

Reliability of anthropometric measurements in the WHO Multicentre Growth Reference Study

WHO MULTICENTRE GROWTH REFERENCE STUDY GROUP^{1,2}

¹Department of Nutrition, World Health Organization, Geneva, Switzerland, and ²Members of the WHO Multicentre Growth Reference Study Group (listed at the end of the first paper in this supplement)

Abstract

Aim: To describe how reliability assessment data in the WHO Multicentre Growth Reference Study (MGRS) were collected and analysed, and to present the results thereof. **Methods:** There were two sources of anthropometric data (length, head and arm circumferences, triceps and subscapular skinfolds, and height) for these analyses. Data for constructing the WHO Child Growth Standards, collected in duplicate by observer pairs, were used to calculate inter-observer technical error of measurement (TEM) and the coefficient of reliability. The second source was the anthropometry standardization sessions conducted throughout the data collection period with the aim of identifying and correcting measurement problems. An anthropometry expert visited each site annually to participate in standardization sessions and provide remedial training as required. Inter- and intra-observer TEM, and average bias relative to the expert, were calculated for the standardization data. **Results:** TEM estimates for teams compared well with the anthropometry expert. Overall, average bias was within acceptable limits of deviation from the expert, with head circumference having both lowest bias and lowest TEM. Teams tended to underestimate length, height and arm circumference, and to overestimate skinfold measurements. This was likely due to difficulties associated with keeping children fully stretched out and still for length/height measurements and in manipulating soft tissues for the other measurements. Intra- and inter-observer TEMs were comparable, and newborns, infants and older children were measured with equal reliability. The coefficient of reliability was above 95% for all measurements except skinfolds whose R coefficient was 75–93%.

Conclusion: Reliability of the MGRS teams compared well with the study's anthropometry expert and published reliability statistics.

Key Words: Anthropometry, bias, measurement error, measurement reliability, precision

Introduction

Measurement reliability is a direct indicator of data quality. Reducing errors in measurement will increase the probability that any relationships among variables in a study are uncovered. Adherence to recommended procedures will reduce bias in measurement and increase the certainty of inferences about similarities/differences with respect to other populations. For these and other reasons, it is generally cost effective to reduce measurement error to recommended minima. Standardized data collection methodology, rigorous training and monitoring of data collection personnel, frequent and effective equipment calibration and maintenance, and periodic assessment of anthropometric measurement reliability were among the quality assurance measures included in the World Health Organization's (WHO) Multicentre Growth Reference Study (MGRS) of infants and children [1].

Anthropometry standardization sessions were conducted with the goal of monitoring anthropometric measurement techniques, identifying sources of error or bias and retraining teams or individuals as necessary.

Only a few growth studies and surveys [2–11] provide detailed descriptions of anthropometric standardization and measurement reliability assessments. The standardization of measurement techniques in anthropometry by Lohman and colleagues in the late 1980s has been a useful guide and reference for the collection of reliable anthropometric measurements [12]. However, there is a lack of uniformity in the methods employed in collecting reliability data and in reporting the statistics and terminology used in reliability assessment [6,7,9,11,13–16].

The objectives of this article are to describe the approach used in the MGRS to collect and analyse

Correspondence: Mercedes de Onis, Study Coordinator, Department of Nutrition, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Tel: +41 22 791 3320. Fax: +41 22 791 4156. E-mail: deonism@who.int

reliability information, to present key results about measurement reliability, and to assess the implications of these results for the MGRS. The analyses are based on data collected during anthropometric standardization sessions held regularly over the duration of data collection in each of the six MGRS sites and on duplicate measurements taken during routine data collection.

Methods

Data collection teams and procedures

Data in the MGRS were collected between 1997 and 2003 in Brazil, Ghana, India, Norway, Oman and the USA [17]. Data collection teams were trained in each site during the study's preparatory phase, at which time measurement techniques were standardized against one of two MGRS lead anthropometrists. During the study, one of these experts visited each site annually to participate in standardization sessions [1]. For the longitudinal component of the study, screening teams measured newborns within 24 h of delivery, and follow-up teams conducted home visits until the children reached 24 mo of age. The follow-up teams also carried out measurements in the cross-sectional component of the MGRS involving children aged 18–71 mo.

The anthropometric variables measured were weight and head circumference at all ages, recumbent length in the longitudinal study, height in the cross-sectional study, and arm circumference, triceps and subscapular skinfolds in all children aged ≥ 3 mo. The methodology and equipment used in taking these measurements have been described in detail elsewhere [1]. Briefly, anthropometric data were collected by observers working in pairs. Each observer independently measured and recorded a complete set of measurements, and the two then compared their readings. If any pair of readings exceeded the maximum allowable difference for a given variable (weight 100 g; circumferences 5 mm; length/height 7 mm; skinfolds 2 mm), the observers again independently measured and recorded a second and, if necessary, a third set of readings for the affected variable(s). The availability of duplicate measurements by two observers allows for the estimation of inter-observer reliability statistics under routine data collection conditions. Since weight was measured with near-perfect precision on digital scales, it was not included in the standardization sessions.

During the standardization sessions, screening teams measured newborns while follow-up teams measured older infants. The children involved in the standardization sessions were not part of the MGRS cohort. During these sessions, the observers measured independently but did not compare values with other

observers, as was done during routine data collection. No inter-site statistical comparisons are presented because no common set of children was measured by observers from different sites. At each site, the screening teams' standardization sessions stopped when the enrolment of newborns ended (duration 12–14 mo), while the follow-up team sessions continued for the entire 3–3½ y of MGRS data collection. Because the USA site did not have access to newborns for the screening team's standardization exercises, the team measured older infants.

Data management

The MGRS data management protocol, which has been described in detail elsewhere [18], highlights the specific measures applied in detecting errors and cleaning the MGRS anthropometry data. For the standardization sessions, study supervisors in each site were responsible for checking the data collected for any recording errors prior to on-site analysis of measurement error. The data were then sent to the study coordinating centre in Geneva, Switzerland, for further quality control checks and monitoring of the performance of observers and site teams. These data were merged within site to create the standardization master files used in the present analyses. Recorded values that varied by more than 4 standard deviations from a given child's mean (estimated from all values recorded by the observers in the session) were considered errors of transcription or the result of causes unrelated to measurement reliability and were reset as missing [8]. For the purpose of this report, data were analysed only from observers who participated in two or more standardization sessions.

Statistical analysis

Reliability statistics reported for the standardization sessions were intra-observer technical error of measurement (TEM), inter-observer TEM and average bias. Inter-observer TEM achieved in routine data collection was also estimated and used to calculate the coefficient of reliability, R , for six anthropometric variables (excluding weight) measured in the MGRS. The key statistics are defined as follows.

Technical error of measurement (TEM) is a measure of error variability that carries the same measurement units as the variable measured, e.g. centimetres of head circumference. Its interpretation is that differences between replicate measurements will be within \pm the value of TEM two-thirds of the time [14]. Similarly, 95% of the differences between replicate measurements are expected to be within $\pm 2 \times$ TEM [9], which is referred to as the 95% precision margin elsewhere in this paper. Intra-observer TEM is estimated from differences between replicate

measurements taken by one observer, while inter-observer TEM is estimated from single measurements taken by two or more observers. The formulae (1)–(4) for these statistics are given below.

Intra-observer TEM for one observer is calculated by:

$$\sqrt{\frac{\sum_{i=1}^N (M_{i1} - M_{i2})^2}{2*N}}, \quad (1)$$

where M_{i1} and M_{i2} are the duplicate measurements recorded by a given observer for the i^{th} child, and N is the number of children measured. It can be generalized to k observers as in (2):

$$\sqrt{\frac{\sum_{j=1}^K \sum_{i=1}^{N_j} (M_{ij1} - M_{ij2})^2}{2*\sum_{j=1}^K N_j}}, \quad (2)$$

where M_{ij1} and M_{ij2} are the duplicate readings recorded by observer j for the i^{th} child, N_j is the number of children measured by observer j , and K is the number of observers taking the measurements.

The inter-observer TEM in standardization data is calculated by:

$$\left\{ \frac{1}{N} \sum_{i=1}^N \frac{1}{(K_i - 1)} \left[\sum_{j=1}^{K_i} Y_{ij}^2 - \frac{\left(\sum_{j=1}^{K_i} Y_{ij} \right)^2}{K_i} \right] \right\}^{1/2}, \quad (3)$$

where Y_{ij} is one of the duplicate measurements taken by observer j for child i (for simplicity in programming the present analyses, the first recorded measurement was selected), K_i is the number of observers that measured child i (this takes care of missing values), and N is the number of children involved. In the routine MGRS data (calculated separately for screening, longitudinal follow-up and cross-sectional survey data), only two observers took measurements, so formula (3) simplifies to:

$$\left\{ \frac{1}{N} \sum_{i=1}^N \left[\sum_{j=1}^2 Y_{ij}^2 - \left(\sum_{j=1}^2 Y_{ij} \right)^2 / 2 \right] \right\}^{1/2}, \quad (4)$$

where N is the total number of children measured in respective master files for each anthropometric variable.

Average bias is estimated as the average difference between measurements taken by an expert and those taken by an observer or observers of the same subjects. A negative-signed average bias estimate indicates that the test group underestimates the measurement, while the opposite indicates overestimation. It is calculated by:

$$\frac{\sum_{i=1}^{N_G} \left[\sum_{j=1}^K (M_{ij1} + M_{ij2}) / (2*K) - (M_{iG1} + M_{iG2}) / 2 \right]}{N_G}, \quad (5)$$

where M_{ij1} , M_{ij2} and M_{iG1} , M_{iG2} are the duplicate readings recorded by observer j and the expert, respectively, for the i^{th} child, N_G is the set of children measured by the expert, and K is the number of observers measuring the same children.

Coefficient of reliability, R , estimates the proportion of the inter-subject variance (total measurement variance) that is not due to measurement error. A reliability coefficient $R=0.8$ means that 80% of the total variability is true variation, while the remaining proportion (20%) is attributable to measurement error, described by Marks and colleagues [8] as imprecision and unreliability. For the MGRS data, R was calculated using the formula:

$$R = 1 - \frac{(TEM(Inter))^2}{SD^2}, \quad (6)$$

where $TEM(Inter)$ refers to the MGRS data TEM as calculated in formula (4), and SD values for each anthropometric variable are taken from the MGRS population at specified ages. For newborns: head circumference 1.27 cm and length 1.91 cm; and for older children: head circumference 1.40 cm, length 2.60 cm, arm circumference 1.30 cm, triceps skinfold 1.80 mm, subscapular skinfold 1.40 mm (12 mo), and height 4.07 cm (48 mo).

In the MGRS, intra-observer TEM could be calculated for the standardization but not the routine study data, while inter-observer TEM was calculated for both data sets. Intra-observer TEM for each team was calculated using data from all the standardization sessions conducted in a given site. The MGRS anthropometry experts' measurements from all sites were combined to calculate the "gold standard" intra-observer TEM. The assessment of bias was restricted to the data collected during the standardization sessions in which an international lead anthropometrist participated.

Several approaches were used to judge the adequacy of measurement in the MGRS, consistent with guidelines suggested in the literature:

- TEM values for observers were considered adequate if they were within ± 2 times the expert's TEM, i.e. the expert's 95% precision margin [19].
- We assessed average bias in terms of magnitude and whether or not site teams systematically over- or underestimated measurements. To be consistent with the criterion used to set the maximum allowable differences between paired observer measurements in the MGRS, bias was

considered to be large if it exceeded the expert's intra-observer TEM $\times 2.8$ [1]. This is equivalent to the limits that were considered to indicate significant deviations from likely "true" values while accommodating the unavoidable imprecision of anthropometric measurements.

- c. Our main criterion for judging adequacy of measurement was the coefficient of reliability, R , because it considers the measurement variance in relation to variability in the measurement. As is the case for other related measures of agreement, e.g. kappa, values of 0.8 and greater may be taken to represent "excellent" agreement and those between 0.61 and 0.8 "substantial" agreement [20].
- d. Finally, we compared the TEMs obtained by the MGRS observers to those reported in the literature.

Results

The number of standardization sessions at each site ranged from five to nine for the screening teams and 14 to 21 for the follow-up teams (Table I). There was also inter-site variation in the number of observers, which was a function of staff turnover (Ghana had the highest turnover and Oman the lowest). The MGRS anthropometry experts participated in 17 of the standardization sessions.

Screening team

Intra-observer TEMs ranged among sites from 0.16 to 0.28 cm for newborn head circumference and from 0.22 to 0.48 cm for length measurements (Table II). In all cases, observer TEMs were within twice the gold standard TEM, that is, within the 95% precision margin. While there was no evidence of bias in the teams' head circumference measurements compared with the expert's, all four sites for which bias was calculated tended to underestimate length, by -0.21 to -0.37 cm.

Inter-observer TEMs for both the standardization and the routine data collected by the screening teams are presented in Table III. TEMs were very similar for the two data sources. Reliability coefficients, estimated for routine data collection, were greater than 95% in all cases. Inter-observer TEMs were not substantially larger than intra-observer TEMs (Table III versus Table II).

Follow-up team

In almost all cases, the follow-up teams' intra-observer TEMs were less than twice the gold standard TEM (Table IV). Only the Norwegian and Omani teams' TEMs exceeded the expert's 95% precision margin (0.24 cm) for head circumference. All bias estimates but one (Brazil, subscapular skinfold) were within the allowable limits of 2.8 times the gold standard TEM for each measurement. However, the sign of the teams' bias estimates showed that they tended to underestimate arm circumference, length and height, and to overestimate skinfold measurements. Estimates of bias in head circumference had a fair balance of positive and negative signs, and were of the lowest overall magnitude.

The three sets of data (standardization, longitudinal and cross-sectional) represented in Table V had similar inter-observer TEMs within each variable and site. The largest disparity in this regard was for triceps skinfold in India with 0.49 mm for the standardization and 0.71 mm for the longitudinal data. The coefficient of reliability was above 0.95 for all variables except the skinfolds for which R ranged from 0.75 to 0.93. A comparison of inter- and intra-observer TEM based on the standardization data revealed very few substantial differences. The expected pattern (inter-observer TEM larger than intra-observer TEM) was systematic for two measurements (the skinfolds) in all sites, and for all measurements in two sites (Brazil and the USA).

The reliability of both newborn and older-child measurements for the MGRS teams was as good as,

Table I. Standardization sessions and observer participation by team and site.

Sites	Newborn screening team			Follow-up team		
	Sessions ^a	Observers	Expert ^b	Sessions ^a	Observers	Expert ^b
Brazil	6	6	0	20	9	1
Ghana	8	9	2	21	15	4
India	9	9	2	19	10	3
Norway	5	5	1	14	9	3
Oman	9	6	3	19	11	4
USA	0 ^c	–	–	17	9	2

^a The screening team sessions are fewer than the follow-up team sessions because newborn screening for the longitudinal study lasted 12–14 mos while the follow-up team worked through the entire 3–3½ y of data collection.

^b These are the sessions in which one of the MGRS international lead anthropometrists participated.

^c The USA did not have access to newborns for the standardization sessions, so the screening team measured older infants.

Table II. Screening team intra-observer technical error of measurement (TEM)^a and bias^b relative to MGRS anthropometry expert in the standardization sessions.

		Site ^c					
		Expert	Brazil (<i>n</i> =20, 60) ^d	Ghana (<i>n</i> =95)	India (<i>n</i> =99)	Norway (<i>n</i> =60)	Oman (<i>n</i> =102)
Head circumference (cm)	TEM	0.16	0.24	0.25	0.16	0.28	0.27
	Average bias	–	–	0.00	–0.09	0.08	0.03
Length (cm)	TEM	0.29	0.22	0.29	0.33	0.48	0.37
	Average bias	–	–	–0.29	–0.21	–0.37	–0.26

^a The expert's TEM is based on the sum of measurements taken in all sites by the MGRS lead anthropometrists participating in standardization sessions. Site teams' intra-observer TEM is calculated using data from all standardization sessions (initial and bimonthly) conducted in respective sites, average of all observers taking part in ≥ 2 bimonthly sessions.

^b Average bias relative to the expert is calculated from the subset of measurements taken in the standardization sessions in which the MGRS lead anthropometrist participated, and thus includes only subjects measured by both the expert and each site's team (*n* per site: Ghana 31; India 30; Norway 20; Oman 42; Brazil did not hold a separate session for the newborn screening team at the initial standardization where the lead anthropometrist participated).

^c The USA was excluded from this analysis because the screening team did not measure newborns in the standardization sessions.

^d Sample size: *n* = 20 infants for head circumference and *n* = 60 for length. The earliest enrolled newborns in Brazil had their first head circumference measurement taken at 7 d. The MGRS protocol was amended, and only then did the screening team begin to take head circumference measurements at birth.

or better than, intra-observer TEM estimates reported in other published studies involving children (Table VI).

Discussion

The measurement and standardization protocols of the MGRS provided a mechanism for continuous monitoring of measurement reliability. This helped to identify and resolve problems by retraining individual observers (during or immediately after each standardization session) or site teams, as happened on specific occasions in Norway and the USA. The sources of error in the MGRS were identified with the express intention of correcting them, going beyond what has been implemented in other studies that documented measurement reliability [5,9,11]. A further unique feature of the MGRS is the documentation of measurement reliability in the very data that

have been used to construct the WHO Child Growth Standards [21].

The standardization sessions and routine data collection settings are difficult to compare. In the former, workers had to collect duplicate measurements on 10 to 20 children in one session and were not allowed to compare and take new measurements when differences were large. In routine data collection, fieldworkers were dealing with just one child at a time and were allowed to compare their values and re-measure if disparities exceeded preset limits. Despite these differences, measurement error was similar in both settings.

A comparison of reliability statistics between the screening and follow-up teams, and between the longitudinal and cross-sectional samples, shows that newborn and older infants were measured as reliably as were older children. Judging by the site teams' intra-observer TEM relative to the expert's 95%

Table III. Inter-observer technical error of measurement (TEM) for the newborn screening teams in the standardization sessions and routine MGRS data collection.

		Site						
		Data source and R coefficient ^a	Brazil ^b (<i>n</i>)	Ghana (<i>n</i>)	India (<i>n</i>)	Norway (<i>n</i>)	Oman (<i>n</i>)	USA (<i>n</i>)
Head circumference (cm)	Standardization	0.42 (20)	0.27 (95)	0.20 (99)	0.25 (60)	0.26 (102)	–	–
	MGRS data	–	0.25 (329)	0.18 (301)	0.24 (300)	0.25 (295)	0.27 (208)	0.22 (1433)
	R coefficient	–	0.96	0.98	0.96	0.96	0.95	0.97
Length (cm)	Standardization	0.32 (60)	0.35 (95)	0.42 (99)	0.48 (60)	0.40 (102)	–	–
	MGRS data	–	0.30 (329)	0.35 (301)	0.39 (300)	0.39 (295)	0.40 (208)	0.34 (1433)
	R coefficient	–	0.98	0.97	0.96	0.96	0.96	0.97

^a Inter-observer TEM was calculated separately for the standardization and routine screening data of the MGRS longitudinal component. The R coefficient is calculated for the latter data set only.

^b Inter-observer TEM and R were not calculated for the Brazilian newborn screening data because the site began to duplicate measurements halfway into recruitment. The early data were thus inappropriate for this analysis.

Table IV. Follow-up team intra-observer technical error of measurement (TEM)^a and bias^b relative to MGRS anthropometry expert in the standardization sessions.

	Expert	Site ^c					
		Brazil (<i>n</i> = 210, 0)	Ghana (<i>n</i> = 234, 138)	India (<i>n</i> = 200, 160)	Norway (<i>n</i> = 162, 80)	Oman (<i>n</i> = 200, 90)	USA (<i>n</i> = 179, 69)
Head circumference (cm)	TEM	0.13	0.23	0.19	0.25	0.29	0.19
	Average bias	0.01	-0.01	-0.16	0.04	-0.18	-0.14
Length (cm)	TEM	0.23	0.37	0.33	0.58	0.43	0.21
	Average bias	0.01	-0.18	-0.15	-0.35	-0.24	-0.70
Arm circumference (cm)	TEM	0.17	0.20	0.20	0.26	0.27	0.15
	Average bias	-0.10	-0.30	-0.24	-0.31	-0.26	-0.37
Triceps skinfold (mm)	TEM	0.42	0.39	0.46	0.61	0.49	0.45
	Average bias	-0.81	0.21	0.45	0.11	0.25	0.11
Subscapular skinfold (mm)	TEM	0.38	0.31	0.32	0.29	0.35	0.41
	Average bias	-1.05	0.28	0.28	0.11	0.03	0.79
Height (cm)	TEM	-	0.26	0.27	0.29	0.27	0.16
	Average bias	-	-0.30	-0.21	-0.20	-0.22	-0.06

^a The expert's TEM is based on the sum of measurements taken in all sites by the MGRS lead anthropometrists participating in standardization sessions. Site teams' intra-observer TEM is calculated using data from all standardization sessions (initial and bimonthly) conducted in respective sites, average of all observers taking part in ≥ 2 bimonthly sessions.

^b Average bias relative to the expert is calculated from the subset of measurements taken in the standardization sessions in which the MGRS lead anthropometrist participated, and thus includes only subjects measured by both the expert and each site's team (*n* per site (*n* height): Brazil 19 (0); Ghana 60 (40); India 40 (30); Oman 50 (30); USA 19 (9)).

^c The second sample size figure is the number of subjects involved in height standardization. Sites normally began to take this measurement at the inception of the cross-sectional study.

Table V. Inter-observer technical error of measurement (TEM) for the follow-up teams in the standardization sessions and the routine MGRS data.

	Data source and R coefficient ^a	Site						
		Brazil (<i>n</i>)	Ghana (<i>n</i>)	India (<i>n</i>)	Norway (<i>n</i>)	Oman (<i>n</i>)	USA (<i>n</i>)	All (<i>n</i>)
Head circumference (cm)	Standardization	0.25	0.24	0.18	0.23	0.29	0.28	–
	Longitudinal	0.23 (5849)	0.23 (6069)	0.23 (5633)	0.25 (5460)	0.26 (5425)	0.23 (3834)	0.24 (32270)
	Cross-sectional	0.25 (1342)	0.23 (1406)	0.21 (1455)	0.24 (1376)	0.29 (1445)	0.28 (1339)	0.25 (8363)
	R coefficient	0.97/0.97	0.97/0.97	0.97/0.98	0.97/0.97	0.97/0.96	0.97/0.96	0.97/0.97
Length (cm)	Standardization	0.40	0.44	0.32	0.48	0.42	0.37	–
	Longitudinal	0.33 (5836)	0.41 (6067)	0.36 (5630)	0.37 (5470)	0.38 (5420)	0.41 (3827)	0.38 (32250)
	Cross-sectional	0.23 (250)	0.34 (286)	0.27 (327)	0.27 (371)	0.35 (356)	0.29 (164)	0.30 (1754)
	R coefficient	0.98/0.99	0.98/0.98	0.98/0.99	0.98/0.99	0.98/0.98	0.98/0.99	0.98/0.99
Arm circumference (cm)	Standardization	0.28	0.27	0.19	0.29	0.26	0.26	–
	Longitudinal	0.26 (4545)	0.25 (4791)	0.26 (4461)	0.29 (4267)	0.23 (4550)	0.26 (3002)	0.26 (25616)
	Cross-sectional	0.22 (1333)	0.20 (1406)	0.18 (1448)	0.22 (1354)	0.21 (1444)	0.23 (1339)	0.21 (8324)
	R coefficient	0.96/0.97	0.96/0.98	0.96/0.98	0.95/0.97	0.97/0.97	0.96/0.97	0.96/0.97
Triceps skinfold (mm)	Standardization	0.67	0.51	0.49	0.83	0.60	0.87	–
	Longitudinal	0.66 (4638)	0.50 (4791)	0.71 (4464)	0.75 (4259)	0.63 (4551)	0.83 (3002)	0.67 (25705)
	Cross-sectional	0.85 (1325)	0.46 (1406)	0.67 (1440)	0.76 (1328)	0.59 (1444)	0.84 (1335)	0.70 (8278)
	R coefficient	0.87/0.78	0.92/0.93	0.85/0.86	0.83/0.82	0.88/0.89	0.79/0.78	0.86/0.85
Subscapular skinfold (mm)	Standardization	0.48	0.42	0.36	0.42	0.41	0.67	–
	Longitudinal	0.47 (4639)	0.42 (4791)	0.45 (4466)	0.43 (4273)	0.46 (4551)	0.69 (3003)	0.48 (25723)
	Cross-sectional	0.59 (1324)	0.44 (1406)	0.44 (1434)	0.39 (1339)	0.49 (1444)	0.62 (1335)	0.50 (8282)
	R coefficient	0.89/0.82	0.91/0.90	0.89/0.90	0.91/0.92	0.89/0.88	0.75/0.80	0.88/0.87
Height (cm)	Standardization	–	0.27	0.23	0.34	0.35	0.33	–
	Cross-sectional	0.15 (1328)	0.39 (1404)	0.23 (1449)	0.34 (1358)	0.26 (1443)	0.32 (1348)	0.29 (8330)
	R coefficient	1.00	0.99	1.00	0.99	1.00	0.99	0.99

^a“Longitudinal” are the data measured by the follow-up team in the MGRS longitudinal component, and “cross-sectional” are data from the MGRS cross-sectional component. The reliability coefficient R was based on the routine MGRS (not standardization) data: the first figure belongs to the longitudinal measurements and the second to the cross-sectional measurements, and the single figure for height refers to the cross-sectional component.

precision margins, the teams’ precision compared favourably with the expert’s for all measurements. There was no consistent pattern in the relationship between intra- and inter-observer variability.

Although the magnitude of bias in the teams’ measurements was overall within allowable limits compared with the expert, distinct negative and positive tendencies were noticeable for all measurements except head circumference. The “problem” measurements were those that involve manipulation of soft tissues (arm circumference and skinfolds) and those that require careful positioning to ensure that the child is fully stretched out for the measurement (length and height). It is worth noting that the same pattern was observed in the Rotterdam standardization session [1] where, compared with the expert, the session’s participants all had negative-signed bias for length, height and arm circumference, and positive-signed bias for the skinfold measurements. In general, the standardization sessions were stressful as the observers had to repeat measurements on often crying and struggling children. Under those conditions, the expert could, with greater self-assurance than the fieldworkers, position the child to full length/height, pause to let the callipers close in on skinfolds before taking the reading, and retain better control of the circumference tape around the child’s arm to avoid compressing the soft tissues. The average bias estimate for subscapular skinfold in Brazil was larger than

the limits set by the expert’s $TEM \times 2.8$ and also in the opposite direction from the other sites. The data used to calculate this estimate were collected at the site’s initial standardization, and the team thereafter received remedial training in the measurement of skinfolds.

Considering our main criterion for assessing measurement reliability in the MGRS data, overall R coefficients were higher than the 90% reliability threshold that Marks and colleagues [8] suggest as adequate for the presentation of growth standards. However, Ulijaszek and Lourie [22], while endorsing that cut-off, recognized the characteristic low reliability of skinfold measurements in young children. Indeed, the MGRS skinfold measurements had R coefficients below 90% but mostly above the threshold of 80% applied to other measures of agreement such as the kappa coefficient cut-off for “excellent” agreement [20]. As others have noted, larger inter-observer reliability is expected in measurements that have characteristically low precision [8]. This is illustrated by the lower intra- than inter-observer TEM for the two skinfold measurements in the MGRS. One suggested approach to improving precision for such measurements is to measure twice and report the average of the two values [5,8]. This is what we did in the MGRS, for all the anthropometric measurements used to construct the WHO Child Growth Standards, with the added assurance that the

Table VI. Comparison of intra-observer TEM between the MGRS and other estimates in the literature (child populations).

Age group and variables	MGRS teams	Published estimates	Source (number in ref. list)
Newborn			
Length (cm)	0.22–0.48	0.79, 1.22	Johnson et al., 1997 [23]
Head circumference (cm)	0.16–0.28	0.28, 0.30	Johnson et al., 1997 [23]
Older children			
Length (cm)	0.23–0.58	0.4, 0.8	Ulijaszek and Lourie, 1994, literature review [22]
Height (cm)	0.16–0.29	0.34	Martorell et al., 1975 [6]
		0.49	Malina et al., 1973, NHES III [5]
Head circumference (cm)	0.13–0.29	0.14	Martorell et al., 1975 [6]
MUAC (cm)	0.15–0.27	0.35	Malina et al., 1973, NHES III [5]
		0.18	Martorell et al., 1975 [6]
Triceps skinfold (mm)	0.39–0.61	0.47	Martorell et al., 1975 [6]
		0.80	Johnston et al., 1972, NHES III [24]
Subscapular skinfold (mm)	0.29–0.41	0.27	Martorell et al., 1975 [6]
		1.83	Johnston et al., 1972, NHES III [24]

NHES III: cycle III of the National Health Examination Survey (USA); MUAC: mid-upper arm circumference.

two measurements were within preset margins of difference [1].

Several published studies and reviews of the anthropometry literature provided intra-observer TEM estimates, and these were compared with the MGRS teams' performance [5,6,22–24]. The MGRS teams' reliability was generally better than the published ranges. However, these comparisons should be viewed with the understanding that the numbers of observers and subjects involved, and the measurement protocols and equipment employed, vary widely among studies. For example, the number of subjects measured in the MGRS standardization sessions is larger than has been reported in most other published studies.

The MGRS presents a number of innovations with regard to reliability assessment in anthropometry. These include the use of standardized measurement protocols and equipment at six country sites, the evaluation of the different site teams' reliability using a common gold standard, and the estimation of measurement reliability in the data that have been used to construct growth standards. Ulijaszek and Kerr [15] proposed using "criterion anthropometrist(s)" for the purpose of overseeing and assuring the standard application of measurement procedures, and to set targets for the level of accuracy that fieldworkers in anthropometry could aim to achieve. The use in the MGRS of the international lead anthropometrist's intra-observer TEM to set cut-offs for precision (the expert's 95% precision margin) and the limits of acceptable bias (2.8 times the expert's TEM) is a significant step in this direction, and one that could be applied in other studies to standardize reliability assessment when a gold standard is available. In the absence of a designated individual to serve as gold standard, the average intra-observer TEM of a well-trained group could be used to set both precision and accuracy targets.

Acknowledgements

This paper was prepared by Adelheid W. Onyango, Reynaldo Martorell, Wm Cameron Chumlea, Jan Van den Broeck, Cora L. Araújo, Anne Baerug, William B. Owusu and Roberta J. Cohen on behalf of the WHO Multicentre Growth Reference Study Group. The statistical analysis was conducted by Alain Pinol and Elaine Borghi.

References

- [1] de Onis M, Onyango AW, Van den Broeck J, Cameron WC, Martorell R, for the WHO Multicentre Growth Reference Study Group. Measurement and standardization protocols for anthropometry used in the construction of a new international growth reference. *Food Nutr Bull* 2004;25 Suppl 1:S27–36.
- [2] Garn S, Shamir Z. *Methods for research in human growth*. Springfield: Charles C. Thomas; 1958.
- [3] Damon A, Stout H, McFarland R. *The human body in equipment design*. Cambridge: Harvard University Press; 1966.
- [4] McGammon R. *Human growth and development*. Springfield: Charles C. Thomas; 1970.
- [5] Malina RM, Hamill PV, Lemeshow S. Selected body measurements of children 6–11 years: United States. *Vital Health Stat Series 11 No. 123*, 1973:38–48.
- [6] Martorell R, Habicht JP, Yarbrough C, Guzman G, Klein RE. The identification and evaluation of measurement variability in the anthropometry of preschool children. *Am J Phys Anthropol* 1975;43:347–52.
- [7] Foster TA, Berenson GS. Measurement error and reliability in four pediatric cross-sectional surveys of cardiovascular disease risk factor variables—the Bogalusa Heart Study. *J Chronic Dis* 1987;40:13–21.
- [8] Marks GC, Habicht JP, Mueller WH. Reliability, dependability, and precision of anthropometric measurements. The Second National Health and Nutrition Examination Survey 1976–1980. *Am J Epidemiol* 1989;130:578–87.
- [9] Chumlea WC, Guo S, Kuczmarski RJ, Johnson CL, Leahy CK. Reliability of anthropometric measurements in the Hispanic Health and Nutrition Examination Survey (HHANES 1982–1984). *Am J Clin Nutr* 1990;51:902S–7S.

- [10] Roche AF. Growth, maturation and body composition: The Fels longitudinal study 1929–1991. New York: Cambridge University Press; 1992.
- [11] Moreno LA, Joyanes M, Mesana MI, González-Gross M, Gil CM, Sarria A, et al. Harmonization of anthropometric measurements for a multicenter nutrition survey in Spanish adolescents. *Nutrition* 2003;19:481–6.
- [12] Lohman TG, Roche AF, Martorell R, editors. Anthropometric standardization reference manual. Champaign: Human Kinetics Books; 1988.
- [13] Habicht JP, Yarbrough C, Martorell R. Anthropometric field methods: criteria for selection. In: Jelliffe DB, Jelliffe EFP, editors. *Nutrition and growth*. New York: Plenum Press; 1979. p. 365–87.
- [14] Mueller WH, Martorell R. Reliability and accuracy of measurement. In: Lohman TG, Roche AF, Martorell R, editors. *Anthropometric standardization reference manual*. Champaign: Human Kinetics Books; 1988. p. 83–6.
- [15] Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. *Br J Nutr* 1999;82: 165–77.
- [16] Johnson TS, Engstrom JE. State of the science in measurement of infant size at birth. *Newborn Infant Nurs Rev* 2002;2: 150–8.
- [17] de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martinez J, for the WHO Multicentre Growth Reference Study Group. The WHO Multicentre Growth Reference Study: Planning, study design and methodology. *Food Nutr Bull* 2004;25 Suppl 1:S15–26.
- [18] Onyango AW, Pinol AJ, de Onis M, for the WHO Multicentre Growth Reference Study Group. Managing data for a multi-country longitudinal study: Experience from the WHO Multicentre Growth Reference Study. *Food Nutr Bull* 2004;25 Suppl 1:S46–52.
- [19] WHO. *Measuring change in Nutritional Status*. Geneva: World Health Organization; 1983.
- [20] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
- [21] WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatr Suppl* 2006;450:76–85.
- [22] Ulijaszek SJ, Lourie JA. Intra- and inter-observer error in anthropometric measurement. In: Ulijaszek SJ, Mascie-Taylor CGN, editors. *Anthropometry: the individual and the population*. Cambridge: Cambridge University Press; 1994. p. 30–55.
- [23] Johnson TS, Engstrom JL, Gelhar DK. Intra- and inter-examiner reliability of anthropometric measurements of term infants. *J Pediatr Gastroenterol Nutr* 1997;24:497–505.
- [24] Johnston FE, Hamill PVV, Lemeshow S. Skinfold thickness of children 6–11 years: United States. *Vital Health Stat Series 11* No. 120, 1972:50–60.