Reliability, Technical Error of Measurements and Validity of Length and Weight Measurements for Children Under Two Years Old in Malaysia

H Jamaiyah*, A Geeta*, M N Safiza, G L Khor***, N F Wong**, C C Kee****, R Rahmah*****, A Z Ahmad**, S Suzana*****, W S Chen******, M Rajaah*, B Adam***

*Clinical Research Centre, Hospital Kuala Lumpur, **Institute for Public Health, Kuala Lumpur, ***Universiti Putra Malaysia, Serdang, ****Institute for Medical Research, Kuala Lumpur, *****Hospital Universiti Kebangsaan Malaysia, Kuala Lumpur, ******Biostats Consult Sdn Bhd

SUMMARY

The National Health and Morbidity Survey III 2006 wanted to perform anthropometric measurements (length and weight) for children in their survey. However there is limited literature on the reliability, technical error of measurement (TEM) and validity of these two measurements. This study assessed the above properties of length (LT) and weight (WT) measurements in 130 children age below two years, from the Hospital Universiti Kebangsaan Malaysia (HUKM) paediatric outpatient clinics, during the period of December 2005 to January 2006. Two trained nurses measured WT using Tanita® digital infant scale model 1583, Japan (0.01kg) and Seca® beam scale, Germany (0.01 kg) and LT using Seca® measuring mat, Germany (0.1cm) and Sensormedics® stadiometer model 2130 (0.1cm). Findings showed high inter and intra-examiner reliability using 'change in the mean' and 'intra-class correlation' (ICC) for WT and LT. However, LT was found to be less reliable using the 'Bland and Altman plot'. This was also true using Relative TEMs, where the TEM value of LT was slightly more than the acceptable limit. The test instruments were highly valid for WT using 'change in the mean' and 'ICC' but was less valid for LT measurement. In spite of this we concluded that, WT and LT measurements in children below two years old using the test instruments were reliable and valid for a community survey such as NHMS III within the limits of their error. We recommend that LT measurements be given special attention to improve its reliability and validity.

KEY WORDS:
Anthropometry, Children, Intra-examiner, Inter-examiner, Reliability, Validity technical error of measurement

INTRODUCTION

Anthropometric measurements, such as weight and length, are used to assess nutritional status and growth in children. All anthropometric measurements require equipment. Whatever equipment is chosen and whoever does the measurement, it will be subjected to some degree of measurement error. These include within and between examiners variability, technical and mechanical limitations. Among the many measurement methods, anthropometry generally demonstrates the largest standard errors and lowest correlation coefficients [1].

Various terms are used to describe anthropometric measurement error. These include reliability and validity[2]. Reliability is the degree to which within-subject variability is present and is due to factors other than 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. The technical error of measurement (TEM) is another accuracy index to express the error margin in anthropometry. It has been adopted by the International Society Standardization Advancement in Kinanthropometry (ISAK) for the accreditation of anthropometrists in Australia[3]. The TEM index allows anthropometrists to verify the accuracy degree when performing and repeating anthropometrical measurements (intra-examiner) and when comparing their measurement with measurements from other anthropometrists (inter-examiner) [3].

The Third National Health and Morbidity Survey, Malaysia 2006[4] which was a nationwide community survey, included a nutritional status assessment component for children. They wanted a validation of the anthropometric measurements that were to be used. This is because despite the importance of accurate and reliable anthropometric measurements, there are relatively few papers [5] addressing reliability and validity issues of these assessments. In fact, search in the Cochrane Reviews Database[6] of the term "anthropometric measurement in children", "anthropometric measurement", and "TEM" failed to yield any matching review paper. This paper attempts to assess the inter- and intra-examiner reliability of weight (WT) and length (LT) measurements and their respective technical error of measurements. We also examined the validity of measurements of WT and LT compared to the measurements using reference instruments that have been used in standard clinical practice in Malaysia.

MATERIALS AND METHODS

This cross-sectional study had a convenient sample of 130 children, age less than 2 years (including infants) who were clinically stable. They were recruited from the Hospital

Universiti Kebangsaan Malaysia (HUKM) paediatric outpatient clinics, during the period of December 2005 to January 2006. The exclusion criteria were children with obvious physical disabilities and body deformation. The sample size was determined as Walter *et al* [7], with two replicates per subject; the expected reliability coefficient (r) of at least 0.8 ($H_1 : \rho_1 = 0.8$), the reliability of 0.7 ($H_0 : \rho_0 = 0.7$) or higher to be minimally acceptable, $\alpha = 0.05$ and $\beta = 0.2$ (corresponds to 80% power); this would require a total number of 117.1 subjects. Using a 10% over–estimate to account for poor response, the final target sample size was 130.

Two trained examiners with background in public health nursing did WT and LT measurements of each subject. The choice of two examiners was deliberate and it was simply because of logistic reasons, and that a pair was the minimum required number for inter examiner reliability. Selection of only one of the examiner for intra examiner reliability was also not influenced in any way. Both examiners were not part of the research team and were therefore study blinded. On the day of the assessment, each examiner performed and recorded the measurements on their own. They were specifically told not to recall their previous readings. The data capture form was designed in such a way that the examiners were asked to fold over recordings of the previous readings immediately after it was recorded to minimise recall bias. The process of measurement is shown in Figure 1.

The LT of the subject was obtained by using two instruments 1) Seca® measuring mat, Germany (0.1cm) as "test" instrument, which was 'improvised' by attaching a non-stretchable tape on the right side of the mat to allow measurements of up to the nearest 0.01 cm on both sides of the mat and 2) Sensormedics® stadiometer model 2130 (0.1cm)[8] as "reference" instrument. The subject body weight was taken using two instruments 1) Tanita® digital infant scale model 1583, Japan (0.01kg) as "test" instrument and 2) the Seca Beam scale Germany (0.01 kg) weighing machine for WT [9] as "reference" instrument. The measurements were done using study specific procedures as described in the Technical Manual of NHMS III [4].

Statistical analysis

Statistical analyses for reliability were done using 'change in the absolute mean', intraclass correlation coefficient (ICC) and Bland and Altman plot [10]. Absolute mean is a crude way for checking for difference or agreement between two readings. We also tested the difference for significance using independent t test and paired t test for absolute means for between and within examiners respectively. Correlation coefficient (r) was used as a more objective way of assessing reliability. It was computed using ICC to demonstrate the strength of the relationship (similarities) between two measurements. The values for reliability coefficient range from 0 to 1. A coefficient of below 0 indicates "no reliability", >0 to <0.2 is slight reliability, 0.2 - <0.4 is fair reliability, 0.4 - <0.6 is moderate, 0.6 - < 0.8 is substantial and 0.8 – 1.0 is almost perfect reliability [11].

Bland and Altman was used to provide an illustration of the spread of differences in readings, the mean difference and the upper and lower limit of agreement both for inter as well as intra-examiner reliability. There is no such 'acceptable' range for Bland and Altman plots. The technical error of measurement (TEM), which is an accuracy index [3] was also calculated. It is essentially the standard deviation between repeated measures. The lower the TEM obtained, the better the reliability. The acceptable ranges for Relative TEM using beginner anthropometrist levels for intra-examiner is < 1.5% and inter-examiner < 2.0% [3]. The formula for TEM calculation [12] is;

$$\frac{\Sigma D^2}{2N}$$

The formula for percentage TEM is as below;

$$\%TEM = \frac{TEM}{\bar{X}} \times 100$$

The formula for coefficient of reliability R is;

$$R = 1 - \left\{ \frac{(TEM)^2}{SD^2} \right\}$$

Using the above formula, the coefficient of reliability R can be determined, which ranges from 0 (not reliable) to 1 (complete reliability). Inter-examiner reliability refers to how consistent/in agreements were the readings from the two examiners on the same subjects. Intra-examiner reliability refers to how consistent/in agreements were the readings from the same examiner on the same subjects but at two different time points.

In addition to these, the coefficient of variation (CV) is calculated to further determine the precision of measurements methods. The CV provides a general "feeling" about the performance of a measurement. CVs of 5% or less generally give us a feeling of good method performance, whereas CVs of 10% and higher are bad [13].

In order to compare the variability of the two methods; LT and WT, the percentage of coefficient of variation (% CV) was calculated using the data from the inter as well as the intra examiner. Percentage of Coefficient of variation is therefore a good indicator to use when comparing methods [14]. For validity, the 'accuracy' of the measurements using the test instruments was compared with their respective measurements using reference instruments. This was done on the basis of an underlying assumption that the reference readings were at least close to, if not the actual 'true' readings. Here again comparison was made using 'change in the mean', ICC and Bland and Altman as described above.

Measures of validity are similar to measures of reliability. With reliability, you compare one measurement of a variable on a group of subjects with another measurement of the same variable on the same subjects. With validity, you also compare two measurements on the same subjects. The first measurement is for the variable you are interested in, which is usually some practical variable or measure. The second

measurement is for a variable that gives values as close as you can get to the true values of whatever you are trying to measure. We call this variable the criterion variable or measure [15].

RESULTS

Sample characteristics

The mean age of 130 children in the study was 279.3 ± 186.4 days. Boys and girls were almost equally represented (57.69% and 42.31% respectively). Malay children formed the majority (70%), followed by Chinese (24.6%) and Indians (3.1%).

Reliability

Inter-examiner reliability

There were three ways in which inter-examiner reliability was examined. The first was by change in the mean. Table I shows that there was an average 0.1 kg difference for WT but no difference detected for LT.

The second method was by correlation coefficient. Results of correlation coefficient of inter examiner analysis using intra-class coefficient (ICC) are as in Table II. The ICC for LT was 0.9880 and for WT was perfect agreement 1.0000 which means strong correlation between readings from the two examiners for WT and LT respectively. These indicated high degree of reliability between the two examiners for both measurements.

The third method was using the Bland and Altman plot. For WT, Fig 2 shows that the measurement taken from examiner #2 is consistent with examiner #1 with an average difference of 0.0 kg, upper limit at 0.2 and lower limit of -0.2 kg. For LT, Fig 3 shows that on average, the measurement taken from examiner 1 is 0.1 cm higher than that of examiner #2. The upper limit of agreement is 4.1 while the lower limit is -3.9 cm. The points were scattered closely to zero which was consistent with ICC analysis of almost perfect agreement.

Intra examiner reliability

Similar analysis was performed for intra-examiner reliability. Absolute differences were very minimal, WT 0.1 kg and LT 0.3 cm (Table I). The ICC for both WT and LT was almost perfect; 0.9900 and 0.9990 respectively indicating the strong correlation between the readings at time1 and time2 from the first examiner for WT and LT respectively. The Bland and Altman plot showed that for WT, the average mean difference across all values of readings were 0.0 kg with upper limit of +0.2 kg and lower limit of -0.2 kg (Fig 4). For LT, the average was -0.1 cm and upper limit of +3.6 cm and lower limit of -3.8 cm (Fig 5).

Coefficient of variation

Variability of readings were minimal for inter-examiner, the CV for LT was 1.8% and for WT was 0.8% which indicated good precision. Similar results were observed for intra-examiner readings at 1.6% and 1.1 % respectively. (See Table II)

Validity

Table I shows comparison of measurements using test instrument versus reference instrument in 129 subjects. There was a mean absolute difference of 0.2 kg for WT but no absolute difference for LT. With an intra-class correlation coefficient close to 1 (Table II), there is a high degree of reliability for both WT and LT measurements obtained using the two instruments. Another way of looking at accuracy is by the Bland & Altman Plot. Fig 6 shows the plot for WT measurement. On average, the measurements taken from the test instrument were consistent with the reference instrument. At maximum, the difference can be up to 0.2 kg and at minimum -0.3 kg. Fig 7 shows the plot for LT measurement. On the average the test instrument is recording 0.2 cm higher than the reference instrument with upper limit of 2.1 and lower limit of -1.7.

Technical Error of Measurement

The result for the TEM and R is tabulated in Table III. The relative TEMs for inter and intra examiners for WT were 0.8% and 1.1% respectively. The relative TEMs for inter and intra examiners for LT were 2.1% and 1.9%. The estimate for inter examiner for LT is marginally acceptable but that of intra slightly exceeded the acceptable values. This study also found that all the R values (for inter and intra, WT and LT) were above the 0.95 suggested cut-off [2]. This means that the human error for measurements in the study was small; all below the acceptable 5% mark.

DISCUSSION

Anthropometric measurement is important in assessing nutritional status. However its interpretation depends greatly on the degree of reliability and validity findings. This study in particular estimated the two basic properties for the weight and length measurement in children below two years old. The commonly used indices for reliability include technical error of measurement (TEM), relative TEM (%TEM), coefficient of reliability (R) and intraclass correlation coefficient (ICC) 2. We report all of the above and in addition we also calculated the coefficient variation (CV) and Bland & Altman plot. The validity aspect was examined in a similar way as above [14].

Changes in the mean, correlation coefficient and Bland and Altman plots revealed a high degree of inter-examiner reliability. The p values of the change in the mean showed no statistical significance. It was also found, by Bland and Altman, that for WT, the two examiners were consistent with an average of 0.0 kg and an upper limit of 0.2 kg to lower limit of -0.2 kg. However it was found that LT measurements had a broader limit of agreement where upper limit of agreement was 4.1 while the lower limit was -3.9 cm and on average, examiner 1 was recording 0.1 cm higher than the examiner 2. This is understood because in children, it is more challenging to keep the child still and stretched for a good assessment of length [4]. For intra-examiner reliability, the change in the mean were minimal; WT 0.1 kg and LT 0.3 cm and both differences were not significant. The ICC for both WT and LT was almost perfect; 0.9900 and 0.9990 respectively. The Bland and Altman plot showed that for WT, the readings were consistent (average mean difference of 0.0 kg) with upper limit of +0.2 kg and lower limit of -0.2 kg. However for LT, the average was -0.1 cm suggestive of some degree of error with limits of ± about 4 cm. This wider limit range in LT measurement was explained earlier.

Table I: Summary Statistics for reliability (inter, intra examiner) and validity (inter instrument)

Inter examiner reliability				
Summary Statistics	Examiner 1 (1)	Examiner 2 (2)	Absolute Mean Diff (1)-(2)	P value
Length, cm				
N	130	129*		
Mean, (SD)	67.9 (9.5)	67.9 (9.2)	0	0.991
Median, (min, max)	67.5 (46.5, 95.0)	68.0 (45.5, 86.6)		
Weight, kg				
N	130	130		
Mean, (SD)	7.6 (2.3)	7.5 (2.3)	0.1	0.955
Median, (min, max)	7.4 (2.2, 15.7)	7.5 (2.2, 15.6)		
Intra examiner reliability in the first examiner				
Summary Statistics	Examiner 1 1st Measurement (1)	2nd Measurement (2)	Change in the mean (1)-(2)	P value
Length, cm				
N	130	129*		
Mean, (SD)	67.9 (9.5)	67.6 (9.2)	0.3	0.452
Median, (Min, max)	67.5 (46.5, 95.0)	67.6 (46.0, 87.0)		
Weight, kg				
N	130	129*		
Mean, (SD)	7.6 (2.3)	7.5 (2.2)	0.1	0.146
Median, (Min, max)	7.4 (2.2, 15.7)	7.4 (2.2, 15.1)		
Inter instrument validity				
Summary Statistics	Reference Instrument (1)	Test Intrument (2)	Change in the mean (1)-(2)	
Length, cm				
N	130	129*		
Mean, (SD)	7.5 (2.3)	7.5 (2.2)	0	
Median, (Min, max)	7.5 (2.2, 15.6)	7.5 (2.2, 15.2)		
Weight, kg				
N	129*	129*		
Mean, (SD)	67.9 (9.2)	67.7 (9.3)	0.2	
Median, (Min, max)	68.0 (45.5, 86.6)	67.7 (46.0, 88.0)		

* 1 subject refused to participate

Table II: Correlation Coefficient (ICC) and Coefficient of Variation of the Inter and Intra Examiners measurements

Variables	**N**	**Inter-examiner**		**N**	**Intra-examiner**	
		ICC	Coefficient of Variation (%)		ICC	Coefficient of Variation (%)
Length, cm	128*	0.9880	1.8%	128*	0.9990	1.6%
Weight, kg	129*	1.0000	0.8%	128*	0.9900	1.1%

* Incompleteness of data recording and patient refusals

Table III: Inter and intra-examiner relative TEM classification results for weights and length measurements

			TEM	**%TEM**	**Classification* of %TEM**	**R ****
1	WT measurement by 1st observer VS WT measurement by 2nd observer both using tanita	Inter examiner WT	0.059708	0.791446	Acceptable (< 2.0%)	0.999329
2	1st WT measurement by 1st observer VS 2rd WT measurement by 1st observer both using tanita	Intra examiner WT	0.082076	1.097031	Acceptable (< 1.5%)	0.998621
3	LT measurement by 1st observer VS LT measurement by 2nd observer both using measurement mat	Inter examiner LT	1.413007	2.083033	Marginally acceptable (< 2.0%)	0.976747
4	1st LT measurement by 1st observer VS 2nd LT measurement by 1st observer both using measurement mat	Intra examiner LT	1.308603	1.932889	Not acceptable (< 1.5%)	0.980048

* using beginner anthropometrist cut-off values for "other measures" (Norton K, Olds T, editors. Antropometrica. Argentina: Biosystem, 2000)
** R is coefficient of reliability

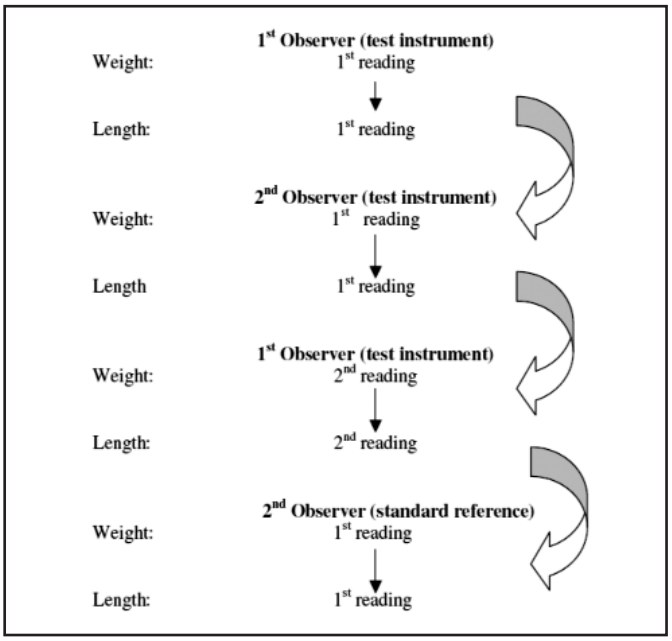


Fig. 1: Process of obtaining length and weight measurements

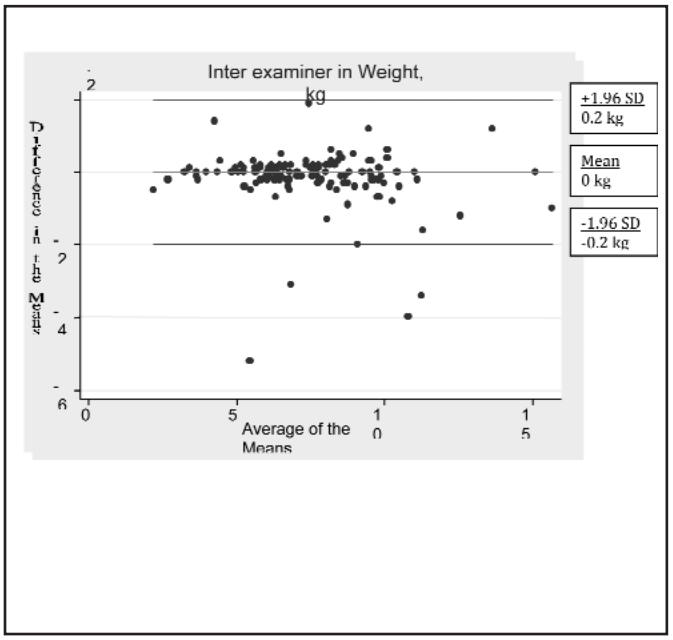


Fig. 2: Bland Altman plot on the weight measurements between examiners

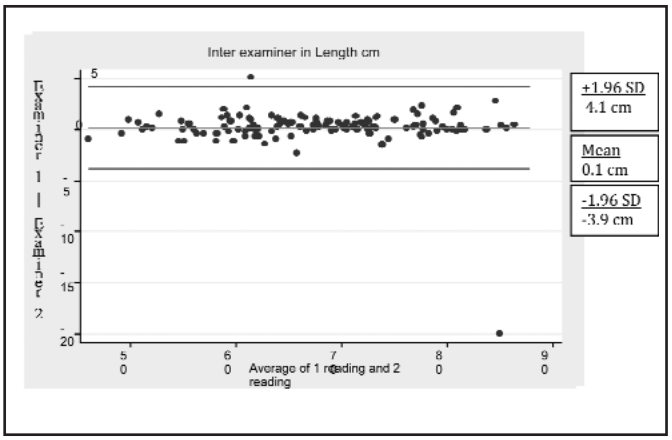
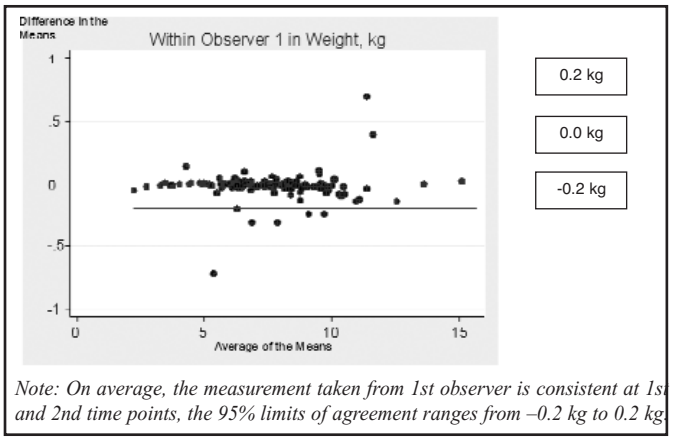
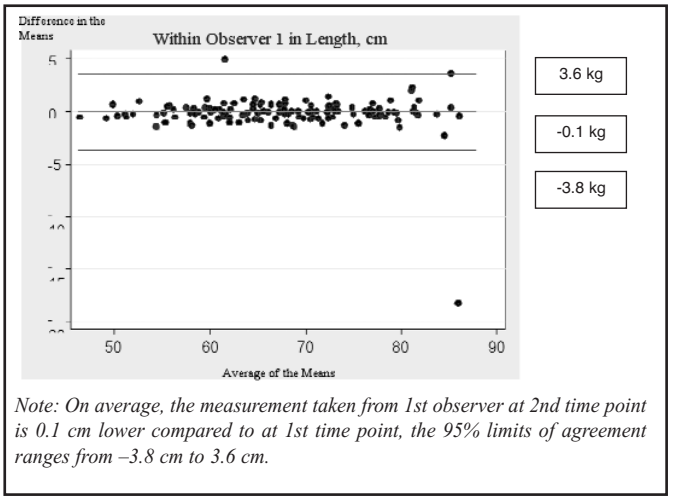


Fig. 3: Bland Altman plot on the length measurements between examiners



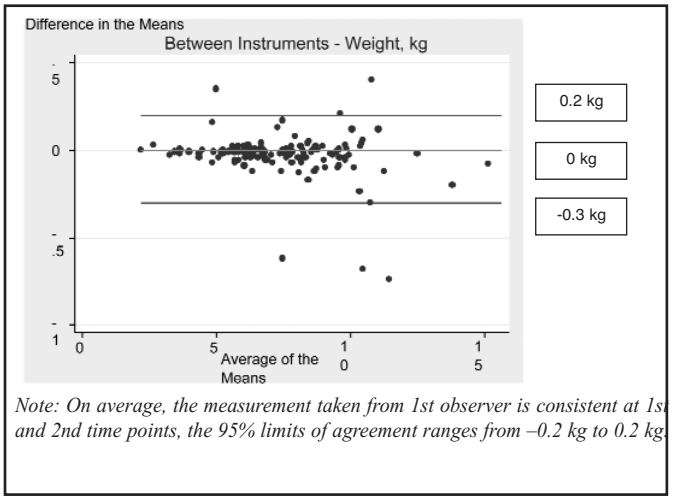
Note: On average, the measurement taken from 1st observer is consistent at 1st and 2nd time points, the 95% limits of agreement ranges from –0.2 kg to 0.2 kg.

Fig. 4: Bland Altman plot on the weight measurements within observer 1



Note: On average, the measurement taken from 1st observer at 2nd time point is 0.1 cm lower compared to at 1st time point, the 95% limits of agreement ranges from –3.8 cm to 3.6 cm.

Fig. 5: Bland Altman plot on the length measurements within observer 1



Note: On average, the measurement taken from 1st observer is consistent at 1st and 2nd time points, the 95% limits of agreement ranges from –0.2 kg to 0.2 kg.

Fig. 6: Bland Altman plot on the weight measurements between instruments

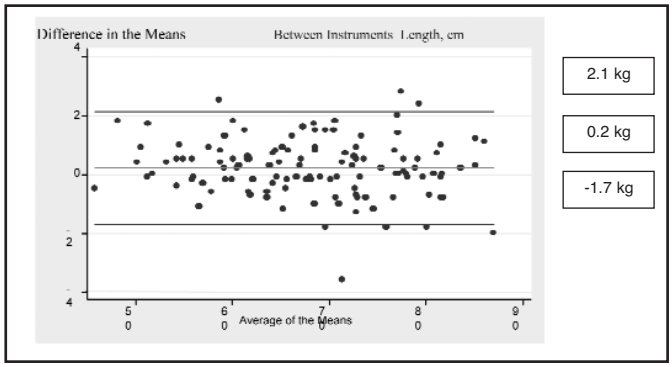


Fig. 7: Bland Altman plot on the length measurements between instruments

All three relative TEM values were within the acceptable limit except for intra examiner LT. Our findings of TEM values for LT at 1.41 cm (inter) and 1.31 cm (intra) were higher compared to the MGRS 6 countries study [15] which reported a range of TEMs for LT 0.23-/0.58 for intra and 0.23-0.35 for inter. This WHO study also found that teams tend to underestimate length and height that are most likely due to difficulties associated with keeping children fully stretched out and keeping still.

From the findings of R, the coefficient of reliability, all these four measurements reported less than 5% errors due to human measurement. These indicated close to excellent intra and inter examiner reliability for both measurements. The CVs for both LT and WT inter and intra were below 5%. This shows the variability was low in this sample which was a puzzling fact for LT. However the authors prefer to conclude that this study shows that LT measurement was less reliable than WT.

On the validity part, there was 0.2 kg difference in mean for WT and no absolute difference for LT. With an intra-class correlation coefficient of close to 1, there was high degree of accuracy between the two measurements for WT and LT. The Bland & Altman plot for WT measurement showed that on average, those taken from the test instrument were consistent with that of the reference instrument. At maximum, the difference was up to 0.2 kg and at minimum -0.3 kg. The Bland and Altman plot for LT showed that on average the test instrument is 0.2 cm higher which indicates some degree of systematic bias. The upper limit was 2.1 cm and the lower limit was -1.7 cm. Since this assessment is not for clinical diagnostic purposes, we were more willing to accept the $\pm$ 2 cm differences in LT, because measuring LT of younger children is rather challenging. WT was more accurate. All these findings could not be compared with that from other studies because there are hardly any papers on validity for similar measurements.

Study Limitations

One of the limitations of the study was that we only had one professional anthropometrist in the team. Professionals add value both during the training of examiners as well as for the quality control during the study. Secondly, we did not have pre-determined training values that the examiner needed to have attained before she is appointed as an examiner for the study. Lastly, these results were internal validity but external generalisation must be cautioned because measurements are often very much operator and instrument dependant.

CONCLUSIONS

The LT measurement was found to be slightly less reliable and less accurate, compared to that of WT measurement. However the authors concluded that both WT and LT measured in infants and children less than two years old using Tanita digital weighing machine and Seca measuring mat is relatively reliable and valid to be used for the purpose of a community survey within the magnitude of errors that was detected. Several recommendations are also given below to further enhance the examiners' measurement techniques to improve precision and accuracy, in particular for LT measurements.

Recommendations

The following recommendations should be taken into consideration to improve LT measurement that appeared to be less reliable and has a non-acceptable level of TEMs. Firstly, we should 'control' the examiners selected to perform these measurements for any surveys; both in terms of quality - ideally to only skilful anthropometrist - and quantity, to reduce inter examiner variability. However, if beginners are needed to act as examiners, then they should be trained and their techniques be assessed using relative TEMs against that of a skillful examiner until they reach acceptable limits or at regular intervals during the survey as a way for quality assurance.

ACKNOWLEDGMENT

We would like to thank the Director-General of Health for permission to conduct the study. Gratitude also goes to Director of Institute of Public Health, Dr Nirmal Singh for his permission to deploy his Institute's staff for the conduct of the study, Director of Clinical Research Centre, Dr Lim Teck Onn for giving us inspiration and advice and to the Director of Hospital Universiti Kebangsaan Malaysia for allowing us to use its premise as our study site. The authors also wish to thank Cik Nurul Naquiyah Kamaruzzaman, Ms Saweah Jono, Ms Zawiyah Md Dom, Ms Norhayati Ahmad, Ms Norizan A. Rashid, Ms Nur Akmar Abd Razak, Ms Noor Akma Hassim. We also thank all those whose names are not mentioned here who rendered their excellent support especially during data collection and data entry.

REFERENCES

1. The New York Obesity Research Centre page. Body Composition Unit. Available at: www.nyorc.org/bcu/labs/anthropometry.html. Accessed September 16, 2006.
2. Ulijaszek S, Kerr D. Anthropometric measurement error and the assessment of nutritional status. British J. of Nutr. 1999; 82: 165-77.
3. Perini TA, de Oliveira GL, Ornellas JS, de Oliveira FP. Technical error of measurement in anthropometry. Rev Bras Med Esporte 2005; 11: 86-90.
4. National Health and Morbidity Survey III (NHMS III) (2006).Technical Manual of Measurements. Institute of Public Health, Malaysia.
5. Johnson W, Cameron N, Dickson *et al*. The reliability of routine anthropometric data collected by health workers: A cross-sectional study. Int J Nurs Stud 2009; 46: 310–16.
6. The Cochrane Collaboration. Available at http://www.cochrane.org/reviews/. Accessed on January 24, 2009.

7. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Statistics in Medicine. 1998; 17: 101-10.
8. Bryne WM, Lenz ER. Reliability of Transportable Instruments for Assessment of Infant Length. J. of Nursing Measurement. 2002; 10: 111-21.
9. Indian Health Service page. American Indian and Alaska Native, Pediatric Height and Weight Study Website (IHS). Available at: http://www.ihs.gov/MedicalPrograms/Anthropometrics/index.cfm?module=train&option=guide&newquery=1. (Accessed October 15, 2005).
10. Qualitative classification of intra-class correlation (ICC) values as degree of agreement beyond chance. Available at http://www.musc.edu/dc/icrebm/index.html. (Accessed September 20, 2006).
11. Bland MJ, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet. 1986; 307-10.
12. Goto R, Mascie-Taylor CGN. Precision of measurement as a component of human variation. J Physiol Anthropol 2007; 26: 253-56.
13. Zady MF. Z-stats 4: mean, standard deviation and coefficient of variation. Wesward QC 1999. Available at http://www.westgard.com/lesson34.htm#coefficient. (Accessed October 7, 2006).
14. Martin B. How could I calculate a within-subject coefficient of variation? Available at http://www-users.york.ac.uk/~mb55/meas/cv.htm. (Accessed October 7, 2006)
15. Hopkins WG. Change in the Mean. Available at http://www.sportsci.org/resource/stats/precision.html. (Accessed January 7, 2009).
16. WHO Multicentre Growth Reference Study Group 2006. Reliability of anthropometric measurements in the WHO Multicentre Growth Reference Study. Acta Paediatrica Suppl 2006; 450: 38-46.