*Ultrasound Obstet Gynecol* 2011; **38**: 681–687
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.8997

# Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements

I. SARRIS*, C. IOANNOU*, M. DIGHE†, A. MITIDIERI‡, M. OBERTO§, W. QINGQING¶, J. SHAH**, S. SOHONI††, W. AL ZIDJALI‡‡, L. HOCH*, D. G. ALTMAN§§ and A. T. PAPAGEORGHIOU*; for the International Fetal and Newborn Growth Consortium for the 21$^{st}$ Century (INTERGROWTH-21st)

**Oxford Maternal & Perinatal Health Institute (OMPHI), Green Templeton College and Nuffield Department of Obstetrics & Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, UK; †Department of Radiology, Ultrasound Section, University of Washington Medical Center, Seattle, WA, USA; ‡Centro de Pesquisas Epidemiologicas, Universidade Federal de Pelotas, Pelotas, Brasil; §Università degli Studi di Torino, Dipartimento di Ostetricia e Neonatologia, Azienda Ospedaliera O.I.R.M. S Anna, Torino, Italy; ¶Beijing Obstetrics & Gynaecology Hospital, Maternal & Child Health Centre, Capital Medical University, Beijing, China; **Aga Khan University Hospital, Nairobi Department of Obstetrics and Gynaecology, East Tower Building, Nairobi, Kenya; ††Ketkar Nursing Home, Sitabuldi Nagpur, India; ‡‡Wattayah Polyclinic, Sultanate of Oman; §§Centre for Statistics in Medicine, University of Oxford, Wolfson College Annexe, Oxford, UK*

**KEYWORDS:** abdominal diameter; biparietal diameter; femur length; growth; pregnancy; training

## ABSTRACT

***Objective*** *To assess whether a standardization exercise prior to commencing a fetal growth study involving multiple sonographers can reduce interobserver variation.*

***Methods*** *In preparation for an international study assessing fetal growth, nine experienced sonographers from eight countries participated in a standardization exercise consisting of theoretical and practical sessions. Each performed a set of seven standard fetal measurements on pregnant volunteers at 20–37 weeks' gestation, and these were repeated by the lead sonographer; all measurements were taken in a blinded fashion. After this the sonographers had hands-on practice and feedback sessions on other volunteers. This process was repeated three times. Measurement differences between sonographers and the lead sonographer, expressed as a gestational-age-specific Z-score, between the first and third scans were compared using the Wilcoxon signed ranks test, and variance was assessed using Pitman's test. Interobserver agreement was also assessed using the intraclass correlation coefficient (ICC), and all images were scored for quality in a blinded fashion.*

***Results*** *At baseline the level of agreement and image scoring were high. A significant reduction in the differences between sonographers and the lead sonographer were seen for fetal biometry overall (head circumference, abdominal circumference and femur length) between the first and third scans (median Z-scores, 0.46 and 0.24;* $P = 0.005$*), and a reduction in the variance was also observed (*$P < 0.001$*). The ICCs for measurement pairs for every fetal measurement showed a clear trend of increasing ICC (better agreement) with consecutive training scan sessions, although no improvement in image scores was seen.*

***Conclusion*** *Even for experienced sonographers, a standardization exercise before starting a study of fetal biometry can improve consistency of measurements. This could be of relevance for studies assessing fetal growth in multicenter sites. Copyright © 2011 ISUOG. Published by John Wiley & Sons, Ltd.*

## INTRODUCTION

When evaluating fetal biometry using ultrasound there is a need to take the measurements in a methodologically consistent manner, both in research studies and in clinical practice. The aim should be to improve the uniformity and quality of the data; decrease bias and diagnostic errors; and minimize systematic user-induced errors[1]. In ultrasound studies, standardized anatomical landmarks are identified, calipers are placed at predefined points and

*Correspondence to:* Dr A. T. Papageorghiou, Oxford Maternal & Perinatal Health Institute (OMPHI), Green Templeton College and Nuffield Department of Obstetrics & Gynaecology, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DU, UK
(e-mail: aris.papageorghiou@obs-gyn.ox.ac.uk)

*Accepted: 10 March 2011*

fetal biometric measurements are taken and, usually, plotted on graphs against expected values for gestational age.

Different strategies have been used to ensure consistency of measurements. One strategy is to employ only one sonographer[2], but this inevitably limits the number of scans possible, risks the possibility of systematic bias and creates a rather artificial scenario that does not reflect normal clinical practice and cannot accommodate the needs of multicenter collaborations. Other studies utilize a number of trained, experienced sonographers[3,4]. While this reflects clinical practice more accurately, interobserver variation may compromise the quality of the data. Some studies use standardization exercises as a means of ameliorating this problem, but may not specify what the exercise involved or how the outcome was assessed[5]. In addition, given that the reliability of measurements depends on the accuracy of the ultrasound images, training assessment and certification programs have been established[6]. To maintain standards, objective scoring tools to assess the quality of images have been used in nuchal translucency measurements[7,8] and have more recently been proposed for fetal biometry[9].

The aim of this study was to assess whether a standardization exercise for a group of already experienced and accredited sonographers prior to starting a research program involving multiple sonographers improves the overall quality of their scanning and decreases interobserver variation.

## METHODS

The International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st) is a large-scale, population-based, multicenter observational project of fetal and newborn growth currently underway in eight hospitals across the world (www.intergrowth21.org.uk). It involves serial fetal growth scans every 5 ± 1 weeks from 14 weeks' gestation, but not beyond 42 weeks. All ultrasound scans are performed using the same commercially available ultrasound machine (Philips HD-9, Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal transducers (C5-2, C6-3 and V7-3). For the purposes of the INTERGROWTH-21st study, the manufacturer has reprogrammed the machine's software to ensure that the measurement values do not appear on screen during the scan.

A standardization training exercise was held in May 2009 at the INTERGROWTH-21st Coordinating Unit (based at the University of Oxford), prior to initiating recruitment into the main study. Nine sonographers from the eight units were invited to take part (henceforth referred to as 'delegates'). All are experienced sonographers, certified in their institutions as competent to perform ultrasound fetal biometry. The purpose of this exercise was to ensure that each delegate became familiar with the study equipment and measurement protocol so that they could perform INTERGROWTH-21st scans themselves in their home institutions and instruct other local team members. The INTERGROWTH-21st protocol was approved by the Oxfordshire Research Ethics Committee C; all the pregnant women involved in this part of the study were volunteers who gave informed consent.

The training consisted of theoretical and practical sessions led by a training team (A.P., I.S., C.I. and a Philips product application specialist) and lasted 3 days. The ultrasound protocol, containing step-by-step instructions on how to use the machine and take measurements, including how to obtain the correct imaging planes and place the calipers, was distributed prior to the course.

The first day was dedicated to lectures explaining the ultrasound protocol, the image scoring and quality-control processes, and an overview of the HD-9 system. The following 2 days were dedicated to hands-on, practical scanning sessions with healthy pregnant women (gestational age range 20–37 weeks based on a first-trimester dating scan) and feedback sessions. During the standardization exercise each delegate performed three consecutive scans, each on a different volunteer; the first two scans were practice scans to become familiar with the machine controls and display, the third was a formal standardization scan. During the 2 days this 'circuit' was repeated three times by all sonographers; in other words each sonographer performed nine scans, of which six were practice scans and three were standardization scans. Different volunteers were recruited for each circuit.

For each of the three standardization scans (henceforth 1st, 2nd and 3rd scan) each delegate performed one complete set of measurements of seven biometric variables: biparietal diameter (BPD), occipitofrontal diameter (OFD), head circumference using the ellipse facility (HC), anteroposterior abdominal diameter (APAD), transverse abdominal diameter (TAD), abdominal circumference using the ellipse facility (AC) and femur length (FL). Detailed definitions of these measurements are available at www.intergrowth21.org.uk (follow the link to 'Study Protocol' and download the *Ultrasound Manual*). Briefly, head measurements were taken in the transthalamic plane and measured 'outer to outer', i.e. with the intersection of the calipers placed on the outer border of the parietal (BPD), occipital and frontal (OFD) bones or on the outer border of the skull (HC using the ellipse facility). Abdominal measurements were taken with the umbilical vein in the anterior third of a transverse section of the fetal abdomen (at the level of the portal sinus) with the stomach bubble visible and with the intersection of the calipers placed on the outer borders of the body outline (skin) for APAD and TAD (taken at 90° to the APAD, across the abdomen at the widest point) or, for AC using the ellipse facility, by placing the line of the ellipse on the outer border of the abdomen. For FL, the femur closest to the probe was measured with its long axis as horizontal as possible. Calipers were placed on the outer borders of the diaphysis of the femoral bone ('outer to outer') and excluding the trochanter. For all measurements the area of interest should fill at least 30% of the monitor. For each biometric variable the blinded recorded measurements were saved directly onto the machine's hard drive along with the corresponding still images.

The measurements were repeated within a few minutes by one of the authors (A.P.), a fetal medicine specialist with extensive experience in ultrasound scanning (henceforth the 'trainer'). He was blinded to all measurements taken by the delegates and also to his own. The trainer did not interfere or correct any of the delegates' measurements. Following each scan, delegates were given feedback on how to improve their image acquisition and measurement techniques. Since it was not practical for all nine delegates to scan the same pregnant woman during each circuit, every volunteer was scanned by only three delegates at a time. Hence, the 27 resultant standardization scans were performed on a total of nine women (three for each circuit).

The measurements and stored images were retrieved after the standardization exercise. In addition to the ellipse measurement, HC was calculated from the head diameter measurements using the formula: $HC = 0.5 \times \pi \times (BPD + OFD)$ and AC was calculated from the abdominal diameter measurements using the formula: $AC = 0.5 \times \pi \times (TAD + APAD)$.

A set of stored images consisted of two head images (one for BPD/OFD and a second for HC using the ellipse), two abdominal images (one for APAD/TAD and a second for AC using the ellipse) and one image of the femur (for FL). All stored images were retrieved at random by one of the authors (C.I.) and scored by another author (I.S.), who was blinded to the identity of the sonographer and the order number (1st, 2nd or 3rd). An image scoring algorithm was used (Table 1) from a method reported by Salomon *et al.*[9]. Briefly, a transverse head image at the BPD plane scores a maximum of 6; a transverse abdominal image at the AC plane a maximum of 6; and an FL image a maximum of 4. To assess intraobserver reproducibility, 30 images were randomly re-retrieved (by C.I.) and blindly re-scored by the same reviewer (I.S.) after 24 h to avoid recall bias. The absolute score difference on test and retest was classified in terms of agreement as follows: 0–1, good; 2, moderate; > 2, poor.

### Statistical analysis

We tested the hypothesis that absolute differences in measurement between trainer and delegate for individual biometric variables may decrease with consecutive scans as a result of training and feedback. In addition we determined whether the variance between delegates of the differences in measurement between trainer and delegate also decreases. For each of the 27 standardization scans there was a set of biometric variables obtained by both delegate and trainer. Measurement difference was expressed using a *Z*-score, defined as the absolute difference of measurements by delegate and trainer divided by the SD of the normal distribution of that specific biometric variable for that specific gestational age[10–12]. *Z*-scores were preferred over absolute differences as women were scanned across a range of gestational ages. Furthermore, expressing measurements as *Z*-scores allows different fetal biometric variables within the same scan to be combined so that the overall consistency of measurements for each scan can be compared. Data were analyzed with SPSS Statistics 18.0 (SPSS Inc., Chicago, IL, USA). Distributions of *Z*-scores were plotted by order of scan: *Z*-scores of the 1st scan were compared with those of the 3rd scan in order to test the absolute (unsigned) measurement differences using the Wilcoxon test. In order to test the variance of the signed measurement differences we used Pitman's test, which allows for pairing in the data[13]. Interobserver variability was also assessed for every delegate–trainer pair for each of the seven variables using intraclass correlation coefficients (ICCs). Image scores between the 1st and 3rd scanning sessions were compared by means of the Wilcoxon signed ranks test, and image-score intraobserver reproducibility was also assessed using the Wilcoxon signed ranks test to compare score distributions on the test and retest exercises.

## RESULTS

Each of the nine delegates carried out three scans, giving 27 scans for analysis. There was a statistically significant reduction in the overall *Z*-score of differences between delegates and trainer in fetal biometry measurements (HC, AC and FL) between the 1st and 3rd scans. This reduction was seen both when measuring the HC and AC using the ellipse facility (median *Z*-score for the 1st and 3rd scans, 0.46 and 0.24, respectively; $P = 0.005$, Figure 1) and when these were calculated from diameter measurements (median *Z*-score for the 1st and 3rd scans, 0.50 and 0.23, respectively; $P = 0.035$). There was also a statistically significant reduction in the overall variance of the *Z*-score of the signed differences between delegates

**Table 1** Image scoring criteria used for the standardization exercise, based on Salomon *et al.* 2006[9]

| *Cephalic plane (maximum 6 points)* | *Abdominal plane (maximum 6 points)* | *Femoral plane (maximum 4 points)* |
|---|---|---|
| Symmetrical plane | Symmetrical plane | Both ends of the bone clearly visible |
| Thalami visible | Stomach bubble visible | Angle < 45° |
| Cavum septi pellucidi visible | Umbilical vein one-third of the way along the abdominal plane (portal sinus) | Femur occupying at least 30% of image |
| Cerebellum not visible | Kidneys not visible | Calipers placed correctly |
| Head occupying at least 30% of image | Abdomen occupying at least 30% of image | |
| Calipers/ellipse placed correctly | Calipers/ellipse placed correctly | |

**Table 2** Intraclass correlation coefficients (ICC) for measurement pairs (delegate–trainer) for each biometric variable across the three standardization scans with the measurement taken by the trainer used as a validation standard

| | *ICC (95% CI)* | | |
|---|---|---|---|
| *Parameter* | *Scan 1* | *Scan 2* | *Scan 3* |
| Biparietal diameter | 0.987 (0.944 to 0.997) | 0.989 (0.953 to 0.998) | 0.996 (0.981 to 0.999) |
| Occipitofrontal diameter | 0.566 (−0.17 to 0.878) | 0.931 (0.726 to 0.984) | 0.998 (0.987 to 1.000) |
| Abdominal circumference | 0.884 (0.565 to 0.973) | 0.956 (0.821 to 0.990) | 0.994 (0.959 to 0.999) |
| Transverse abdominal diameter | 0.806 (0.374 to 0.952) | 0.756 (0.278 to 0.938) | 0.960 (0.843 to 0.991) |
| Anteroposterior abdominal diameter | 0.845 (0.478 to 0.962) | 0.721 (0.193 to 0.928) | 0.969 (0.876 to 0.993) |
| Femur length | 0.972 (0.319 to 0.995) | 0.974 (0.892 to 0.994) | 0.994 (0.976 to 0.999) |
| Head circumference | 0.980 (0.914 to 0.995) | 0.982 (0.928 to 0.996) | 0.998 (0.993 to 1.000) |

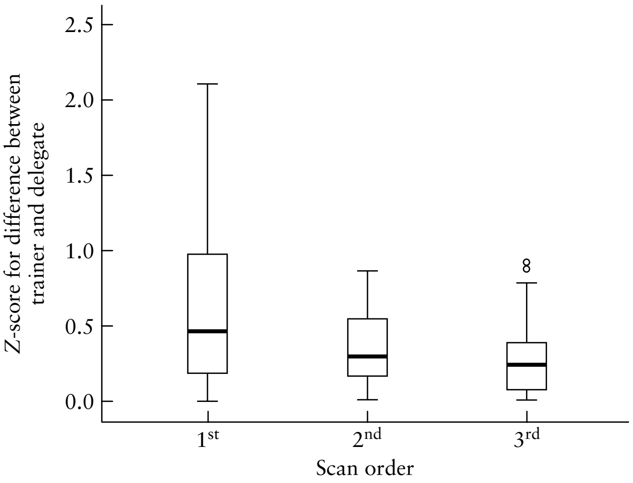

**Figure 1** Differences in measurement of head circumference (HC) and abdominal circumference (AC) between trainer and delegate, expressed as a $Z$-score, by order of scan (overall $Z$-scores measuring HC and AC with the ellipse facility). Median (black bars), interquartile range (IQR, boxes), values within 1.5 IQR (whiskers) and values exceeding 1.5 IQR (circles) are shown. Difference between first and third scans, $P = 0.005$.

and trainer in fetal biometry (HC, AC and FL) between the 1st and 3rd scans ($P < 0.001$).

For each individual biometric variable there was a clear trend of falling delegate–trainer differences with successive scanning sessions (Figure 2), but statistical significance was reached only for the HC. Table 2 summarizes the ICCs for the delegate–trainer measurement pairs and their 95% CIs for all the biometric variables across the three scanning sessions. There was a clear trend of rising ICCs with successive scanning sessions, suggesting that the accuracy of the delegates' measurements improved compared with those of the trainer.

The delegates' median image scores showed no trend across the three sessions (Table 3). On test and re-test of a sample of 30 images, there were no significant differences in score distributions for any biometric variable (Wilcoxon $P$ between 0.16 and 1.00). There was good test–retest agreement for 29 out of 30 images (97%) and moderate agreement for one image. These results suggest that image scoring by a single reviewer was reproducible.

**Table 3** Image scores during the standardization exercise

| | *Image score (median (range))* | | |
|---|---|---|---|
| *Parameter* | *Scan 1* | *Scan 2* | *Scan 3* |
| Head circumference | | | |
| From BPD and OFD | 5 (4–6) | 5 (5–6) | 6 (4–6) |
| By ellipse method | 6 (4–6) | 5.5 (5–6) | 5 (4–6) |
| Abdominal circumference | | | |
| From APAD and TAD | 5 (4–6) | 5.5 (4–6) | 5 (2–6) |
| By ellipse method | 5.5 (4–6) | 5 (4–6) | 5 (3–6) |
| Femur length | 4 (4–4) | 4 (3–4) | 4 (3–4) |

Difference between 1st and 3rd scans, $P = 0.785$. APAD, anteroposterior abdominal diameter; BPD, biparietal diameter; OFD, occipitofrontal diameter; TAD, transverse abdominal diameter.

## DISCUSSION

This study has shown that a standardization exercise for a group of experienced and accredited sonographers before starting a multicenter study led to significant improvement in the consistency of measurements, with improvements in both the median differences and their variance. Although this might appear to be inherently not surprising, few studies employ such exercises or describe them in any detail. To our knowledge this is the first study to quantify the effect that a standardization exercise has on actual measurement reproducibility among well-trained sonographers.

The institutions participating in INTERGROWTH-21st are diverse and employ different protocols for scanning women in their routine clinical practice. This is common in multicenter studies and could lead to systematic errors[14]. For data to be comparable across observers and sites all ultrasound measurements must be standardized in a consistent manner to allow data across sites to be pooled. Whatever the chosen methodology used for measurement, an important aspect of data collection is ensuring that measurements are made consistently[10–12]. Although sonographers taking part in the INTERGROWTH-21st study were trained to each country's national standards and perform a large number of scans each year, we hypothesized that a standardization exercise could lead to greater uniformity in measurement.

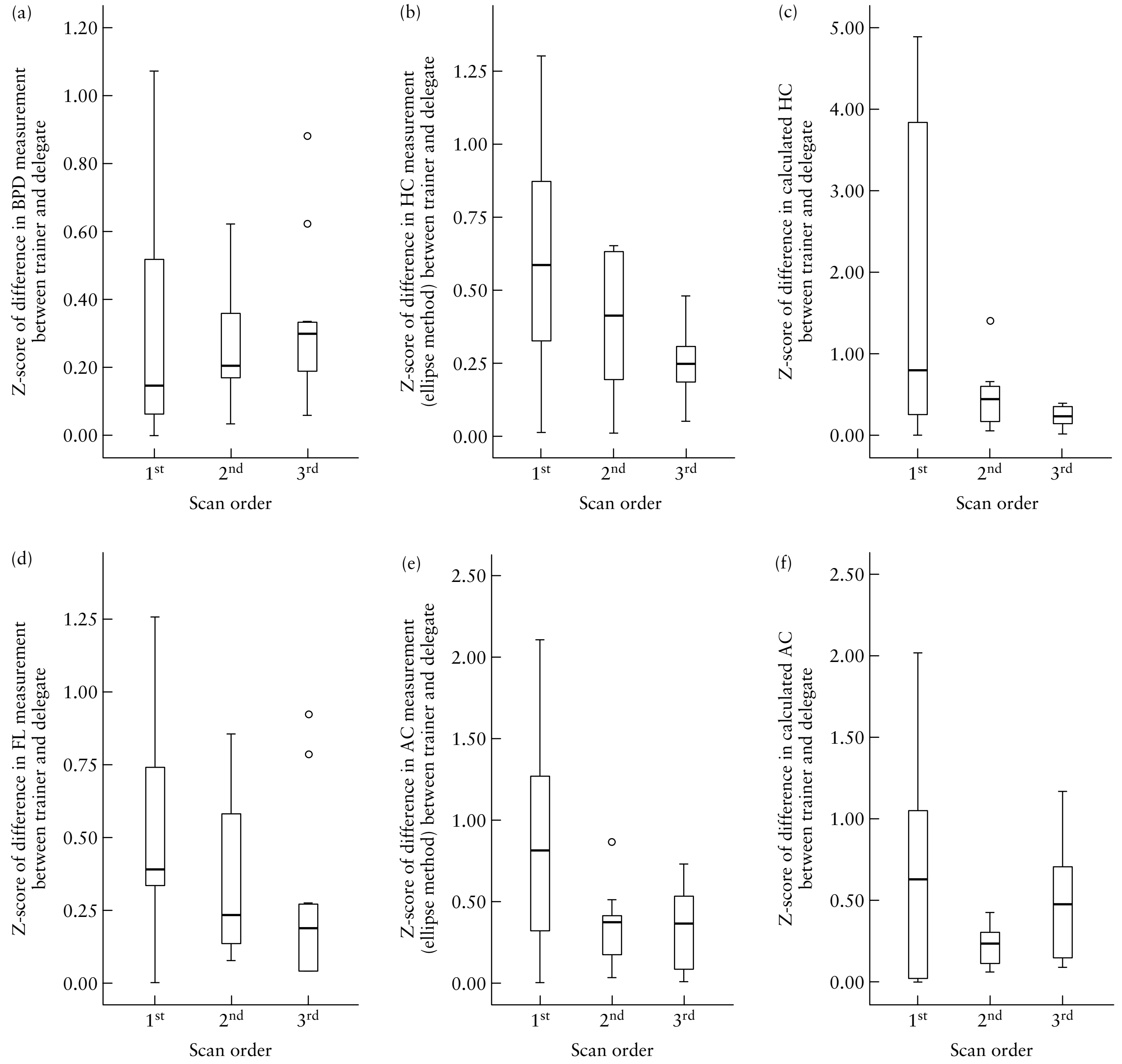

**Figure 2** Differences in measurements between trainer and delegate of: (a) biparietal diameter (BPD) ($P = 0.859$); (b) head circumference (HC) by ellipse method ($P = 0.066$); (c) HC calculated from measurement of head diameter ($P = 0.038$); (d) femur length (FL) ($P = 0.086$); (e) abdominal circumference (AC) by ellipse method ($P = 0.139$); (f) AC calculated from measurement of abdominal diameter ($P = 0.859$) expressed as $Z$-scores, by order of scan. Median (black bars), interquartile range (IQR, boxes), values within 1.5 IQR (whiskers) and values exceeding 1.5 IQR (circles) are shown. $P$-values are for difference between the 1st and 3rd scans.

This study confirms this, and evaluates the performance of the exercise.

The training period aimed to familiarize sonographers with the study equipment and how to measure fetuses in a standardized manner using the study protocol. To evaluate any improvement over time, each delegate was tested against the trainer three times during training. Measurements were compared and the corresponding images scored independently and blindly.

Even though the accuracy of the measurements improved over the three scanning sessions, the image scores did not, although it is possible that a difference in scoring performance could have been demonstrated if the number of observations had been larger. There are a number of possible explanations. One explanation has to do with the level of experience at the beginning of the exercise: for example, a study assessing the abilities of trainee doctors in performing scans in emergency gynecology showed improvement after training, as assessed by a different scoring method[15]. Our study was different in that all the delegates were already experienced and it may be for this reason that no improvement in image scoring was seen. Another possible explanation is that image scoring may not be sensitive enough to assess the finer details that cause small measurement differences; although scoring ultrasound images has been shown to be a reproducible way of evaluating quality[9], few studies have used it as a means of assessing measurement performance prospectively in training[15]. However, it is currently the only validated tool described in the literature for objective assessment of image scoring.

A standardization program prior to starting a study involving multiple sonographers was recently reported and involved training and testing inexperienced sonographers in order to standardize their ability to perform scans adequately[16]. Our study was different in that participating sonographers were already competent, and a training exercise could have been deemed unnecessary – in keeping with other studies[3,4]. However, we have demonstrated that standardization is still beneficial for experienced sonographers. Although at baseline the level of agreement and image scoring were high, there was a significant reduction in the differences between the sonographers and the trainer, and a significant reduction in the variance. It is important to note that studies measuring fetal biometry often report small changes in size as meaningful. Our findings could have important implications for such studies, since more precise measurements could increase the ability to detect significant differences. Knowing the magnitude of inconsistencies within and between sonographers and how much this contributes to observed measurement differences (i.e. intra- and interobserver variation and possible bias) should be considered when interpreting data[16–18].

Our study has limitations. Although the trainer scanned all women included in the standardization, so that the delegates could be evaluated against him, not all delegates scanned all women. It is possible, though unlikely, that maternal characteristics could be associated with differences in reproducibility of measurements[3]. Ideally, all the delegates should have scanned all the women three times, but it was felt unreasonable to ask them to be scanned so many times. A further limitation is that the trainer was utilized as the gold standard and representative of the 'true' biometric measurement. There is no way of verifying that these measurements are indeed a 'truer' representation. It is also possible that the trainer was improving over the course of the exercise, and was not a fixed reference point of competency and quality. However, the trainer was thoroughly familiar with the INTERGROWTH-21st protocol, equipment, image quality and scoring algorithms, and has considerable experience in obstetric ultrasound at a tertiary level; it was considered that these measurements would be as close to the 'true' value as achievable. An alternative approach would have been to use the average of multiple measurements taken for each woman by several delegates. We felt that practically this was not the best way as variation in the mean measurement value compared to the 'true' one would vary as delegates became more proficient throughout the exercise, and that it was preferable to have a potential fixed systematic bias introduced by the trainer acting as the gold standard.

From the outset, all delegates were deemed competent to perform fetal biometric measurements and they had high levels of agreement at the baseline comparison. Despite this, even with our relatively short training exercise, we demonstrated narrowing differences with consecutive scans. The possible reasons for this could be familiarization with the equipment used, the precise study protocol, or focusing on image quality via the scoring system. To ensure high-quality measurements throughout the INTERGROWTH-21st study, ongoing quality control measures are in place in addition to this standardization exercise, and these will be reported at a later stage.

## CONCLUSION

Even for experienced sonographers a standardization exercise before starting a study of fetal biometry using multiple sonographers can improve the consistency of the measurements. This could be of great relevance to studies using fetal growth as a primary outcome.

## ACKNOWLEDGMENTS

We are grateful to the pregnant volunteers who patiently underwent the ultrasound scans for this exercise. Aris Papageorghiou and Christos Ioannou are supported by the Oxford Partnership Comprehensive Biomedical Research Centre with funding from the Department of Health NIHR Biomedical Research Centres funding scheme. This project was supported by the INTERGROWTH-21st Grant ID# 49038 from the Bill & Melinda Gates Foundation to the University of Oxford, for which we are very grateful.

We would like to thank Philips Healthcare for providing the specially modified HD-9 ultrasound machines and technical assistance.

## REFERENCES

1. Dudley NJ, Chapman E. The importance of quality management in fetal measurement. *Ultrasound Obstet Gynecol* 2002; **19**: 190–196.
2. Martins WP, Ferriani RA, Nastri CO, Filho FM. First trimester fetal volume and crown–rump length: comparison between singletons and twins conceived by in vitro fertilization. *Ultrasound Med Biol* 2008; **34**: 1360–1364.
3. Harstad TW, Buschang PH, Little BB, Santos-Ramos R, Twickler D, Brown CE. Ultrasound anthropometric reliability. *J Clin Ultrasound* 1994; **22**: 531–534.
4. Gomez O, Martinez JM, Figueras F, Del Rio M, Borobio V, Puerto B, Coll O, Cararach V, Vanrell JA. Uterine artery Doppler at 11–14 weeks of gestation to screen for hypertensive disorders and associated complications in an unselected population. *Ultrasound Obstet Gynecol* 2005; **26**: 490–494.
5. Choong S, Rombauts L, Ugoni A, Meagher S. Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions. *Ultrasound Obstet Gynecol* 2003; **22**: 571–577.
6. American Institute of Ultrasound in Medicine. AIUM Practice Guideline for the performance of an antepartum obstetric ultrasound examination. *J Ultrasound Med* 2003; **22**: 1116–1125.

7. Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. Nuchal translucency audit: a novel image-scoring method. *Ultrasound Obstet Gynecol* 1998; **12**: 398–403.
8. Snijders RJ, Thom EA, Zachary JM, Platt LD, Greene N, Jackson LG, Sabbagha RE, Filkins K, Silver RK, Hogge WA, Ginsberg NA, Beverly S, Morgan P, Blum K, Chilis P, Hill LM, Hecker J, Wapner RJ. First-trimester trisomy screening: nuchal translucency measurement training and quality assurance to correct and unify technique. *Ultrasound Obstet Gynecol* 2002; **19**: 353–359.
9. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol* 2006; **27**: 34–40.
10. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 2. Head measurements. *Br J Obstet Gynaecol* 1994; **101**: 35–43.
11. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 4. Femur length. *Br J Obstet Gynaecol* 1994; **101**: 132–135.
12. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 3. Abdominal measurements. *Br J Obstet Gynaecol* 1994; **101**: 125–131.
13. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research* (4th edn). Blackwell Science: Oxford, 2002; 203–204.
14. Deter RL, Harrist RB, Hadlock FP, Carpenter RJ. Fetal head and abdominal circumferences: I. Evaluation of measurement errors. *J Clin Ultrasound* 1982; **10**: 357–363.
15. Salomon LJ, Nassar M, Bernard JP, Ville Y, Fauconnier A; Société Française pour l'Amélioration des Pratiques Echographiques (SFAPE). A score-based method to improve the quality of emergency gynaecological ultrasound examination. *Eur J Obstet Gynecol Reprod Biol* 2009; **143**: 116–120.
16. Neufeld LM, Wagatsuma Y, Hussain R, Begum M, Frongillo EA. Measurement error for ultrasound fetal biometry performed by paramedics in rural Bangladesh. *Ultrasound Obstet Gynecol* 2009; **34**: 387–394.
17. Mongelli M, Ek S, Tambyrajia R. Screening for fetal growth restriction: a mathematical model of the effect of time interval and ultrasound error. *Obstet Gynecol* 1998; **92**: 908–912.
18. Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Predanic M, Streltzoff J, Skupski DW. Intraobserver and interobserver reproducibility of fetal biometry. *Ultrasound Obstet Gynecol* 2004; **24**: 654–658.