Reliability, technical error of measurements and validity of instruments for nutritional status assessment of adults in Malaysia

Geeta A , Jamaiyah H, Safiza M N, Khor G L, Kee C C, Ahmad A Z, Suzana S, Rahmah R, Faudzi A

ABSTRACT

Introduction: The Third National Health and Morbidity Survey Malaysia 2006 includes a nutritional status assessment of children. This study aimed to assess the inter- and intra-examiner reliability, the technical error of measurement and the validity of instruments for measuring weight, height and waist circumference.

Methods: A convenience sample of 130 adults working in a selected office setting was chosen to participate in the study, subject to the inclusion and exclusion study criteria. Two public health nurses, trained to follow a standard protocol, obtained the weight, height and waist circumference measurements. The weight was measured using the Tanita HD-318 digital weighing scale to the nearest 0.1 kg, and Seca Beam Scale to the nearest 0.01 kg. The height was measured using the Seca Bodymeter 206 and Stadiometer, both to the nearest 0.1 cm. The waist circumference was measured using the Seca circumference measuring tape S 201, to the nearest 0.1 cm.

Results: The intra-examiner reliability in descending order was weight and height followed by waist circumference. The height measurement, on average, using the test instrument, reported a recording of 0.4 cm higher than the reference instrument, with the upper and lower limits at 2.5 cm and 1.6 cm, respectively. The technical error of measurement and coefficient of variation of weight and height for both inter-examiner and intra-examiner measurements were all within acceptable limits (below five percent).

Conclusion: The findings of this study suggest that weight, height and waist circumference measured in adults aged 18 years and above, using the respective abovementioned instruments, are reliable and valid for use in a community survey. Limiting the number of examiners, especially for waist circumference measurements, would yield a higher degree of reliability and validity.

Keyword: anthropometry, height measurement, nutritional status assessment, waist circumference measurement, weight measurement

Singapore Med J 2009;50(10):1013-1018

INTRODUCTION

Anthropometry is a relatively simple, quick and inexpensive means of nutritional assessment that can be used in clinical and community settings, as well as for research in laboratories and field facilities.[1] However, anthropometry methods have inherent limitations, such as the need for trained observers, and the relatively high between-measurement technical and mechanical limitations.[2] Among the various measurement methods, anthropometry techniques usually demonstrate the largest standard error and the lowest correlation coefficients when compared against other techniques. Various terms are used to describe anthropometric measurement error. These include the terms, reliability and validity.[1] Reliability is the degree to which within-subject variability is present and is due to factors other than the variance of measurement error or physiological variation. The second type of measurement error, validity, is the extent to which the "true" value of a measurement is attained.

There are many articles describing anthropometric assessment methods and interpretation; however, there are very few that discuss reliability and validity issues and the extent to which these factors can influence both the measurement and interpretation of nutritional status.[1] This study was designed to address the issue of the lack of evidence on the reliability and validity of instruments used for weight (WT), height (HT) and waist circumference

Clinical Research Centre, Kuala Lumpur Hospital, Jalan Pahang, Kuala Lumpur 50586, Malaysia

Geeta A, BSc, MSc
Research Officer

Jamaiyah H, MD, MCommHealth, MSc
Head, Clinical Epidemiology Unit

Institute of Public Health, Ministry of Health Malaysia, Jalan Bangsar, Kuala Lumpur 50590, Malaysia

Safiza MN, BSc, MSc
Dietician

Ahmad AZ, BSc, MSc
Dietician

Faudzi A, MD, MPH
Epidemiologist

Department of Nutrition and Dietetics, Faculty of Medicine and Health Sciences, University Putra Malaysia, Serdang 43400, Malaysia

Khor GL, BSc, MSc, PhD
Professor

Epidemiology and Biostatistics Unit, Institute for Medical Research, Jalan Pahang, Kuala Lumpur 50588, Malaysia

Kee CC, BSc
Nutritionist

Department of Nutrition and Dietetics, Faculty of Allied Health Sciences, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abd Aziz, Kuala Lumpur 50300, Malaysia

Suzana S, BSc, MSc, PhD
Associate Professor

Department of Paediatrics, Universiti Kebangsaan Medical Centre, Jalan Tenteran, Bandar Tun Razak, Kuala Lumpur 56000, Malaysia

Rahmah R, MD, MMed
Associate Professor

Correspondence to:
Ms Geeta Appannah
Tel: (60) 3 4044 3060
Fax: (60) 3 4044 3080
Email: geeta@crc.gov.my

Table I. Mean, median and absolute mean difference for inter- and intra-examiner reliability.

Summary statistics	Inter-examiner		Intra-examiner		Absolute mean difference
	Observer 1 (1)	Observer 2 (2)	Time 1 (3)	Time 2 (4)	(1)−(2) / (3)−(4)
Waist circumference (cm)					
No. of subjects	130	129*	130	130	
Mean ± SD	83.9 ± 12.11	84.0 ± 12.3	83.9 ± 12.1	84.0 ± 12.1	−0.1
Median (range)	82.2 (62.7–123.1)	82.2 (62.3–124.0)	82.2 (62.7–123.1)	82.3 (62.8–122.1)	
Weight (kg)					
No. of subjects	130	130	130	130	
Mean ± SD	59.7 ± 13.7	59.7 ± 13.6	59.7 ± 13.7	59.9 ± 13.6	0.0
Median (range)	57.1 (36.6–122.4)	57.2 (36.6–122.5)	57.1 (36.6–122.4)	57.1 (36.6–122.4)	
Height (cm)					
No. of subjects	130	130	130	130	
Mean ± SD	157.2 ± 8.5	157.2 ± 8.4	157.2 ± 8.5	157.2 ± 8.4	0.0
Median (range)	156.0 (138.1–181.7)	155.9 (137.7–181.6)	156.0 (138.1–181.7)	155.9 (137.9–181.6)	

*one subject refused to participate
SD: standard deviation

(WC) measurements and their respective technical errors of measurement. These issues need to be established prior to the actual use of these instruments for adults aged 18 years and older in the Third National Health and Morbidity Survey (NHMS III).

METHODS

This is a cross-sectional study, where a convenience sample of 130 working adults was taken from a selected office setting. The sample size was determined based on Walter et al's functional approximation method to calculate the required number of subjects in a reliability study.[3] The data collection was conducted in December 2005. The inclusion criteria were adults aged ≥ 18 years and the ability to stand upright. The exclusion criteria were pregnant mothers, mothers whose postnatal period was ≤ two months and subjects with obvious physical disability or body deformation that inhibits the ability to stand upright.

Two public health nurses conducted the measurements of HT, WT and WC. Each subject was examined four times for HT and WT, but only three times for WC. The study protocol was as follows: the first examiner measured the subject for HT, WT and WC; the same subject was then measured for the same variables by the second examiner; thereafter, the subject returned to the first examiner for a repeat measurement of the three variables; and lastly, the subject returned to the second examiner to measure only the HT and WT using the reference instruments. Both the examiners were not part of the study team, and hence were not aware of the study objectives. They were requested not to recall their previous readings. The data capture form was also designed in such a way that all the recordings of previous readings were obscured immediately after each recording, to minimise recall bias.

Both the examiners were trained to adhere to the standard procedure outlined in the Clinical Manual of the NHMS III. The choice of reference instruments was in accordance with the recommendations by the World Health Organization Expert Committee on Physical Status[4] and also on the basis that they were widely used in local health facilities. WT was measured to the nearest 0.1 kg accuracy, using the Tanita HD-318 digital weighing scales (Tanita Corporation, Tokyo, Japan) as the test instrument and the Seca Beam Balance 700 (Seca GmBH & Co Kg, Hamburg, Germany) as the reference instrument. HT was measured to the nearest 0.1 cm, from the subject's head to toe in an upright standing position with five points of his body touching the wall, using the Seca 206 mechanical measuring tape (Seca GmBH & Co Kg, Hamburg, Germany) as the test instrument and the Seca Stadiometer 282, (Seca GmBH & Co Kg, Hamburg, Germany) as the reference standard. WC was measured to the nearest 0.1 cm at the end of normal expiration for adults and elderly, using the Seca S 201 circumference measuring tape (Seca GmBH & Co Kg, Hamburg, Germany). The circumference tape was applied horizontally, midway between the lowest rib margin and the iliac crest, and the circumference over the waist was measured with the examiner seated in front of the subject.

For reliability and validity, the findings of the statistical analyses were reported using the absolute mean difference, correlation coefficient (r) and the Bland-Altman plot.[5] The r was computed to demonstrate the strength of the relationship (similarities) between two measurements. Intraclass correlation (ICC) was used for this purpose. The values for the reliability coefficient ranged from 0 to 1, where ICC < 0 indicated "no reliability", ≥ 0 but < 0.2 "slight reliability", 0.2 to < 0.4 "fair reliability", 0.4 to < 0.6 "moderate reliability", 0.6 to < 0.8 "substantial

Table II. Inter-and intra-examiner relative technical error of measurement classification results for weight and length measurements.

Measurement	TEM	% TEM	Classification[9] of % TEM	r
Weight				
Inter-examiner	0.45	0.75	Acceptable (≤ 2.0%)	0.999
Intra-examiner	0.31	0.53	Acceptable (≤ 1.5%)	0.999
Height				
Inter-examiner	0.32	0.20	Acceptable (≤ 2.0%)	0.996
Intra-examiner	0.32	0.20	Acceptable (≤ 1.5%)	0.999
Waist circumference				
Inter-examiner	0.77	0.91	Acceptable (≤ 2.0%)	0.999
Intra-examiner	0.44	0.52	Acceptable (≤ 1.5%)	0.999

TEM: technical error of measurement

reliability", and 1 "almost perfect reliability".[6] The Bland-Altman plot was used to provide an illustration of the spread of differences in the readings, the mean difference, and the upper and lower limits of agreement for both inter- and intra-examiner reliabilities. The validity was assessed using the findings from the statistical analysis, and also via the Bland-Altman plot. Inter-examiner reliability refers to the consistency of the readings of the same subjects between the two examiners, while intra-examiner reliability refers to the consistency of the readings of the same subject by the same examiner.

The technical error of measurement (TEM) is an accuracy index and represents the measurement quality and control dimension. It is the most common way to express the error margin in anthropometry and has been adopted by the International Society for the Advancement of Kinanthropometry for the accreditation of anthropometrists. It is essentially the standard deviation between repeated measures. The TEM index allows anthropometrists to verify the degree of accuracy when performing and repeating anthropometrical measurements (intra-examiner) and when comparing their measurement with measurements from other anthropometrists (inter-examiner). After the calculation of the relative TEM for both intra-examiner and inter-examiner variation analysis, the values were classified (Table II). The lower the TEM obtained, the better was the accuracy of the examiner in performing the measurement. In addition, the coefficient of variation (CV) was calculated to further determine the precision of the method of measurements. The CV provides a general "feeling" about the performance of a method. CVs ≤ 5% generally implied a good method performance, while CVs ≥ 10% did not.[7] The percentage of the CV is therefore a good indicator to use when comparing methods.[8]

RESULTS

The mean age and standard deviation of the 130 adults involved in this study was 36 ± 10.9 years, with a median (range) of 36 (18–64) years. The age distribution was not normally distributed, i.e. there were more respondents in the younger age group with a small p-value from the Shapiro-Wilk test. More than two-thirds were female. Malays formed the majority at 83%. In terms of education, the majority had a secondary to tertiary education. The mean, median and range of measurements and absolute mean difference for the first and second examiners are illustrated in Table I. There was no difference in the mean measurements of both WT and HT, and the absolute mean difference in the WC measurements was only 0.1 cm. This indicated a good agreement between the two examiners. The r results of the inter-examiner analysis using ICC were 0.9990 for WC, 0.9990 for WT and 0.9960 for HT.

The Bland-Altman plot was a useful means to reveal a relationship between the differences and the averages, to look for any systematic bias and to identify possible outliers. If there was a consistent bias, it could be adjusted by subtracting the mean difference using the findings from the test instrument. If the differences within the mean ± 1.96 standard deviation are not clinically important, the two methods may be used interchangeably. Some degree of random error was observed (Fig. 1). On the average, the WC measurement taken from the first examiner was consistently 0.2 cm higher than the second examiner across all the range of averages of their paired readings on the same subjects. The lower limit of difference was –1.9 cm, while the upper was 2.3 cm.

For WT measurements, most points seemed to cluster around the horizontal line of zero average (Fig. 2). However, because there were two extreme values, one each at –5 and + 5 kg, the lower and upper limits were pulled to –1.3 kg and +1.3 kg, respectively. In terms of HT, on average, the measurements taken by the first examiner were consistent with that of the second examiner across all the range of average values (Fig. 3). The lower limit of difference was –0.9 cm and the upper limit was 0.9 cm. There was some evidence of random error for HT measurements.

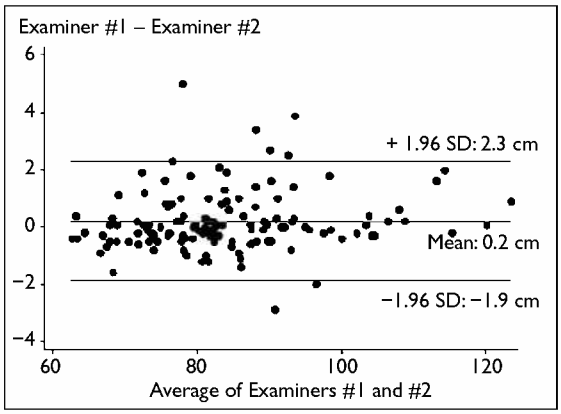

Fig. 1 Bland-Altman plot of the differences of waist circumference measurements of the examiners compared to the average of their paired readings.

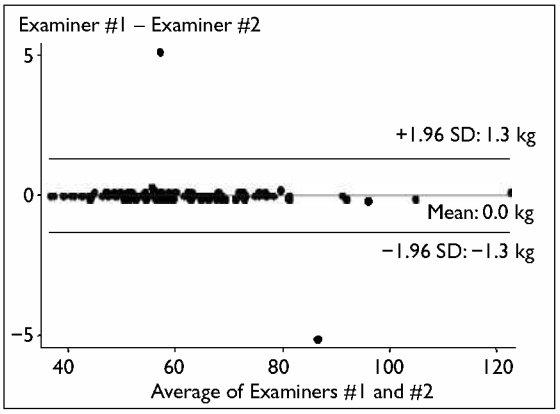

Fig. 2 Bland-Altman plot of the differences of weight measurements of the examiners compared to the average of their paired readings.

For intra-examiner analysis, the mean ± standard deviation, median and absolute mean differences are shown in Table I. For WC, the difference in the mean was too small to be detected, and for HT, there was no difference detected at all. In terms of r of the two readings from the same examiner on the same subject, the ICC was equal to or very close to one, indicating an almost perfect correlation.

For WC, on average, the first measurement taken was 0.1 cm higher than the second. The lower and upper limits of difference ranged from –1.1 cm to +1.3 cm. It was also noted that as the values of readings became bigger, the difference between readings was also larger. In terms of WT measurements, on average, the readings taken by the first examiner at Time 1 were consistent with that at Time 2. However, one outlier value had pulled the lower limit of difference to –0.9 kg and the upper limit to 0.9 kg. Examination of the HT measurements taken at both times showed that the readings of the two were consistent. However, the points were more dispersed; the lower limit of difference between the two was at –0.9 cm and the higher was at 0.9 cm.

The results for the TEM are tabulated in Table II. The respective relative TEMs for inter- and intra-examiners for WT were 0.75% and 0.53%, while that for HT were 0.20% and 0.20%, and for WC were 0.91% and 0.52%. This study also found that all the r values (for inter- and intra-examiners, WT, HT and WC) were almost 1.0, in accordance with the suggested cut-off.[1] This indicates that the human error for measurements in this study was small, all below the acceptable 5% mark.

The validity or accuracy was assessed by comparing these "test" instruments against that of the reference instruments. Only WT and HT measurements were assessed for the inter-instrument validity. There were very minimal differences (range 0.1–0.4) found in the

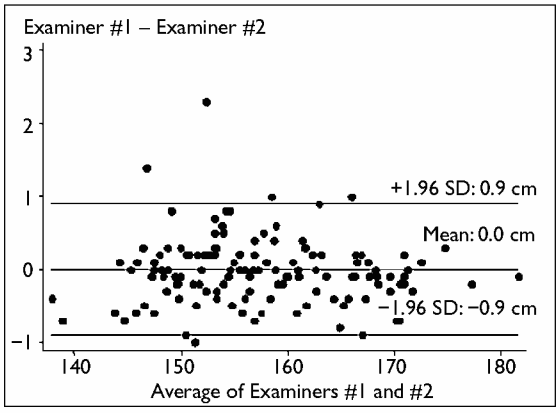

Fig. 3 Bland-Altman plot of the differences of the height measurements of the examiners compared to the average of their paired readings.

absolute mean between the two pairs of instruments, indicating a good agreement between the two (Table III). The ICC for WT and HT between the instruments was relatively high. On the average, the measurements taken using the test instrument were consistent with that of the reference instrument, with the lower and upper limits of agreement ranging from –2.0 kg to 2.0 kg for WT. For HT, the measurement taken from the test instrument was 0.4 cm higher than the reference instrument, and the limits of agreement ranged from –1.6 cm to 2.5 cm. This indicated the possibility of a systematic error in the HT measurement by using difference instruments.

Table IV shows that the various CVs were below 5%, which indicates that the readings obtained were not too varied in nature, across the three measurements for inter- and intra-examiners and across the two different instruments. However, it was also noted that the CV value was better if a single examiner was involved. One of the reasons the differences were observed in this study could be due to the lack of target training values imposed on the examiners prior to conducting the study. Training in itself

Table III. Mean, median, and absolute mean difference for inter-instrument validity.

Summary statistics	Test instrument (1)	Reference instrument (2)	Absolute mean difference (1) − (2)
Weight (kg)			
No. of subjects	130	130	
Mean and standard deviation	59.7 ± 13.6	59.8 ± 13.7	0.1
Median (range)	57.2 (36.6–122.5)	57.1 (36.6–122.8)	
Height (cm)			
No. of subjects	130	130	
Mean and standard deviation	157.2 ± 8.4	156.8 ± 8.3	0.4
Median (range)	155.9 (137.7–181.6)	155.3 (137.5–181.0)	

Table IV. Coefficient of variation of waist circumference, weight and height for the inter- and intra-examiner and inter-methods.

Variable	No. of subjects	Coefficient of variation (%)		
		Inter-examiner	Intra-examiner	Inter methods
Waist circumference (cm)	130	0.9	0.5	Not done
Weight (kg)	130	0.7	0.4	1.2
Height (cm)	130	0.2	0.2	0.5

is essential as it influences the degree of measurement error and interpretation, especially if there is a high inter-examiner variation. Besides training, some targets need to be set *a priori*, and the failure to achieve that would disqualify a person from being an examiner.

DISCUSSION

Overall, both the reliability and validity were good for all the measurements in this study. For inter-examiner measurements, the ICC was excellent for WT, HT and WC, which is in the order for the highest degree of reliability. The Bland-Altman plots showed that on average, the differences were small; WC 0.2, WT 0 and HT 0. The upper and lower limits of differences (± 2 cm for WC and WT, + 1 cm for HT) were also not clinically important. The WT measurements showed excellent results, with a zero difference across all average values in their Bland-Altman plot. There seemed to be no clear evidence of systematic bias for WT. However, for WC, there was a "funnel" effect seen in Fig. 1, which shows that the wider the WC, the bigger the difference in the measurements between the two examiners. For the HT measurement, there was some evidence indicating random error.

For intra-examiner reliability, the ICC coefficient showed a high degree of reliability (almost perfect). However, the Bland-Altman plot showed that within the examiner, WT and HT were consistent (mean = 0), while WC had a mean difference of 0.1. The lower and upper limits of difference were ± 1.3 cm for WC, + 1 kg for WT and ± 1 cm for HT, all of which were also not clinically important. For inter-instrument validity, the findings were excellent for both WT and HT. The mean (standard deviation) difference for WT was 0 (± 2) kg and that for HT was 0.4 (± 2.5) cm (taking the bigger value of the two limits); this indicates that the test instruments were almost comparable with the reference instruments, especially for WT measurements.

Technical errors of measurements in this study were minimal, and periodic controls, in terms of measurement quality, allowed for better accuracy and reliability in the measurement determinations.[9] Although it was found that there was some evidence of systematic bias towards the positive side in the HT measurement between the two instruments, the authors suggest for this to be noted and accepted, because the difference would not have any impact of clinical significance on the survey findings.

For WC, there are 14 different site descriptions for the site of measurement. Some of the methods used are slightly different from the others, and are thus organised into four groups by specific anatomical landmarks: (1) immediately below the lowest ribs; (2) at the narrowest waist; (3) the midpoint between the lowest rib and iliac crest; and (4) immediately above the iliac crest.[10] There is no universally-accepted method of measuring WC as the values at these four sites differ in magnitude, depending on the gender, and are highly reproducible and correlated with the total body and trunk adiposity in a gender-dependent manner.[10] A potential source of measurement error for all the WC sites is present, suggesting the need for training before and during data collection.

It is unfortunate that the authors were only able to find a few studies to enhance the write-up section in the

discussion. This is probably due to a lack of standardised terminology with which to describe the reliability of measurement in a clear manner.[11] The reference values for WT, HT and WC were also not reported in most of the publications as more emphasis was given to skin fold and breadth measurements.[11,12] In addition, most of the anthropometric measurement studies were done mainly in children, and hence, the intra- and inter-observer errors obtained in these studies may not be applicable to similar studies done on other age groups.[11] Due to the constraints of the recommended value of r in the literature, Himes suggested that researchers conduct their own reliability studies and determine the levels of r at their own discretion.[13]

In conclusion, this study found that the measurements of WC, WT and HT had a high degree of reliability, for both intra- and inter-examiner. The validities of WT and HT were also excellent. It was reassuring that the CV was found to be consistently below 5%. For both inter- and intra-examiner measurements, the order of the degree of reliability would be WT, HT and WC. This study also indicates that the intra-examiner measurement on WC (upper and lower limits, 1.3 to −1.1) was better than the inter-examiner measurement (upper and lower limits, 2.3 to −1.9). The findings of this study suggest that WC, WT and HT measurements in adults aged 18 years and above using the Seca circumference tape, Seca Bodymeter and Tanita weighing scale are reliable instruments to be used in a community survey. We would like to stress the critical importance of training, both before and during the course of data collection in surveys, to minimise potential errors and, where possible, to limit to a single/minimum number of examiner(s) to reduce inter-examiner differences. The examiner(s) should be trained and later assessed by qualified anthropometrists against some predetermined target training values. It is also recommended that during the course of the data collection, some onsite assessments be done on a random basis to maintain the quality and accuracy of the measurements.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Nirmal Singh, Director of the Institute for Public Health, for allowing his staff to participate in the study, both as respondents and as project team members. Special thanks to Dr Lim Teck Onn, Director of the Network of Clinical Research Centres, for providing inspiration and technical advice; Chen WS and Rajaah M for their assistance in data analysis; Dr Jamalludin AR, Dr Ruzita AT and Wong NF for their support during the planning of the project; and lastly, Nurul Naquiyah K, Saweah J, Zawiyah MD, Norhayati A, Norizan AR and Nur Akmar AR for their cooperation.

REFERENCES

1. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. British J Nutr 1999; 82:165-77.
2. New York Obesity Research Center. Body composition unit. Available at: www.nyorc.org/anthropometry.html. Accessed September 16, 2006.
3. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998; 17:101-10.
4. WHO Expert Committee. Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. Technical Report Series No. 854. Geneva: World Health Organization, 1995.
5. Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
6. Medical University of South Carolina (MUSC) (2006). Available at: www.musc.edu/dc/icrebm/index.html. Accessed September 20, 2006.
7. Zady MF. Z-4: mean, standard deviation, and coefficient of variation. Madison: Westgard QC. Available at: www.westgard.com/lesson34.htm#coefficient. Accessed October 7, 2006.
8. Bland M. How could I calculate a within-subject coefficient of variation? Available at: www-users.york.ac.uk/~mb55/meas/cv.htm. Accessed October 7, 2006.
9. Talita AP, Glauber LO, Juliana SO, Fatima PO. Technical error of measurement in anthropometry. Rev Bras Med Esporte 2005; 11:86-90.
10. Wang J, Thornton JC, Bari S, et al. Comparisons of waist circumferences measured at 4 sites. Am J Clin Nutr 2003; 77:379-84.
11. Ulijaszek SJ, Mascie-Taylor CGN, eds. Anthropometry: the Individual and the Population. Cambridge: Cambridge University Press, 2005.
12. Pelletier DL, Low JW, Msukwa LAH. Sources of measurement variation in child anthropometry in the Malawi maternal and child nutrition survey. Am J Hum Biol 1991; 3:227-37.
13. Himes JH. Reliability of anthropometric methods and replicate measurements. Am J Phys Anthropol 1989; 79:77-80.