Anthropometric Measurement Standardization in the US-Affiliated Pacific: Report from the Children's Healthy Living Program

FENFANG LI,1,6 LYNNE R. WILKENS,2 RACHEL NOVOTNY,1 MARIE K. FIALKOWSKI,1 YVETTE C. PAULINO,4 RANDALL NELSON,5 ANDREA BERSAMIN,8 URSULA MARTIN,7 JONATHAN DEENIK,3 AND CAROL J. BOUSHEY2

1 Department of Human Nutrition, Food and Animal Sciences, University of Hawai‘i at Manoa, Honolulu, Hawai‘i
2 Cancer Center, University of Hawai‘i at Manoa, Honolulu, Hawai‘i
3 Department of Tropical Plants and Soil Science, University of Hawai‘i at Manoa, Honolulu, Hawai‘i
4 School of Nursing and Health Sciences, University of Guam, Mangilao, Guam
5 Land Grant Program, Northern Marianas College, Saipan, Northern Mariana Islands
6 Department of Alaska Native Health, University of Alaska, Fairbanks, Alaska
7 Land Grant Program, American Samoa Community College,Pago Pago, American Samoa

Objectives: Anthropometric standardization is essential to obtain reliable and comparable data from different geographical regions. The purpose of this study is to describe anthropometric standardization procedures and findings from the Children's Healthy Living (CHL) Program, a study on childhood obesity in 11 jurisdictions in the US-Affiliated Pacific Region, including Alaska and Hawai‘i.

Methods: Zerfas criteria were used to compare the measurement components (height, waist, and weight) between each trainee and a single expert anthropometrist. In addition, intra- and inter-rater technical error of measurement (TEM), coefficient of reliability, and average bias relative to the expert were computed.

Results: From September 2012 to December 2014, 79 trainees participated in at least 1 of 29 standardization sessions. A total of 49 trainees passed either standard or alternate Zerfas criteria and were qualified to assess all three measurements in the field. Standard Zerfas criteria were difficult to achieve: only 2 of 79 trainees passed at their first training session. Intra-rater TEM estimates for the 49 trainees compared well with the expert anthropometrist. Average biases were within acceptable limits of deviation from the expert. Coefficient of reliability was above 99% for all three anthropometric components.

Conclusions: Standardization based on comparison with a single expert ensured the comparability of measurements from the 49 trainees who passed the criteria. The anthropometric standardization process and protocols followed by CHL resulted in 49 standardized field anthropometrists and have helped build capacity in the health workforce in the Pacific Region. Am. J. Hum. Biol. 00:000–000, 2015.

© 2015 Wiley Periodicals, Inc.
particularly true in the US-Affiliated Pacific Region (USAP), where professional anthropometrists are rare and childhood obesity rates are high (Bruss et al., 2010; de Onis et al., 2010; Novotny et al., 2013a; Novotny et al., 2014; Ogden et al., 2014; Paulino et al., 2015).

The purpose of this paper is to describe the anthropometric standardization process, protocol, lessons learned, and findings of the Children's Healthy Living (CHL) Program for Remote Underserved Minority Populations of the Pacific. CHL is a multicomponent program in the USAP designed to prevent child obesity by building regional capacity, including a multicenter trial testing a community-based environmental intervention to prevent childhood obesity and to promote healthful behaviors (Novotny et al., 2013b; Wilkens et al., 2013). The 11 jurisdictions of CHL in the USAP region include Alaska, American Samoa, the Commonwealth of the Northern Mariana Islands (CNMI), Guam, Hawai'i, and the Federated States of Micronesia (FAS), which include the Republic of Palau, the Republic of the Marshall Islands (RMI), and the Federated States of Micronesia (Chuuk, Kosrae, Pohnpei, and Yap) (Fig. 1).

METHODS

To ensure a unified, standardized protocol for field data collection, 2–5 days of measurement and standardization trainings were held for each jurisdiction before measurement and data collection were started. The guidelines for the conduct of anthropometric standardization training were adapted from Lohman et al. (1988) and were incorporated into CHL standard operating procedures, which were then compiled into a field manual and provided to all trainees. In addition to CHL staff, local agencies (e.g., Head Start, The Supplemental Nutrition Program for Woman, Infants and Children [WIC]) in some jurisdictions had staff attend these trainings as well. The CHL expert measurement team, which included the Program Director (RN) and Assistant Program Director (MKF), were in charge of training and reviewing the measurements in each jurisdiction's training sessions. The Program Director (RN) served as the expert anthropometrist and the Assistant Program Director (MKF) served as the standard recorder for each training session. The expert anthropometrist for the CHL standardization process is a professor in human nutrition who was trained and standardized by a physical anthropologist and has conducted more than 15 studies with anthropometry in the past three decades and has trained hundreds of measurers (Guerrero et al., 2008; Novotny et al., 2004; Novotny et al., 2007a; Novotny et al., 2007b; Novotny et al., 2013b).

All applicable institutional and governmental regulations regarding the ethical issues of human volunteers were followed. The standardization trainings for measurement were approved by the University of Hawai‘i Institutional Review Board (IRB). Jurisdiction-level approvals were also obtained from the University of Alaska Fairbanks and the University of Guam institutional review boards. Parent consent forms were obtained for all children
before each training session, and assents were obtained from all children before measurement occurred.

**Anthropometric measurement**

Child height was measured by a Portable Adult/Infant Measuring Unit stadiometer (Model PE-AIM-101, Perspective Enterprises, Portage, MI). This device has a flat vertical surface on which a measuring rule is attached. The stadiometer also has an attached movable headpiece with a screw to hold it in place. Per the protocol, the portable stadiometer was to be positioned on a level floor without carpeting, and flush against a wall or other flat surface. Height was measured with the children barefoot, or with light socks (Alaska), to the nearest 0.1 cm. The instrument was calibrated with a 140 cm aluminum rod before each measurement session or every time the stadiometer was moved.

Child weight was measured using a portable SECA 876 scale (SECA 876, Hamburg, Germany). The scales came equipped with four adjustable feet and a bull’s eye spirit level. The scales were placed on a noncarpeted floor and leveled as needed to ensure accuracy. Children were weighed with bare feet, or with light socks (Alaska), and wearing only light weight clothing (e.g., shorts and t-shirt). Weight was measured to the nearest 0.1 kg. The scale was calibrated with a 4.5 kg weight before each measurement session or every time the scale was moved.

Child waist circumference was measured with a SECA 201 circumference measuring tape (SECA 201, Hamburg, Germany) to the nearest 1 mm. The measuring tape needed to be flush to the skin at the child’s umbilicus, which provides similar measures in children as the iliac crest (Lohman et al., 1988). Measurers waited to make sure the child was breathing normally prior to taking the measurement and was in standard position (standing on both feet, looking ahead, arms folded across chest). The measurement was taken at the midpoint between inspiration and expiration since breath control is difficult in young children. The measuring tape was calibrated against the 140 cm aluminum rod before every measurement session.

For each training session, the expert team explained the anthropometry protocol, which the trainees had reviewed beforehand. The NHANES III Anthropometric Procedure video was shown and trainees practiced measuring each other to familiarize themselves with the procedure and equipment before measuring children. The trainees then measured the designated anthropometry components of the child volunteers. Two trainees were paired as a team to measure and record each anthropometric component for each child participant. One trainee took the measurements and the other recorded the values, followed by providing the values to the measurer to view and confirm agreement or to correct or redo the measure. Then, the two trainees traded roles. Every trainee eventually assumed the role of measurer and recorder for each child. The goal was for each trainee to measure eight children aged 2–8 years old at a session; effort was made to recruit a minimum of 10 children, as a few children at each session would become fatigued before being measured by all trainees. Upon completion of the session, measurements were then compared against those of the expert anthropometrist.

Three anthropometric components were standardized at each training session: weight, height, and waist circumference. Each child was measured at least three times for each component by each measurer, including the expert anthropometrist. If no two measurements were within two units (0.2 kg for weight and 0.2 cm for height and waist circumference), the measurer was instructed to repeat the measurement until there were at least two measurements within two units. Staff were allowed to cross out measurements they felt the least confident about as long as at least three measurements remained. The average of all available measurements of an anthropometric component was used in the assessment (de Onis et al., 2004).

**Anthropometry standardization procedures and protocols**

Zerfas criteria, developed to compare the measurements of trainees against the measurements of the expert anthropometrist, were used to assess each trainee (Zerfas, 1985). Initially, the trainee measurement for each child was rated (Step 1) and then the trainee performance across children was rated (Step 2). In Step 1, the trainee measurements for each component and each child were rated based on the bias, defined as the difference (d) in the mean values of the component for the expert and for the trainee. Zerfas provided two tiers of rating: the Standard (more stringent) criteria or the Alternate (more lenient) criteria (Table 1). Each rating system categorizes each measurement based on magnitude of the bias into one of the following four categories: Good, Fair, Poor, and Blunder Errors. In the Zerfas manual, there is no specific criteria given for waist circumference; however, as other anthropometric components that are given in mm units (height and arm circumference) have the same Zerfas criteria, those Zerfas criteria were applied to waist circumference in this study.

In Step 2, the individual measurement ratings for an anthropometric component were aggregated across children to assess the trainee’s overall performance. The Zerfas criteria at this step categorizes the overall performance as Pass, Borderline Pass, No Pass, or No Pass by

---

**Table 1. Evaluation of measurement error for the Children’s Healthy Living Program anthropometry trainees for each child by the Zerfas criteria (modified from Zerfas, 1985)**

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Zerfas Criteria</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
<th>Blunder (gross error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height and arm circumference (cm)</td>
<td>AZ</td>
<td>0.6–0.9</td>
<td>1.0–1.9</td>
<td>2.0–2.9</td>
<td>&gt;2.0</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>SZ</td>
<td>0.2</td>
<td>0.3–0.4</td>
<td>0.5–0.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

*SZ, standard Zerfas; AZ, alternate Zerfas.

The Zerfas criteria for height and arm circumference were applied to waist circumference in this study.
Table 2 gives the number of unsuccessful measurements for each trainee for the Anthropometric component. For components where there was no "blunder," the overall evaluation was determined by the number of unsuccessful measurements at Step 1, which were those that fell into the category of "Fair" or "Poor." The number of unsuccessful measurements is compared to the expected number based on a binomial distribution with parameters $P = 0.05$ and $M = \text{number of children measured.}$. "No Pass" is assigned when $n_F$ is greater than or equal to $n_4$, the minimum number meeting the criteria $\text{Prob}(X \geq M, P) < 0.05$. "Pass" is assigned when $n_F$ is less than or equal to $n_3$, the maximum number meeting the criteria $\text{Prob}(X \leq M, P) < 0.05$. "Borderline Pass" is assigned when $n_3 < n_F < n_4$.

Table 2 gives the number of unsuccessful measurements for each trainee. The number of unsuccessful measurements for a trainee are based either on the standard or alternate Zerfas criteria in Table 1, respectively, depending on whether a more stringent or lenient approach is desired. The numbers in Table 2 mirror those in Zerfas (1985) and were expanded to larger values of $M$. It is advantageous to have more than eight children measured at a training session, as one unsuccessful measurement with five to seven children will usually result in a no pass rating at Step 2 for that component.

We chose to accept a rating of "Pass" or "Borderline Pass" as our criterion for qualifying to measure the anthropometric component in the field and a rating of "No Pass" or "No Pass by Blunder" as our criterion for not qualifying for that component. Only trainees who measured at least eight children for each component at the training session were evaluated. We also used the more stringent standard Zerfas criteria at a trainee's first session and the more lenient alternate Zerfas criteria at subsequent sessions, in order to ensure that trainees received a sufficient level of experience with measurement (i.e., at least two sessions unless the trainee was very experienced at the outset). If a trainee failed the standardization for any of the three anthropometric components, he/she had to participate in another training session. Only anthropometric components that were not passed were required to be repeated. At an individual's second or any other subsequent training session, the Alternate Zerfas criteria were used. Trainees passing any of the three anthropometric components by the Standard Zerfas criteria at his/her first training session or by the Alternate Zerfas criteria at any subsequent training session were qualified and permitted to measure in the field. No one was permitted to measure in the field that did not pass by at least the Alternate Zerfas criteria at his/her second or subsequent training sessions.

Data analysis

Anthropometric standardization results from each training session were sent to the CHL Coordinating Center (Honolulu, HI) to be analyzed. A SAS program (SAS Institute Inc., Cary, NC) was developed to evaluate each standardization training session. An evaluation report was produced for each training session that provided...
information on each trainee’s ratings for each measure based on the Zefras criteria.

The technical error of measurement (TEM) is commonly reported as a measure of imprecision in anthropometric assessment (Chumlea et al., 1990; de Miguel-Setay, et al. 2014; Johnson et al., 1997; Marks et al., 1989; Moreno et al., 2003; Perini et al., 2005; Stomfai et al., 2011; Uljasahek and Kerr, 1999; WHO Multicentre Growth Reference Study Group, 2006). It is the square root of the measurement error variance. Intra-rater TEM is a measure of the imprecision about the rater-specific means and is computed as

\[ TEM_{\text{intra}} = \sqrt{\frac{\sum_{c=1}^{C} \sum_{t=1}^{T_c} \sum_{r=1}^{R_c} (M_{ctr} - M_{ct})^2}{CT_c(R_c - 1)}} \]

where \( M_{ctr} \) is the mean for child \( c \) (\( c = 1 \ldots C \)) for trainee \( t \) (\( t = 1 \ldots T_c \)) for replicate \( r \) (\( r = 1 \ldots R_c \)), and \( M_{ct} \) is the mean of the \( R_c \) replicates for the \( T_c \) trainees that measured child \( c \): \( M_{ct} = \frac{\sum_{r=1}^{R_c} M_{ctr}}{R_c} \). Therefore, there is one \( M_{ct} \) per trainee and child combination.

Inter-rater TEM is a measure of the imprecision of the group-specific means and is computed as

\[ TEM_{\text{inter}} = \sqrt{\frac{\sum_{c=1}^{C} \sum_{t=1}^{T_c} (M_{ct} - M_c)^2}{CT_c}} \]

where \( M_c \) is defined as above and \( M_{ct} \) is the mean across the \( T_c \) trainees of \( M_{ct} : M_c = \frac{\sum_{t=1}^{T_c} M_{ct}}{T_c} \). Therefore, there is one \( M_c \) per child.

Total TEM is the aggregate measure of imprecision and is computed as

\[ TEM_{\text{total}} = \sqrt{TEM_{\text{intra}}^2 + TEM_{\text{inter}}^2} \]

A related statistic, the coefficient of reliability, \( R \), estimates the proportion of the total variance that is not due to measurement error (Chumlea et al., 1990; Johnson et al., 1997; Marks et al., 1989; Moreno et al., 2003; Perini et al., 2005; Stomfai et al., 2011; Uljasahek and Kerr, 1999; WHO Multicentre Growth Reference Study Group, 2006). It is defined as

\[ R = 1 - \frac{TEM_{\text{total}}^2}{SD^2} \]

where \( SD^2 = \frac{\sum_{c=1}^{C} \sum_{t=1}^{T_c} \sum_{r=1}^{R_c} (M_{ctr} - M_c)^2}{CT_c(R_c - 1)} \).

Average bias was assessed in terms of magnitude and whether or not the trainees systematically overestimated or underestimated measurements compared to the expert anthropometrist (WHO Multicentre Growth References Study Group, 2006).

**RESULTS**

**Anthropometric standardization results**

First session results. Among the 79 trainees, 58% (n = 46) passed weight and another 19% (n = 15) passed weight at borderline using the Standard Zefras criteria. For height, only 13% (n = 10) passed and another 23% (n = 18) passed at borderline. For waist, only 8% (n = 5) passed and another 10% (n = 8) passed at borderline.

Although the Alternate Zefras criteria were not used for the evaluation of an individual’s first training session, the results would be much improved had the Alternate Zefras criteria been used. For example, for weight, 73% (n = 58) would have passed and another 20% (n = 16) would have passed at borderline. For height, 49% (n = 39) would have passed and an additional 27% (n = 21) would have passed at borderline. For waist, 21% (n = 17) would have passed and an additional 25% (n = 20) would have passed at borderline.

Second session results. Among the 79 trainees, 57 (72%) participated in a second training session. Measurement improvement was seen in all three components (Fig. 2). The percentage of trainees that passed or passed at borderline for waist by the Standard Zefras criteria increased by 50%, from 14% in the first session to 21% in the second training. For height, there was a 59% increase in the pass rate (from 32% to 51%). For weight, there was a 24% increase (from 74% to 92%). In contrast, increase in rates of passing or passing at borderline for all three components was substantially smaller when the Alternate Zefras criteria was used, e.g., 3% for weight, 26% for height, and 45% for waist.

Overall session results. Among the 79 trainees, a total of 49 (62%) passed the criteria and were qualified to measure all three anthropometric components in the field. Among those 49 trainees, 2 (4%) passed at their first training session, 23 (47%) passed by their second session, and 17 (35%) passed at their third session. The rest, seven (16%), passed at a session/s subsequent to their third session.

**Intra- and inter-rater TEM, average bias, and reliability score**

Intra- and inter-rater TEMs, average bias, and the coefficient of reliability score (R) statistics were calculated.
first for all 79 trainees, then for the 49 trainees who passed the criteria for all three anthropometric components and the 30 who did not pass (Table 3). Blunder types of errors were removed from these calculations. The intra-rater TEMs for the components were relatively similar between the 49 who passed the criteria and the 30 who did not pass, indicating that the reliability of measurements within trainees about their respective means was small, regardless of whether they passed the Zerfas criteria. However, the inter-rater TEMs were much larger for the 30 who did not pass than for the 49 who did pass for all components, indicating that the measurements for the trainees who passed were close to a common mean for each child, while those that did not pass were not converging to a common mean. Intra-rater TEMs of the expert anthropometrist were 0.20 for height, 0.17 for waist, and 0.02 for weight. The coefficient reliability scores were higher than 0.99 for all three components among those 49 who passed, compared to 0.95 for height, 0.80 for waist, and 0.82 for weight among those 30 trainees who did not pass. Average bias for both the height and waist was negative, −0.32 for height and −0.19 for waist among those 49 trainees who passed the criteria, indicating that in general, trainees tended to underestimate height and waist measures. The average absolute biases for both height and waist were larger for those that did not pass than those who did, as would be expected based on the Zerfas criteria. Average bias for weight was small at 0.01 among all trainees.

**Lessons learned**

Measurement from a minimum of eight children allows for the most efficient situation to evaluate the Zerfas criteria among trainees. To achieve the minimum number of measures, children were often measured multiple times; therefore, techniques to engage children in the training sessions were important. Successful strategies included starting the measurement with children who were more eager to be measured, or starting the measurement with an anthropometric component that required minimum touching (e.g., weight), or letting the child pick a measure to start. In addition, giving each child a small reward, e.g., a colorful sticker, after each measure, was a useful strategy.

Although having at least eight children for a training session was found to be beneficial, increasing the number of trainees in each session beyond 10 was difficult to manage. With more trainees, it became a challenge to ensure that all trainees measured the minimum number of children that were also measured by the expert anthropometrist within a reasonable time period. The development of a grid that trainees could use to track the children they had measured in comparison to the expert anthropometrist was found to be useful and motivating both to children and trainees. Lastly, practicing recording, with verification by the measurer, was just as important as measuring. Most of the blunders were due to recording errors.

**DISCUSSION**

CHL's anthropometric standardization process and protocols were innovative and unique in several ways. The process included the use of standardized measurement protocols and equipment at 11 US-Affiliated Pacific jurisdictions, the evaluation of measurement teams from all jurisdictions against a single expert anthropometrist, and the use of Zerfas criteria. Due to the multicenter design of CHL and the unique geographical distances, the risk of measurement error due to variation in technique and skill was high, and adherence to the study protocol between members of the field staff was important. For this reason, an emphasis was placed on the standardization of anthropometric measurements to minimize measurement error.

Criteria for anthropometric assessment have been put forward by Zerfas, using a repeated-measure protocol (Ulijaszek, 1997; Zerfas, 1985). The trainee and trainer measure the same individuals until the difference between the trainee and trainer is “good” by standard Zerfas criteria, or at the very least, “fair” by the alternate Zerfas criteria (Ulijaszek, 1997; Zerfas, 1985) (Table 1). Our experience was that the standard Zerfas criteria were very strict and few trainees could pass their first standardization training using this criteria, e.g., only two of the 79 trainees passed all three anthropometric

---

**Table 3. Intra- and inter-rater (trainee in this study) technical error of measurement (TEM), bias relative to the expert anthropometrist, and reliability score from the 29 anthropometry standardization sessions from the Children’s Healthy Living program**

<table>
<thead>
<tr>
<th>Components</th>
<th>Reliability statistics</th>
<th>Expert anthropometrist</th>
<th>Among all 79 trainees</th>
<th>Among 49 passed the criteria</th>
<th>Among 30 not passing the criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Intra-TEM 0.20a</td>
<td>0.28</td>
<td>0.27</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inter-TEM 1.44b</td>
<td>0.30</td>
<td>0.30</td>
<td>2.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average bias</td>
<td>−0.33</td>
<td>−0.32</td>
<td>−0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R coefficient</td>
<td>0.984</td>
<td>0.999</td>
<td>0.951</td>
<td></td>
</tr>
<tr>
<td>Waist</td>
<td>Intra-TEM 0.17</td>
<td>0.32</td>
<td>0.30</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inter-TEM 2.02</td>
<td>0.49</td>
<td>0.49</td>
<td>4.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average bias</td>
<td>−0.20</td>
<td>−0.19</td>
<td>−0.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R coefficient</td>
<td>0.936</td>
<td>0.994</td>
<td>0.801</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>Intra-TEM 0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inter-TEM 1.69</td>
<td>0.06</td>
<td>0.06</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average bias</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R coefficient</td>
<td>0.938</td>
<td>0.999</td>
<td>0.821</td>
<td></td>
</tr>
</tbody>
</table>

aBlunder type of errors, defined as the mean difference between the expert anthropometrist and a trainee of 2 cm or larger for height and waist circumference, or 0.5 kg or larger for weight were excluded for the calculation of all statistics presented in this table.
bThe intra-TEM of the expert anthropometrist was based on all measurements taken by the expert anthropometrist across the 29 training sessions for a specific measurement component.
cThe intra-TEM among the trainees based on all measurements taken by all trainee across the 29 training sessions for a specific measurement component.
dThe inter-TEM among all trainees, excluding the expert anthropometrist.

---

American Journal of Human Biology
components at their first training session by the standard Zerfas criteria.

Waist was the most difficult measurement for the trainees to pass, and weight was the least difficult. This is in accordance with findings from other studies (WHO Multicentre Growth Reference Study Group, 2006; de Miguel-Etayo et al., 2014) which reported that the “problem” measurements were those that require careful positioning of the child. Accurate weight measurement required that the scale be level and the child stand still in the middle of the digital scale to provide accurate measures, a technique which was easily learned by the trainees. Accurate height measurement required that the child’s posture be monitored and the headboard of the stadiometer be manipulated, a skill that took practice. Accurate waist measurement required that the trainee manage the posture and breathing of the child, manipulate the tape to ensure it is flat, level, and neither too lax nor too taut, a technique that required the most practice to learn. Most trainees had to learn to loosen the cinching of the tape and ensure that children were not holding their breath or holding in their abdomens.

Measurements from the 49 trainees who passed all criteria to become field anthropometrists were precise, as measured by the intra-rater TEM, which were less than twice the expert’s TEM in all three components and were comparable to what were reported in the literature (Lohman et al., 1988). This is also shown in the $R$ coefficients for all three anthropometric components, which were higher (all $>99\%$) than the 90% reliability threshold suggested by Marks and colleagues (1989) and what has been published in the literature (Chumlea et al., 1990; de Miguel-Etayo, et al., 2014; Johnson et al., 1997; Moreno et al., 2003; Perini et al., 2005; Stomfai et al., 2011; Ulijaszek and Kerr, 1999; WHO Multicentre Growth Reference Study Group, 2006). This might be explained by the CHL protocol that for each anthropometric component, the trainee measure the child at least three times with at least two measurements within two units (0.2 kg for weight and 0.2 cm for height and waist circumference).

The negative-signed bias for both height and waist indicated that trainees in general tended to underestimate those two measures. This may be explained by the lack in experience in positioning children or manipulating the instrument as compared to the expert anthropometrist, in particular, gaining full extension for height, with the head in the Frankfort plane, and waiting for the children to be sufficiently relaxed to not hold their breath and not cinching the tape too tight, for waist circumference. Nevertheless, the magnitude of biases for all three components in the team’s measurements were much smaller than the maximum allowable difference set by the WHO Multicentre Growth Reference Study study at 2.8 times the expert’s TEM (WHO Multicentre Growth Reference Study Group, 2006).

This is one of the first anthropometric standardization protocols that used a single expert anthropometrist. This protocol requires a great deal of effort, and in this case travel, on the part of the expert team. It was unclear at the outset whether this approach was necessary to ensure accurate measurement by anthropometrists in diverse settings with little direct supervision. Our experience demonstrates that this training is optimal, particularly for our large and diverse region and team composition. Precision and accuracy were high among those 49 trainees who passed the criteria for all three measures. The distinction between trainees who did and did not pass the Zerfas criteria appears to be in accuracy more than in variability and precision of measurements. The small intra-rater TEMs for the 30 trainees who did not pass indicate that they generally obtained similar repeat measurements for a single child (i.e., low variability); however, the large inter-rater TEMs indicate that their measures were not close to a common mean for that child (i.e., low accuracy). Conversely, the 49 trainees who passed were able to achieve consistent repeat measurements about a common mean. The 49 trainees started with very different technique and skill levels and types of measurement errors, but all moved closer in their values to the experts’ values and eventually passed, implying that a specific measurement technique can be achieved.

One of the goals of CHL’s anthropometry measurement training was to build capacity and infrastructure for data system development and monitoring for nutrition and health (Novotny et al., 2013b; Wilkens et al., 2013). This was particularly important because there is currently no uniform anthropometry monitoring in the region, as NHANES does not include this region, and population obesity rates are higher than the contiguous US and known global prevalence rates (Bruss et al., 2010; Leon-Guerrero et al., 2008; Novotny et al., 2014; Ogden et al., 2014). The anthropometric standardization process and protocols followed by CHL have helped build capacity in the health workforce in the Pacific Region. The 49 standardized field anthropometrists are a first step of this capacity building.

Author Contribution

F. Li participated in study design, led the data analysis, interpretation of data and writing of the manuscript, and had primary responsibility for the integrity of the data, accuracy of the data analysis and the final article. L.R. Wilkens led the study analytic strategy, interpretation of data, and took responsibility for the integrity and the accuracy of the data analysis. R. Novotny led the study, acquisition of the data and interpretation of data; M.K. Fialkowski co-led the study design, acquisition of the data and interpretation of data. Y.C. Paulino participated in design, data acquisition and led interpretation of Guam data. R. Nelson participated in design, led in data acquisition and interpretation of Commonwealth of the Northern Marianas data. A. Bersamin participated in design, led in data acquisition and interpretation of American Samoa data. J. Deenik participated in design, led in data acquisition and interpretation of Alaska data. U. Martin participated in design, led in data acquisition and interpretation of American Samoa data. J. Deenik participated in design and led in interpretation of data of Freely Associated States (Republic of Palau, Republic of Marshall Islands, and Federated States of Micronesia (in this study Yap, Kosrae, and Pohnpei)). C.J. Boushey participated in study design and data interpretation. All authors critically reviewed and approved the final manuscript.

REFERENCES
