or inaccurate values before sending them for data entry.

Data entry is performed at each site using the INTERGROWTH-21st centrally coordinated, online, data management system with built-in range and consistency checks. Local data managers ensure that data entry does not lag behind data collection by more than 3 days. Recorded values are flagged if they are out of the expected range, giving rise to checks for

recording possible data entry errors and raising queries if necessary. The data manager at the INTERGROWTH-21st Project Coordinating Unit checks all flagged values for the following: consistency between anthropometrists (i.e. within maximum allowed difference); consistency with other anthropometric indicators (e.g. to check if this is just a 'big' baby); consistency with previous measurements of the same infant (for the follow-up study), and that there are no data entry errors. Additional quality control activities performed on the INTERGROWTH-21st data are documented elsewhere in this supplement.[8]

Discussion

These rigorous standardisation and quality control measures are employed to ensure that data of the highest quality are acquired to construct the prescriptive standards for optimal fetal and preterm postnatal growth, as well as newborn nutritional status. The statistical methods used herein to assess the accuracy and precision of the anthropometrists are robust for use in large multicentre studies.[4] The high frequency of compulsory standardisation sessions for all anthropometrists at each of the eight study sites keeps their skills sharp. These sessions are also an opportunity to identify any departures from prescribed measurement techniques in the protocols and to take corrective measures if required.

Reliability data are crucial for the interpretation of anthropometric assessments of growth and nutritional status in children.[9,10] However, little has been published on the subject among newborns and infants because of their fragility. Our accuracy and precision data for newborns are part of a unique set of important health information. The TEM values for HC in Table 1 range from approximately 0.3 to 0.4 cm. Corresponding values in the literature are approximately 0.2 to 0.4 cm in clinical samples and from 0.1 to 0.2 cm for all children in the Fels Longitudinal Study.[11–13]

Standardisation data from MGRS[4] for recumbent length are very similar to those reported in Table 2 for the same measurement. The TEM values for the expert anthropometrist (the same individual in both studies) and for the four observers in MGRS range from approximately 0.3 to 0.5 cm, compared with 0.3 to almost 0.6 cm in Table 2. The MGRS data for bias relating to recumbent length are also very similar to the corresponding data in Table 2. These findings are much lower than reliability data for length in clinical samples.[14]

The findings in Tables 1 and 2 are further verified by the plots of the TEM and bias values by site in Figures 1 and 2, respectively. The TEM values (Figures 1A and 2A) for the second session, HC and length are smaller and more linear at each study site. This improvement in measuring technique is also clear in the bias plots for the second session (Figure 2A and 2B): the values at each site are more linear and closer to zero than those from the first session. In MGRS, similar levels of reliability for newborn recumbent length measurements were documented over 18 months of data collection across the participating sites.[4] Comparative data for HC reliability in newborns are scarce.

Newborns and infants are the most difficult group to measure because of their very small size and fragility, and the relatively large hands of those who are measuring them. Hence, accuracy and precision data are necessary to ensure that measurements are of sufficiently

high quality to allow accurate interpretation of growth and nutritional status. This applies particularly to preterm infants. To maintain data quality, the INTERGROWTH-21st Project Coordinating Unit monitors the percentage of repeat measurements at each site. A high level could indicate poor measurement technique on the part of one or more measurers; a low level might indicate a lack of independence between two or more measurers. The maximum allowable differences are 0.5 cm for HC and 0.7 cm for length. Data from MGRS indicate that a repeat rate of about 5% is anticipated.[4] If the rate deviates significantly from this percentage, the data manager informs the INTERGROWTH-21st Anthropometry Group Leader who then investigates the reason for this deviation with the local lead anthropometrist. Similar limits on the allowable differences between independent repeat measures have been used in MGRS, the National Health and Nutrition Examination Survey[15] and other large, multicentre/national studies.

In conclusion, the implementation of these standardisation sessions and quality control measures across study sites in the INTERGROWTH-21st Project provides a useful means of controlling inter-observer and intraobserver variability during data collection. Newborns are difficult to measure and anthropometric data collected from them is therefore prone to error. However, stringent quality control protocols can effectively manage the normal errors that occur during data collection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This Project was supported by the INTERGROWTH-21st Grant ID# 49038 from the Bill & Melinda Gates Foundation to the University of Oxford, for which we are very grateful.

A full list of Members of the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) and its Committees appears in the preliminary pages of this supplement.

References

1. de Onis M, Garza C, Victora CG, Bhan MK, Norum K. The WHO Multicentre Growth Reference Study (MGRS): rationale, planning, and implementation. Food Nutr Bull. 2004; 25:S1–90.
2. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, et al. for the International Fetal and Newborn Growth Consortium (INTERGROWTH-21st). The objectives, design and implementation of the INTERGROWTH-21st Project. BJOG. 201310.1111/1471-0528.12047
3. Cheikh Ismail L, Knight HE, Bhutta Z, Chumlea WC. for the International Fetal Newborn Growth Consortium (INTERGROWTH-21st). Anthropometric protocols for the construction of new international fetal and newborn growth standards: the INTERGROWTH-21st Project. BJOG. 201310.1111/1471-0528.12125
4. de Onis M, Onyango AW, Van den Broeck J, Chumlea WC, Martorell R. Measurement and standardisation protocols for anthropometry used in the construction of a new international growth reference. Food Nutr Bull. 2004; 25(1 Suppl):S27–36. [PubMed: 15069917]
5. Malina, RM.; Hamill, PVV.; Lemeshow, S. Vital Health and Statistics Series 11. USDHHS; Washington, D.C: US Government Printing Office; 1973. Selected measurements of children 6–11 years.

6. World Health Organization. Measuring Change in Nutritional Status, Annex 1: In Standardization Procedures for the Collection Of Weight and Height Data in the Field. Geneva: World Health Organization; 1983.
7. Daly, LE.; Bourke, GJ. Interpretation and Uses of Medical Statistics. 5. Oxford, UK: Blackwell Science; 2000.
8. Ohuma EO, Hoch L, Cosgrove C, Knight HE, Cheikh Ismail L, Juodvirsiene L, et al. for the International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Managing data for the international, multicentre INTERGROWTH-21st Project. BJOG. 201310.1111/1471-0528.12080
9. Garn, SM.; Shamir, Z. Methods for Research in Human Growth. Springfield, IL: Charles C Thomas; 1958. 1958
10. Cameron, N. The measurement of human growth. London: Croom Helm; 1984.
11. Lohman, TG.; Roche, AF.; Martorell, R. Anthropometric Standardization Reference Manual. Champaign, IL: Human Kinetics Books; 1958. 1988
12. Roche, AF.; Sun, SS. Human growth: assessment and interpretation. Cambridge: Cambridge University Press; 2003.
13. West J, Manchester B, Wright J, Lawlor DA, Waiblinger D. Reliability of routine clinical measurements of neonatal circumferences and research measurements of neonatal skinfold thicknesses: findings from the Born in Bradford study. Paediatr Perinat Epidemiol. 2011; 25:164–71. [PubMed: 21281329]
14. Johnson TS, Engstrom JL, Warda JA, Kabat M, Peters B. Reliability of length measurements in full-term neonates. J Obstet Gynecol Neonatal Nurs. 1998; 27:270–6.
15. U.S. Department of Health and Human Services. National Center for Health Statistics. NHANES III Anthropometric Procedures Video. Washington, DC: U.S. Government Printing Office; 1996. Stock No. 017-022-01355-5

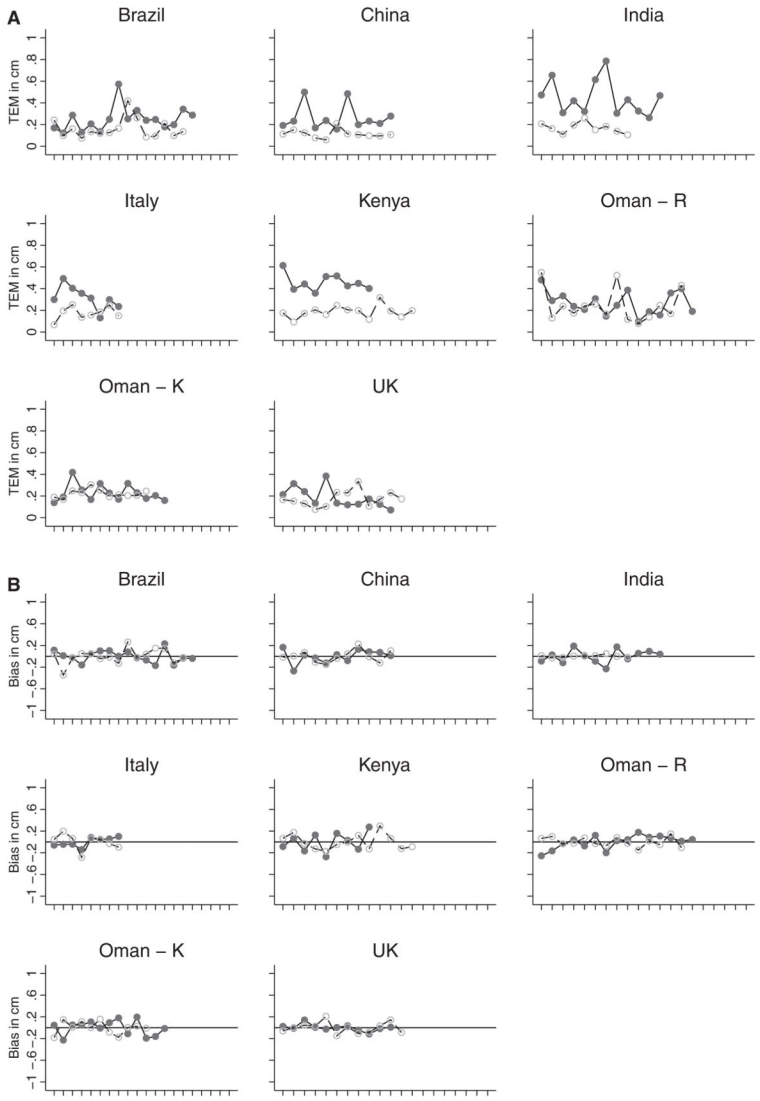

Figure 1.

(A) Technical error of measurement (TEM) for head circumference by country at first standardisation session (solid line with solid circles) and second standardisation session (dashed line with open circles). (B) Bias of observers with their overall mean in each country for head circumference at first standardisation session (solid line with solid circles) and second standardisation session (dashed line with open circles). Standardisation sessions were conducted at two hospitals in Oman: Oman – R, Royal Hospital; Oman – K, Khoula Hospital.

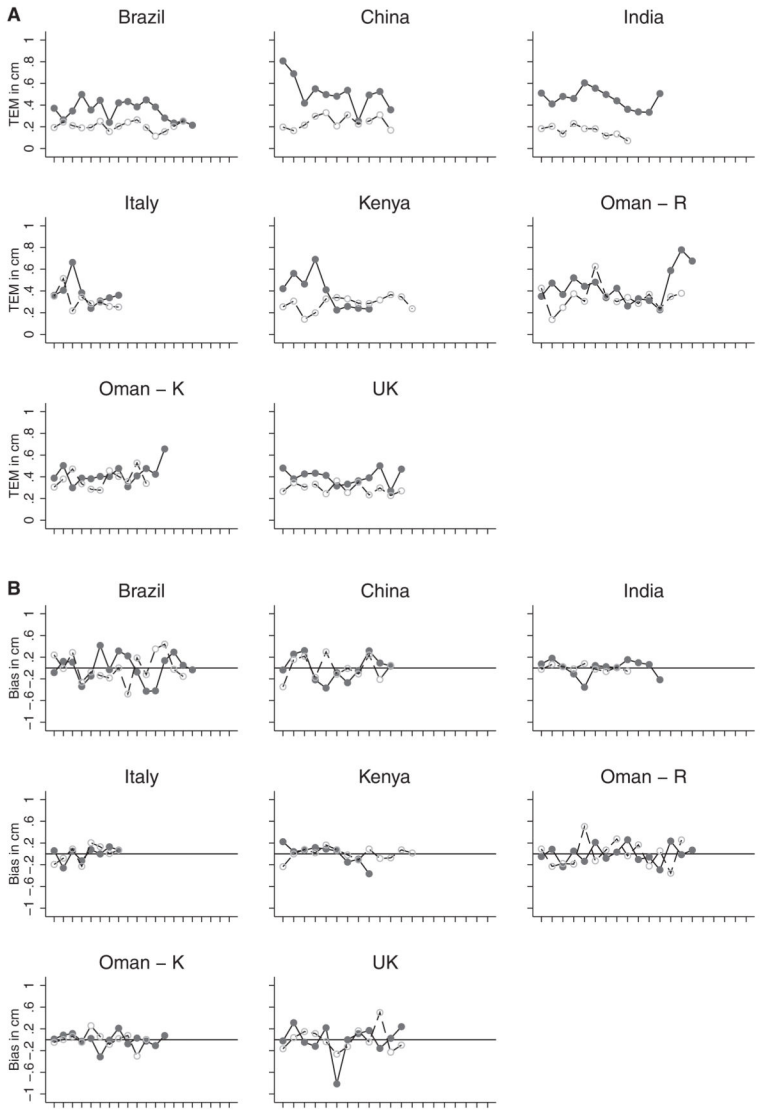

Figure 2.

(A) Technical error of measurement (TEM) for length by country at first standardisation session (solid line with solid circles) and second standardisation session (dashed line with open circles). (B) Bias of observers with their overall mean in each country for length at first standardisation session (solid line with solid circles) and second standardisation session (dashed line with open circles). Standardisation sessions were conducted at two hospitals in Oman: Oman – R, Royal Hospital; Oman – K, Khoula Hospital.

Table 1

Precision and accuracy from the initial standardisation session in Nairobi, Kenya: head circumference data

	TEM*	*F* test		Sign test***	Bias			Bias		
		Lead anthropometrist**	Overall mean**		Lead anthropometrist****	*F* test*****	Sign test******	Overall mean****	*F* test*****	Sign test******
Group 1										
WHO lead anthropometrist	0.227	–	$P > 0.25$	$P > 0.05$	–	–	–	−0.059	$P > 0.25$	$P > 0.05$
Observer 1	0.304	$0.05 < P < 0.10$	$P > 0.25$	$P > 0.05$	0.178	$0.05 < P < 0.10$	$P < 0.05$	0.119	$0.10 < P < 0.25$	$P < 0.05$
Observer 2	0.272	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	0.095	$P > 0.25$	$P > 0.05$	0.036	$P > 0.25$	$P > 0.05$
Observer 3	0.276	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	0.113	$0.10 < P < 0.25$	$P < 0.05$	0.054	$P > 0.25$	$P > 0.05$
Observer 4	0.248	$P > 0.25$	$P > 0.25$	$P > 0.05$	−0.050	$P > 0.25$	$P > 0.05$	−0.109	$P > 0.25$	$P < 0.05$
Observer 5	0.315	$0.05 < P < 0.10$	$P > 0.25$	$P > 0.05$	0.018	$P > 0.25$	$P > 0.05$	−0.041	$P > 0.25$	$P > 0.05$
Group 2										
WHO lead anthropometrist	0.133	–	$P > 0.25$	$P > 0.05$	–	–	–	−0.052	$P > 0.25$	$P > 0.05$
Observer 1	0.171	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	0.100	$P < 0.01$	$P > 0.05$	0.048	$P > 0.25$	$P > 0.05$
Observer 2	0.174	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	0.055	$0.05 < P < 0.10$	$P > 0.05$	0.003	$P > 0.25$	$P > 0.05$
Observer 3	0.207	$0.01 < P < 0.05$	$P > 0.25$	$P < 0.05$	0.213	$P < 0.01$	$P < 0.05$	0.160	$P > 0.25$	$P < 0.05$
Observer 4	0.475	$P < 0.01$	$P < 0.01$	$P > 0.05$	−0.071	$P < 0.01$	$P < 0.05$	−0.125	$0.01 < P < 0.05$	$P < 0.05$
Observer 5	0.167	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	0.012	$P > 0.25$	$P > 0.05$	−0.040	$P > 0.25$	$P > 0.05$

* Technical error of measurement (TEM): $\sqrt{\sum(d_i^2/2n)}$; where d_i is the difference between the i^{th} participant's test and retest measurements by the observer and n is the number of measured participants.

** *F* ratio for precision: $\text{Observer's}\sum(d_i^2)/\text{Lead anthropometrists's}\sum(d_i^2)$. When overall mean is the gold standard, d_i in the denominator is the difference between the i^{th} participant's overall mean of retest measurements.

*** Precision sign test: binomial proportion p, where $p = x/n$, and x is the frequency of the observer's retest scores that are higher (or lower) than the corresponding test scores. Significance is based on exact confidence limits for proportions when $n \leq 75$ (see table B.11 in Daly and Bourke[7]).

**** Average bias: Observer $\Sigma\Delta_i/n$; where Δ_i is the difference between the observer's mean and the lead anthropometrist's (or overall) mean measurement for the i^{th} participant.

***** *F* ratio for bias: $\text{Observer's}\sum(\Delta_i^2)/\text{Lead anthropometrist's}\sum(\Delta_i^2)$ or overall means $\sum(d_i^2)$ (same denominator as the precision *F* ratio).

****** Bias sign test: binomial proportion p, where $p = x/n$, and x is the frequency of the observer's means that are above (or below) the lead anthropometrist's mean or overall mean. Significance is based on exact confidence limits for proportions when $n \leq 75$ (see table B.11 in Daly and Bourke[7]).

Table 2

Precision and accuracy from the initial standardisation session in Nairobi, Kenya: length data

	TEM*	*F* test		Sign test***	Bias			Bias		
		Lead anthropometrist**	Overall mean**		Lead anthropometrist****	*F* test*****	Sign test******	Overall mean****	*F* test*****	Sign test******
Group 1										
WHO lead anthropometrist	0.371	–	$P > 0.25$	$P > 0.05$	–	–	–	−0.069	$P > 0.25$	$P > 0.05$
Observer 1	0.596	$0.01 < P < 0.05$	$0.01 < P < 0.05$	$P > 0.05$	−0.348	$0.01 < P < 0.05$	$P < 0.05$	−0.416	$0.01 < P < 0.05$	$P < 0.05$
Observer 2	0.314	$P > 0.25$	$P > 0.25$	$P < 0.05$	−0.055	$P > 0.25$	$P > 0.05$	−0.124	$P > 0.25$	$P < 0.05$
Observer 3	0.368	$P > 0.25$	$P > 0.25$	$P < 0.05$	0.288	$P > 0.25$	$P < 0.05$	0.219	$P > 0.25$	$P < 0.05$
Observer 4	0.312	$P > 0.25$	$P > 0.25$	$P < 0.05$	0.166	$P > 0.25$	$P < 0.05$	0.067	$P > 0.25$	$P < 0.05$
Observer 5	0.349	$P > 0.25$	$P > 0.25$	$P > 0.05$	0.432	$0.05 < P < 0.10$	$P < 0.05$	0.355	$0.10 < P < 0.25$	$P < 0.05$
Group 2										
WHO lead anthropometrist	0.311	–	$P > 0.25$	$P > 0.05$	–	–	–	0.160	$P > 0.25$	$P < 0.05$
Observer 1	0.377	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	−0.288	$0.10 < P < 0.25$	$P < 0.05$	−0.127	$P > 0.25$	$P > 0.05$
Observer 2	0.534	$0.01 < P < 0.05$	$P > 0.25$	$P < 0.05$	−0.186	$P > 0.25$	$P < 0.05$	−0.015	$P > 0.25$	$P > 0.05$
Observer 3	0.513	$0.01 < P < 0.05$	$P > 0.25$	$P > 0.05$	−0.227	$P > 0.25$	$P < 0.05$	−0.067	$P > 0.25$	$P > 0.05$
Observer 4	0.543	$P < 0.01$	$0.10 < P < 0.25$	$P > 0.05$	−0.080	$0.05 < P < 0.10$	$P > 0.05$	0.080	$0.05 < P < 0.10$	$P > 0.05$
Observer 5	0.409	$0.10 < P < 0.25$	$P > 0.25$	$P > 0.05$	−0.193	$P > 0.25$	$P < 0.05$	−0.032	$P > 0.25$	$P > 0.05$

* Technical error of measurement (TEM): $\sqrt{\sum(d_i^2/2n)}$; where d_i is the difference between the i^{th} participants test and retest measurements by the observer and n is the number of measured participants.

** *F* ratio for precision: $\text{Observer's}\sum(d_i^2)/\text{Lead anthropometrists's}\sum(d_i^2)$. When overall mean is the gold standard, d_i in the denominator is the difference between the i^{th} participant's overall mean of retest measurements.

*** Precision sign test: binomial proportion *p*, where $p = x/n$, and *x* is the frequency of the observer's retest scores that are higher (or lower) than the corresponding test scores. Significance is based on exact confidence limits for proportions when $n \leq 75$ (see table B.11 in Daly and Bourke[7]).

**** Average bias: Observer $\Sigma\Delta_i/n$; where Δ_i is the difference between the observer's mean and the lead anthropometrist's (or overall) mean measurement for the i^{th} participant.

***** *F* ratio for bias: $\text{Observer's}\sum(\Delta_i^2)/\text{Lead anthropometrist's}\sum(\Delta_i^2)$ or overall means $\sum(d_i^2)/(\text{same denominator as the precision } F \text{ ratio})$.

****** Bias sign test: binomial proportion *p*, where $p = x/n$, and *x* is the frequency of the observer's means that are above (or below) the lead anthropometrist's mean or overall mean. Significance is based on exact confidence limits for proportions when $n \leq 75$ (see table B.11 in Daly and Bourke[7]).