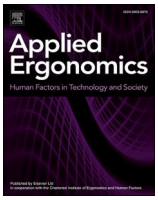
Applied Ergonomics 116 (2024) 104218

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo



Intra-rater and inter-rater reliability of 3D facial measurements

Kayna Hobbs-Murphy [a,*], Isabel Olmedo-Nockideneh [b], William J. Brazile [b], Kristen Morris [a], John Rosecrance [b]

[a] Department of Design and Merchandising, Colorado State University, 1574 Campus Delivery, Fort Collins, CO, 80523-1574, United States
[b] Department of Environmental and Radiological Health Sciences, Colorado State University, 1681 Campus Delivery, Fort Collins, CO, 80523-1681, United States

ARTICLE INFO

Keywords:
Three-dimensional body scanning
Facial anthropometry
Anthropometric measurement reliability

ABSTRACT

Three-dimensional (3D) body scanning technology has applications for obtaining anthropometric data in human-centered and product development fields. The reliability of 3D measurements gathered from 3D scans must be assessed to understand the degree to which this technology is appropriate for use in place of manual anthropometric methods. The intra- and inter-rater reliabilities of 3D facial measurements were assessed among four novice raters using 3D landmarking. Intraclass correlation coefficient (ICC) statistics were calculated for the 3D measurement data collected in three phases to assess baseline reliabilities and improvements in reliabilities as the result of additional training and experience. Based on the results of this study, the researchers found that the collection of 3D measurement data, by multiple raters and using 3D landmarking methods, yielded a high percentage of ICC statistics in the good to excellent (>0.75 ICC) reliability range. Rater training and experience were important considerations in improving intra- and inter-rater reliabilities.

1. Introduction

In an ergonomic product development process, anthropometric data of the head and face are used to help sizing and design of important headwear products such as spectacles (Kouchi and Mochimaru, 2004), bicycle helmets (Pang et al., 2018; Skals et al., 2016), medical headsets (Lacko et al., 2017), and personal protective equipment (Coblentz et al., 1991; Lee et al., 2013). Three-dimensional (3D) body scanning allows for the rapid, contact-free collection of anthropometric measurement data. 3D scanning (i.e., 3D scan) creates a "virtual twin" (Kuehnapfel et al., 2016, p. 1) of the scanned person/object, which allows researchers to revisit the 3D scan as needed to collect detailed measurement information. Measurement data collected from 3D scans (i.e., 3D measurements) provide more contextual anthropometric information about facial surface dimensions than manually-collected measurement data, such as lengths along the surface of the face (Bailar et al., 2007; benAzouz et al., 2006). Due to this higher level of anthropometric context, researchers (Ban and Jung, 2020; Coblentz et al., 1991; Goto et al., 2019; Lacko et al., 2015; Lee et al., 2018; Park et al., 2021) have begun to rely on 3D body scanning tools to collect anthropometric data from the head.

3D measurement research features a wide range of methods including different 3D tools, various forms of data collection methods, and several appropriate statistical analyses (Bragança et al., 2016). A systematic review of literature by Viviani et al. (2018) revealed that intra- and inter-rater reliabilities of anthropometric data are seldom reported in anthropometric research, despite these being the "most troublesome source of anthropometric error" (p. 7). Previous researchers have evaluated intra- and inter-rater reliability of 3D facial measurements using several combinations of these 3D tools, methods, and analyses (Aynechi et al., 2011; Coward et al., 1997; Düppe et al., 2018; Franco de Sá Gomes et al., 2019; Kim et al., 2018; Kuehnapfel et al., 2016; Wong et al., 2008). Authors of a meta-analysis of 3D facial measurement reliability have suggested that researchers may be able to reliably analyze 3D facial anthropometry without manually landmarking the face prior to 3D scanning (Gibelli et al., 2020). However, Fagertun et al. (2014) noted that 3D landmarks around the jaw are often difficult to place due to a lack of bony definition around the jaw area. Furthermore, facial accessories such as facial hair and eyeglasses often occlude and make 3D landmark placement difficult in several areas of the face (Srinivasan and Balamurugan, 2014). Modabber et al. (2016) and Ayaz et al. (2020) found that a combination of manually landmarking the face prior to 3D scanning and 3D landmarking the scans helped improve data reliability and contributed to increased data collection accuracy in their studies. Despite these findings, there is interest in the 3D measurement research community to move away from

* Corresponding author.
E-mail address: Kayna.Hobbs-Murphy@colostate.edu (K. Hobbs-Murphy).

https://doi.org/10.1016/j.apergo.2023.104218
Received 20 July 2023; Received in revised form 30 November 2023; Accepted 21 December 2023
Available online 5 January 2024

manual landmarking in favor of time efficiency and reduced contact with human participants (Bragança et al., 2016; Franco de Sá Gomes et al., 2019; Gibelli et al., 2020). Furthermore, as 3D scanning technology advances and diversifies, reliability assessments must be done to ensure the appropriateness of 3D scanning in facial anthropometric research (Aynechi et al., 2011; Bragança et al., 2016; Düppe et al., 2018; Franco de Sá Gomes et al., 2019; Kim et al., 2018; Kuehnapfel et al., 2016; Modabber et al., 2016; Wong et al., 2008).

The aim of the current study was to assess the intra- and inter-rater reliability of 3D facial measurements gathered from 3D scan data by four novice anthropometric raters using 3D landmarking methods. In the present study, intra-rater reliability (intraRR) was defined as the degree of agreement among collections of a 3D measurement performed on the same subject by a single rater, and inter-rater reliability (interRR) as the degree of agreement across all raters who collect 3D measurements on the same subjects. The four research questions (RQ) for accomplishing the aim of this study included.

RQ1: What percentage of good to excellent (>0.75 ICC statistic) intra-rater reliability (on average across four raters) is able to be achieved by the final phase of data collection?
RQ2: What percentage of good to excellent (>0.75 ICC statistic) inter-rater reliability is able to be achieved by the final phase of data collection?
RQ3: In percentage terms and averaging across four raters, how much does intra-rater reliability improve over two phases of data collection?
RQ4: In percentage terms, how much does inter-rater reliability improve on average over three phases of data collection?

Previously, researchers have used intraclass correlation coefficient (ICC) statistics (Koo and Li, 2016) to evaluate 3D measurement data reliability (Franco de Sá Gomes et al., 2019). Therefore, RQ1 and RQ2 address intra- and inter-rater reliabilities respectively by observing ICC statistics calculated from 3D measurements. Further, previous researchers have suggested that providing anthropometric training to raters prior to data collection may improve reliability scores (Düppe et al., 2018). Additionally, Androutsos et al. (2020) and de Miguel-Etayo et al. (2014) observed improved interRR as the result of multiple group training sessions for raters. Therefore, the authors were interested in observing improvements in intra- and inter-rater reliabilities for RQ3 and RQ4 by observing improvements in ICC statistics calculated from 3D measurements collected multiple times.

2. Methods

3D scan data for this study were purchased from Human Solutions of North America, Inc., an anthropometric data collection company with over 30 years of experience in 3D scanning and sizing surveys (Human Solutions, n.d.). The company collected 3D facial scans from participants using a handheld Artec Eva 3D (Model Eva, Senningerberg, Luxembourg) structured-light scanner (Artec3D, n. d.). Human Solutions technicians collected scans from volunteer participants at their headquarters in Morrisville, North Carolina. Each scan participant was asked to wear a swim cap during the scan process so that the shape of their head may be accurately scanned. If a participant had long hair, they were asked to gather their hair to a bun style on the top of their head and pull the bun through a hole at the top of the swim cap. All participants were asked to remove glasses and jewelry for the scan process. After participants were scanned, the Human Solutions technicians processed each facial scan by removing stray artifacts and making the 3D models watertight and ready for analysis. The researchers purchased 2022 total 3D scans from Human Solutions, of which 30 (1.5%) scans were randomly selected for this intraRR and interRR study.

There were two general stages to the research procedure: first, training the raters on 3D landmark placement, and second, 3D measurement collection. Within the 3D measurement collection stage, the procedure was to 1) use a software plug-in wizard to view the 3D measurements generated between 3D landmarks, 2) modify 3D landmark location (if necessary), and 3) export final 3D measurements for intraRR and interRR analysis (detailed below). Data collection was completed over three phases to assess improvements in reliability over time.

2.1. Rater training on 3D landmark placement

Four raters (Rater A, B, C, and D) were involved in 3D measurement data collection for this research. At the time of data collection, two raters were undergraduate students and two were graduate students. Many research efforts rely on student research assistants; thus, the present research provides information about non-expert data collection reliability that is applicable to research efforts driven by student and expert staff alike. The two graduate student raters were recruited through the authors' programs of study, and were compensated as part of their contractual research appointment. The two undergraduate students were recruited using a combination of institutional job posting sites and author recommendation. Post-interview and hire, the undergraduate students were compensated $20 per hour for their work as raters in this 3D measurement data collection effort.

Previous researchers have suggested that providing anthropometric training to raters prior to data collection may improve reliability scores (Bragança et al., 2016; Düppe et al., 2018). Therefore, before raters collected any 3D measurement data, they were provided with a guide to the 3D facial landmarks and measurements with which to familiarize themselves. The guide included 18 3D facial landmarks (illustrated in Fig. 1, described in Table 1), which served as the endpoints to the 27 3D facial measurements collected as data in this research study.

Next, the raters were asked to watch a video tutorial on how to place 3D landmarks in Anthroscan, Human Solutions' proprietary 3D scan software, Anthroscan (Version 3.6.1, Kaiserslautern, Germany). In the video tutorial, the primary investigator explained how to use the mouse to zoom in, move the 3D scan, place a 3D landmark, and visually check each 3D measurement for accuracy. In this way, landmarking was done solely in 3D (3D landmarking) without manual landmarking done on the face prior to 3D scanning.

In the case where 3D landmarks were occluded by hair (facial or head) and/or glasses, raters were trained to discern if the 3D landmark could be carefully placed, or if to omit landmark placement in the occluded area(s). When raters felt confident in their understanding of the 3D facial landmarks, they were asked to take a quiz testing their ability to recognize the 3D landmarks. The 18-question, multiple-choice quiz consisted of pictures of isolated facial landmarks from Fig. 1 and one correct 3D landmark name among four choices for each question. If raters did not select all 18 correct answers on the 3D landmark quiz, they reviewed the aforementioned 3D landmark and measurement guide and video tutorial, and re-attempted the quiz until full credit was received. Once raters received full credit on the 3D landmark quiz, they were prompted to begin 3D landmarking facial scans for training and then for 3D measurement data collection. Lastly, raters met with the senior researcher (first author) prior to starting the data collection process, to address questions. Throughout the research process, the raters were asked about the difficulty of placing 3D landmarks via an email questionnaire, with questions including "What did you find difficult about placing the landmarks?" and "What landmarks were the hardest to place?". These qualitative data were used to give context to interRR and intraRR scores for specific 3D measurements (i.e., if landmarks around the jaw were difficult to place, 3D measurements around the jaw may have lower reliability).

2.2. 3D measurement collection

After the raters were trained in Anthroscan (Version 3.6.1,

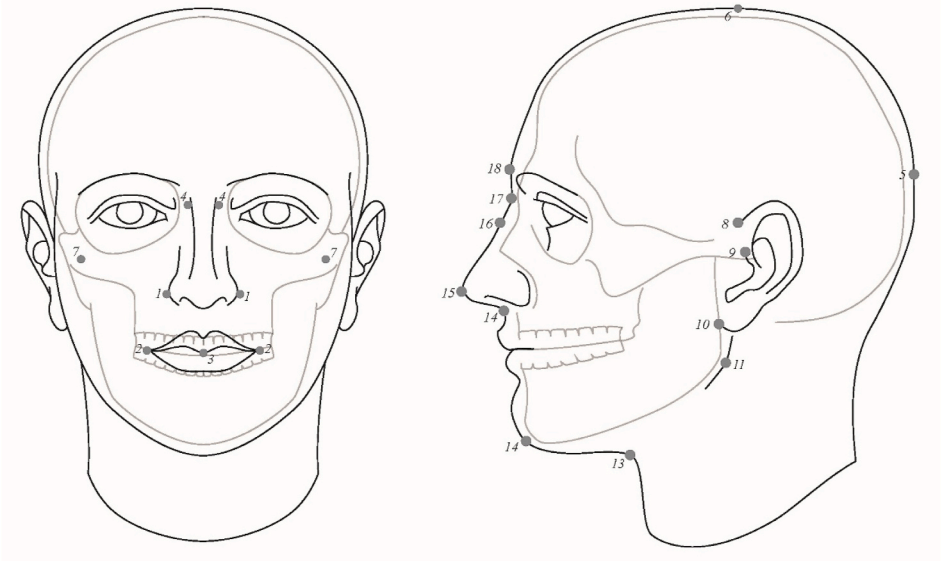


Fig. 1. Illustration of 3D facial landmarks placed on each 3D scan, used to collect 3D measurements seen in Fig. 2.

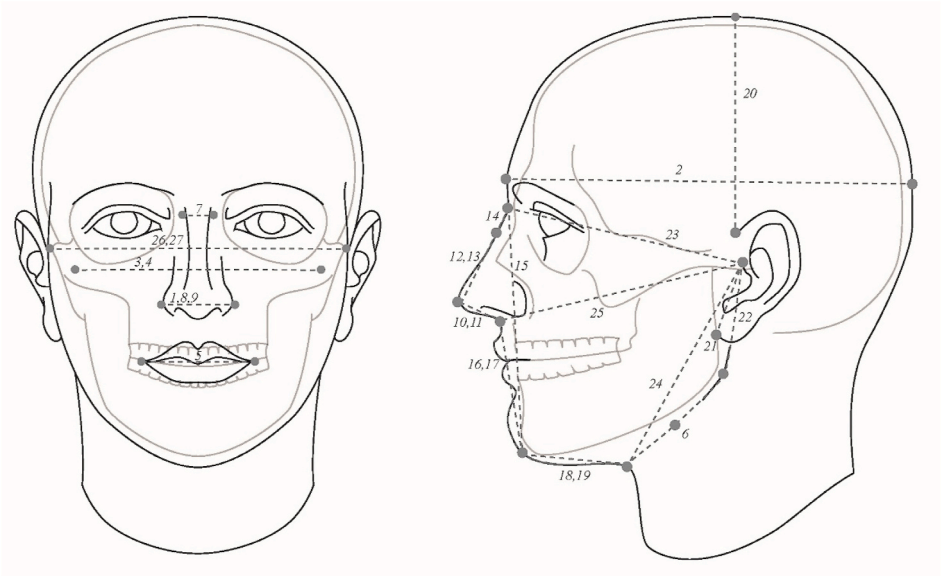


Fig. 2. Illustration of 3D facial measurements collected from each 3D scan.

Kaiserslautern, Germany) on how to manually place the 3D landmarks on the 3D face scans, the raters extracted 3D measurement data from each 3D scan using a measurement extraction program, referred to as the 'wizard'. The wizard, a custom-made measurement software plug-in program for Anthroscan (Bonin et al., 2019; Traumann et al., 2019), was developed in collaboration with Human Solutions. The wizard was developed specifically for this project because at the time of data collection, Anthroscan software did not yet allow for collection of 3D anthropometric data from 3D face scans. The wizard was programmed to automatically generate linear (direct from point to point) and contour (over the surface of the face) lengths between all 27 landmarks which were exported to a data sheet in Microsoft Excel. For most measurements, the wizard operated by gathering the shortest distance between two identified landmarks. In rare cases where the shortest distance between two landmarks did not provide an accurate measurement (such as Tragion to Tragion Contour or TrTr_C), the measurement was collected through a specified middle s landmark (such as the Sellion landmark for TrTr_C) or collected along a specific axis (X, Y, or Z). The 27 3D facial measurements collected using the Anthroscan wizard are illustrated in Fig. 2 and were used to assess intraRR and interRR ICC. The name of each measurement, whether the measurement was collected in a linear and/or contour fashion, and the abbreviated measurement name are described in Table 2. The present research study collected 27 3D measurements, compared to 11 3D measurements collected by Franco de Sá Gomes et al. (2019). The researchers chose these 27 facial measurements based on past large-scale studies on facial anthropometrics (Clauser et al., 1988; Gordon et al., 1989, 2014; Zhuang et al., 2005, 2007, 2008, 2010; Zhuang and Bradtmiller, 2005).

The raters operated the Anthroscan wizard by clicking through every 3D face landmark to extract each 3D measurement. At each 3D landmark, the rater assessed the measurement and used the measurement to check the landmark location. Meaning, once the measurement was known and able to be visualized, the rater was able to use the measurement to determine if the location of the 3D landmark was placed accurately. The measurements provided context for the raters to understand if the 3D landmark was accurate, or if the location of the

Table 1
3D landmark names (corresponding to Fig. 1) and an indication of single (center) or left and right (L&R) marking status.

Number in Fig. 1	Landmark Name	Single or L&R
1	Alare	L&R
2	Cheilion	L&R
3	Cheilion Center	Single
4	Nasal Root	L&R
5	Back of Head	Single
6	Top of Head	Single
7	Zygomatic	L&R
8	Otobasion	L&R
9	Tragion	L&R
10	Earlobe juncture	L&R
11	Gonion	L&R
12	Submandibular	Single
13	Menton	Single
14	Subnasale	Single
15	Pronasale	Single
16	Dorsal Hump	Single
17	Sellion	Single
18	Glabella	Single

Table 2
Measurement name (corresponding to Fig. 2), measurement type (linear, contour, or both), and abbreviated measurement name.

Number in Fig. 2	Measurement Name	Measurement Type	Abbreviated Name
1	Alare to Alare	Contour	AA_C
2	Back of Head to Glabella	Contour	BGl_C
3, 4	Bizygomatic Width	Both	BiW_L & BiW_C
5	Cheilion to Cheilion	Contour	ChCh_C
6	Gonion to Submandibular	Contour	GoSub_C
7	Nasal Root Breadth	Linear	NRB_L
8, 9	Pronasale to Alare	Both	ProA_L & ProA_C
10, 11	Pronasale to Subnasale	Both	ProS_L & ProS_C
12, 13	Sellion to Pronasale	Both	SelP_L & SelP_C
14	Sellion to Dorsal Hump	Contour	SelDH_C
15	Sellion to Menton	Linear	SelM_L
16, 17	Subnasale to Menton	Both	SnasM_L & SnasM_C
18, 19	Submandibular to Menton	Both	SmanM_L & SmanM_C
20	Top of Head to Otobasion	Contour	TrHO_C
21	Tragion to Earlobe Juncture	Contour	TrEJ_C
22	Tragion to Gonion	Contour	TrGo_C
23	Tragion to Sellion	Contour	TrSel_C
24	Tragion to Submandibular	Contour	TrSman_C
25	Tragion to Subnasale	Contour	TrSnas_C
26, 27	Tragion to Tragion	Both	TrTr_C & TrTr_L

landmark needed to be slightly moved. If a landmark was not able to be confidently identified and placed due to occlusion (facial or head hair, glasses, etc.), the corresponding 3D measurement(s) were not obtained from that 3D scan. When the rater was satisfied with the final landmark placement, the rater re-ran the Anthroscan wizard to get the final 3D measurements. These final 3D measurements were used to calculate the ICC statistics and analyze the agreement within (intraRR) and between (interRR) raters for each 3D facial measurement.

2.3. Three-phase data collection

3D measurement data were collected from a total of 30 3D scans (i.e., 30 scan subjects) in a three-phase data collection process. 3D measurement data were collected by the same set of raters throughout the study on the same scans, allowing for analysis of the rate of improvement in intraRR and interRR over time as the result of additional training and experience. IntraRR and interRR were both assessed in Phase 1 of the study by asking all four raters to collect 3D measurements from 10 scans (subjects #1–10), three times each. InterRR (but not intraRR) was assessed in Phase 2 of the study by asking all four raters to collect 3D measurements from 10 new scans (subjects #11–20), only one time each. Lastly, intraRR and interRR were again both assessed in Phase 3 of the study by asking all four raters to collect 3D measurements from 10 new scans (subjects #21–30), three times each. The outline of the three phases of data collection is provided in Table 3.

2.4. Statistical analysis

Intraclass correlation coefficient (ICC) statistics were calculated to analyze the agreement within (intraRR) and between (interRR) raters for each 3D facial measurement. Intraclass correlation coefficients are preferred (over *inter*class) when variables being measured are of a common class (McGraw and Wong, 1996), which was true in the present study. ICC statistics were calculated for all 27 3D measurements within raters (intraRR) and between raters (interRR), as applicable to the three phases (Table 3). In the case of a missing measurement value for one or more raters, the data point was assigned as 0 (millimeters) to allow for the assessment of agreement in landmark placement (or lack thereof) for all measurements. Based on published ICC statistic guidelines (Koo and Li, 2016), a two-way mixed effects model with "mean of the k raters" type (concerning mean value as basis for assessment) and "absolute agreement" data definition (concerning the same "score" or finding) was chosen (pp. 157–159). Analyses were conducted using RStudio (R Core Team, 2022) with packages tidyverse, irr, lpSolve (Berkelaar, 2022; Gamer et al., 2019; Wickham et al., 2019). The irr package was chosen over other available packages that analyze ICC statistics as it allowed the researchers to denote ICC characteristics in coding, such as the "two--way" model and the "agreement" definition.

3. Results

At the start of Phase 1, each rater required approximately 10 min to gather the 3D measurements from each 3D scan. By the end of Phase 3, each rater required about 5 min to gather the 3D measurements from

Table 3
Outline of three-phase 3D measurement data collection used to assess intraRR (Phase 1 and Phase 3), and interRR (Phase 1, Phase 2, and Phase 3).

	Reliability assessed	Data Collection Procedure		
		1st Collection	2nd Collection (repeated?)	3rd Collection (repeated?)
Phase 1	intraRR and interRR	Collected 3D measurement data from random scans #1–10 (1st time)	Yes, the collection of 3D measurement data from random scans #1–10 was repeated a 2nd time	Yes, the collection of 3D measurement data from random scans #1–10 was repeated a 3rd time
Phase 2	interRR	Collected 3D measurement data from random scans #11-20	No, the collection of measurement data from random scans #11–20 was not repeated	No, the collection of measurement data from random scans #11–20 was not repeated
Phase 3	intraRR and interRR	Collected 3D measurement data from random scans #21–30 (1st time)	Yes, the collection of 3D measurement data from random scans #21–30 was repeated a 2nd time	Yes, the collection of 3D measurement data from random scans #21–30 was repeated a 3rd time

each 3D scan. When data collection was complete, ICC statistics were calculated using 3D measurement data collected in metric units. The guidelines used for evaluating reliability, as related to ICC statistics, followed Koo and Li (2016). Negative ICC statistics indicate poor reliability and are often common when the number of data points analyzed is small (Liljequist et al., 2019; Taylor, 2009), which was true in the present study. Therefore, ICC statistics below 0.50 (including negative values) indicated poor reliability. ICC statistics between 0.50 and 0.75 indicated moderate reliability. ICC statistics between 0.75 and 0.90 indicated good reliability. ICC statistics greater than 0.90 indicated excellent reliability. The 95% confidence interval (CI) for ICC statistics are presented throughout the results, allowing for reliability to be assessed on a range of reliability categories, if applicable (Koo and Li, 2016). Negative ICC statistics were only observed in the 95% confidence intervals of the ICC statistics. The ICC statistical ranges and the corresponding reliability demonstrated in this study are quantified in Table 4. The results of this study are reported chronologically, starting with Phase 1, then Phase 2, and finally Phase 3.

3.1. Phase 1

3.1.1. IntraRR

In Phase 1, the range of intraRR ICC statistics was −0.28 to 0.99. Each rater's Phase 1 intraRR ICC statistic and 95% CI by abbreviated measurement location are presented in Fig. 3, with dashed lines indicating ICC statistical range limits for reliability categories presented in Table 4. In Fig. 3 and all subsequent figures, negative ICC values (observed only within the 95% CI) are illustrated by the solid 95% CI line passing over the y-oriented grid-line corresponding with 0.00 ICC. Each rater had at least one ICC statistic indicating poor intraRR. 3D measurements with poor intraRR were primarily unique to each rater, with Bizygomatic Linear Width (BiW_L) being the only 3D measurement for which two raters scored in the poor intraRR range.

3.1.2. InterRR

In Phase 1, the range of interRR ICC statistics was −0.20 to 0.98. Phase 1 interRR ICC statistics and 95% CI by abbreviated measurement location are illustrated in Fig. 4, with dashed lines indicating ICC statistical range limits for reliability categories. ICC statistics indicated poor interRR for 18.52%, moderate interRR for 14.81%, good interRR for 51.85%, and excellent interRR for 14.81% of the 27 3D measurements.

3.2. Phase 2

3.2.1. InterRR

In Phase 2, the range of interRR ICC statistics was 0.37–0.99. Phase 2 interRR ICC statistics and 95% CI by abbreviated measurement location are illustrated in Fig. 5, with dashed lines indicating ICC statistical range limits for reliability categories. ICC statistics indicated poor interRR for 14.81% of the 27 3D measurement locations, but not the same locations with poor interRR in Phase 1. ICC statistics indicated moderate interRR for 22.22%, good interRR for 18.51%, and excellent interRR for 44.44% of the 27 3D measurements.

Table 4
ICC statistic range and corresponding reliability, following guidelines from Koo and Li (2016).

ICC statistic range	intraRR and interRR
<0.50	Poor
0.50–0.75	Moderate
0.75–0.90	Good
>0.90	Excellent

3.3. Phase 3

3.3.1. IntraRR

In Phase 3, the range of intraRR ICC statistics was 0.38–0.99. Each rater's Phase 3 intraRR ICC statistic and 95% CI by abbreviated measurement location are illustrated in Fig. 6, with dashed lines indicating ICC statistical range limits for reliability categories. Two raters each had one ICC statistic indicating poor intraRR, while the two other raters had no ICC statistics indicating poor intraRR. 3D measurement locations with poor intraRR were unique to each rater and unique to Phase 3 in that they were not the same 3D measurement locations with poor intraRR in Phase 1.

3.3.2. InterRR

In Phase 3, the range of interRR ICC statistics was 0.04–0.99. Phase 3 interRR ICC statistics and 95% CI by abbreviated measurement location are illustrated in Fig. 7, with dashed lines indicating ICC statistical range limits for reliability categories. ICC statistics indicated poor interRR for 14.81% of the 27 3D measurement locations, (one of the same locations with poor interRR as Phase 2 and two of the same locations as Phase 1). ICC statistics indicated moderate interRR of 11.11%, good interRR for 29.63%, and excellent interRR of 44.44% of the 27 3D measurements.

4. Discussion

Intra- and inter-rater reliability were evaluated based on 3D facial measurements collected by the 3D landmarking of 3D facial scans by four raters using Human Solutions' Anthroscan software. The digital identification of 3D facial landmarks (Fig. 1, Table 1) resulted in the collection of 27 3D facial anthropometric measurements per scan, per phase (Fig. 2, Table 2). In response to RQ1, 90.74% good to excellent (>0.75 ICC) intra-rater reliability was achieved by Phase 3 on average across all four raters. In response to RQ2, 74.07% good to excellent (>0.75 ICC) inter-rater reliability was achieved by Phase 3.

In the only comparable reliability research using 3D facial scans collected by the Artec Eva 3D scanner, Franco de Sá Gomes et al. (2019) collected 11 3D measurements using 3D landmarking. Using an ICC statistic cutoff of >0.75 as excellent, researchers reported that 72.73% of their measures had excellent intraRR (only one rater collected 3D measurement data, averaging over multiple raters was not required) and 54.55% of their measures had excellent interRR over a single phase of data collection (Franco de Sá Gomes et al., 2019). In the present study, the four raters achieved an average of 90.74% intraRR above 0.75 by the third phase of data collection and achieved 74.07% interRR above 0.75 by the third phase of data collection. The raters in the present study collected over double the number of 3D measurements collected by Franco de Sá Gomes et al. (2019) (27 vs. 11). Therefore, in comparing the present study to Franco de Sá Gomes et al. (2019), the ICC statistics in the present study indicate a high percentage of intraRR and interRR.

4.1. Improvements in IntraRR

Each rater's Phase 1 and Phase 3 ICC statistics and intraRR respectively compared are shown in Figs. 8–15 (Rater A: Figs. 8 and 9, Rater B: Figs. 10 and 11, Rater C: Figs. 12 and 13, Rater D: Figs. 14 and 15). Each figure caption describes the percent of improved, constant, and/or decreased intraRR values for each rater. In response to RQ3, ICC scores for intraRR improved by 58.34% (averaged across four raters) over the two phases in which intraRR was assessed (Phase 1 and Phase 3). There was a higher proportion of excellent intraRR for all 3D measurements in Phase 3 than in Phase 1. Furthermore, fewer ICC statistics in the poor intraRR range for all raters were found in Phase 3. These findings indicate that additional training and experience with the 3D landmarking and 3D measurement collection process may have improved each rater's intraRR over time. In different terms, training to improve, and/or experience with, the data collection process may have improved

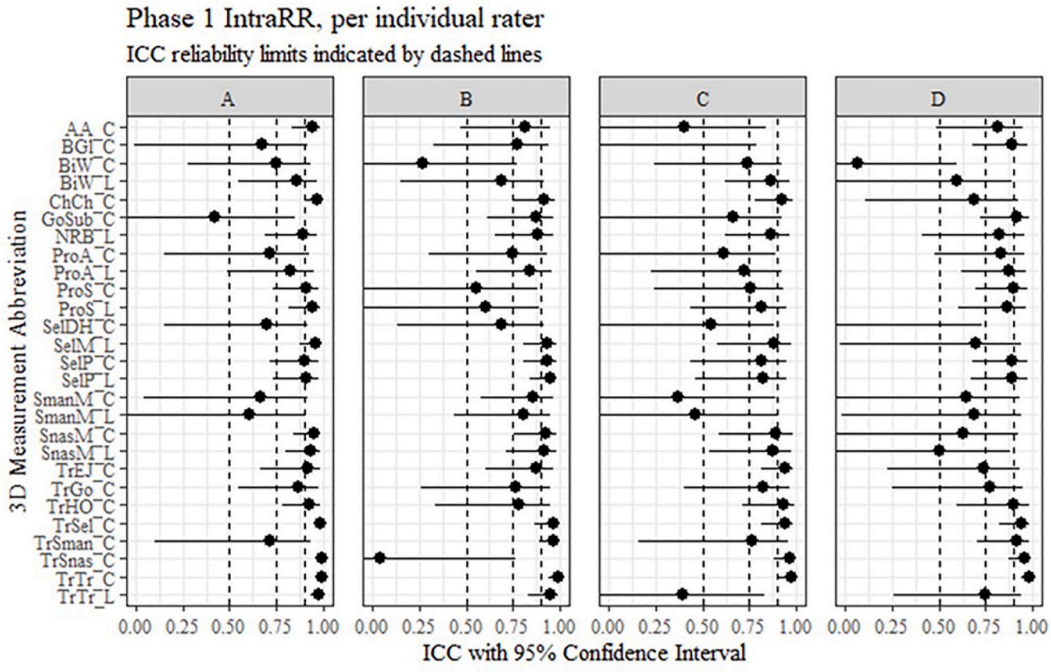


Fig. 3. Phase 1 intraRR ICC statistics and 95% confidence interval, for each individual rater (Rater A, Rater B, Rater C, and Rater D). Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.50–0.75, good 0.75–0.90, and excellent >0.90).

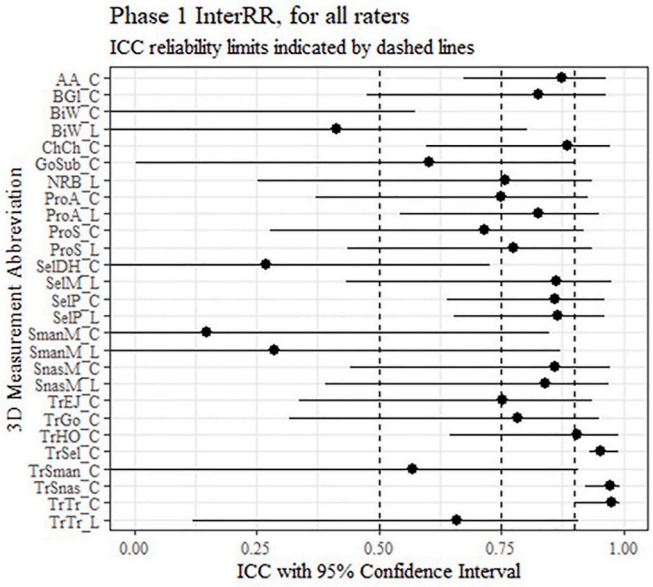


Fig. 4. Phase 1 interRR ICC statistics and 95% confidence interval, for across all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.50–0.75, good 0.75–0.90, and excellent >0.90).

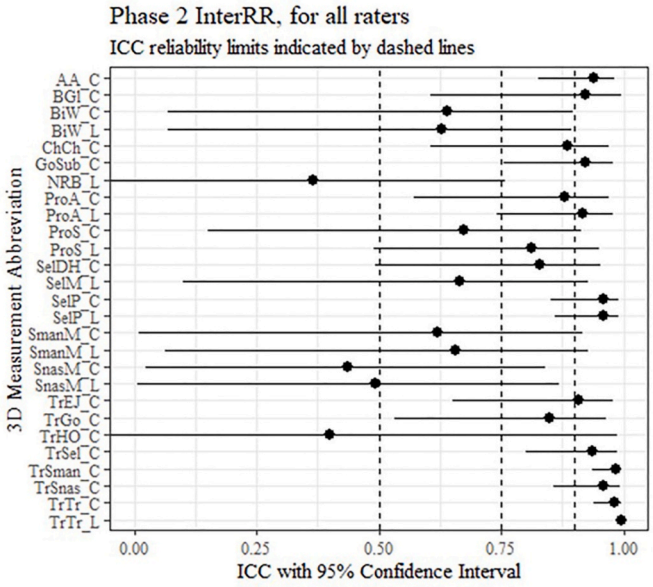


Fig. 5. Phase 2 interRR ICC statistics and 95% confidence interval, across all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.50–0.75, good 0.75–0.90, and excellent >0.90).

each rater's ability to collect sets of consistent measurement data. Similarly, previous anthropometric researchers have attributed improved intraRR to multiple training sessions for raters (Androutsos et al., 2020; de Miguel-Etayo et al., 2014).

4.2. Improvements in InterRR

The interRR ICC statistics across all three phases at each 3D measurement location are compared in Figs. 16 and 17. In response to RQ4, ICC scores for inter-rater reliability improved by 42.59% averaging over the three phases of data collection. Between Phases 1 and 2, interRR was improved for 15 (55.56%), remained constant for 7 (25.93%), and decreased for 5 (18.52%) of the 27 3D measurement locations. Between Phases 2 and 3, interRR was improved for 8 (29.62%), remained constant for 14 (50.00%), and decreased for 5 (18.52%) of the 27 3D measurement locations. More improvements in interRR were seen between Phases 1 and 2 than between Phases 2 and 3. Between Phase 2 and Phase 3, there was a) no increase in the overall proportion of ICC statistics indicating excellent interRR, b) an increase in the overall proportion of ICC statistics indicating good interRR, and c) no increase or

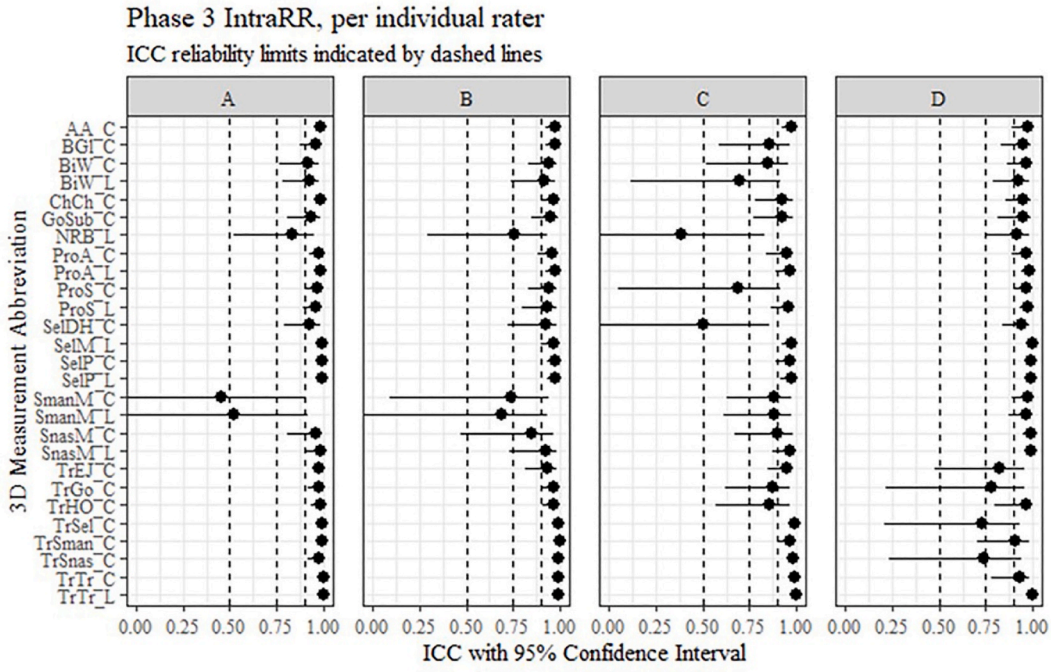


Fig. 6. Phase 3 intraRR ICC statistics and 95% confidence interval, for each individual rater (Rater A, Rater B, Rater C, and Rater D). Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

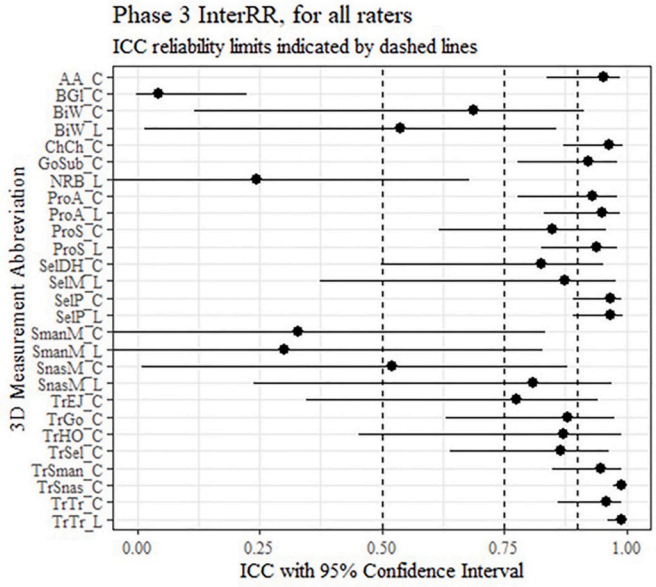


Fig. 7. Phase 3 interRR ICC statistics and 95% confidence interval, across all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.50–0.75, good 0.75–0.90, and excellent >0.90).

decrease in the overall proportion of ICC statistics indicating poor interRR.

4.3. Training and experience

Rater training is a key factor affecting error in anthropometric studies (Bragança et al., 2016; Viviani et al., 2018). In this research, improved ICC statistics over each phase suggest that additional rater experience positively impacted intraRR and interRR. As ICC statistics indicated more dramatic improvements in intraRR between phases than interRR, it is possible that raters gained skill in placing 3D facial landmarks in the same place multiple times, while perhaps not fully matching other raters' landmark placement locations. Stagnant proportions of excellent and poor interRR between Phases 2 and 3 may suggest that additional group training, specifically regarding how to place landmarks in the same way as other raters, was needed to increase interRR in the case of 3D measurement data collection. Previous anthropometric researchers have attributed improved interRR to multiple group training sessions for raters (Androutsos et al., 2020; Bragança et al., 2016; de Miguel-Etayo et al., 2014). Overall, based on the results of this study, the authors suggest that additional rater training and experience with the placement of 3D landmarks resulted in a higher proportion of ICC statistics indicating good to excellent intraRR and interRR for 3D measurement data collection.

4.4. 3D landmark placement

In the post-data collection questionnaire regarding landmark placement difficulty, raters mentioned three specific facial landmarks as difficult to digitally place: the gonion, the zygomatic arches, and the submandibular. The gonion and submandibular points are prominent on the jaw and neck, and raters mentioned that some 3D scan subjects in the sample had body fat occluding the shapes of the jaw and neck. Regarding the reliability and variation of 3D landmark placement, researchers found that jaw landmarks resulted in the lowest reliability and highest variation in placement, due to lack of bony definition around the jaw area for some 3D scan subjects with higher amounts of flesh in that area (Fagertun et al., 2014). The landmarks on the zygomatic arches were also noted as being difficult to place due to lack of bony definition in zygomatic (cheekbone) definition.

Previous researchers have suggested that manually landmarking the participant's face with visual dots (using marker, eyeliner, or small stickers) prior to the 3D scan process allows for increased reliability in 3D gathered measurements, given that data collectors are able to palpate bony landmarks (Aynechi et al., 2011; Bragança et al., 2016; Franco de Sá Gomes et al., 2019; Gibelli et al., 2020; Modabber et al., 2016). Manually-placed landmarks will be visible on the 3D scan, allowing

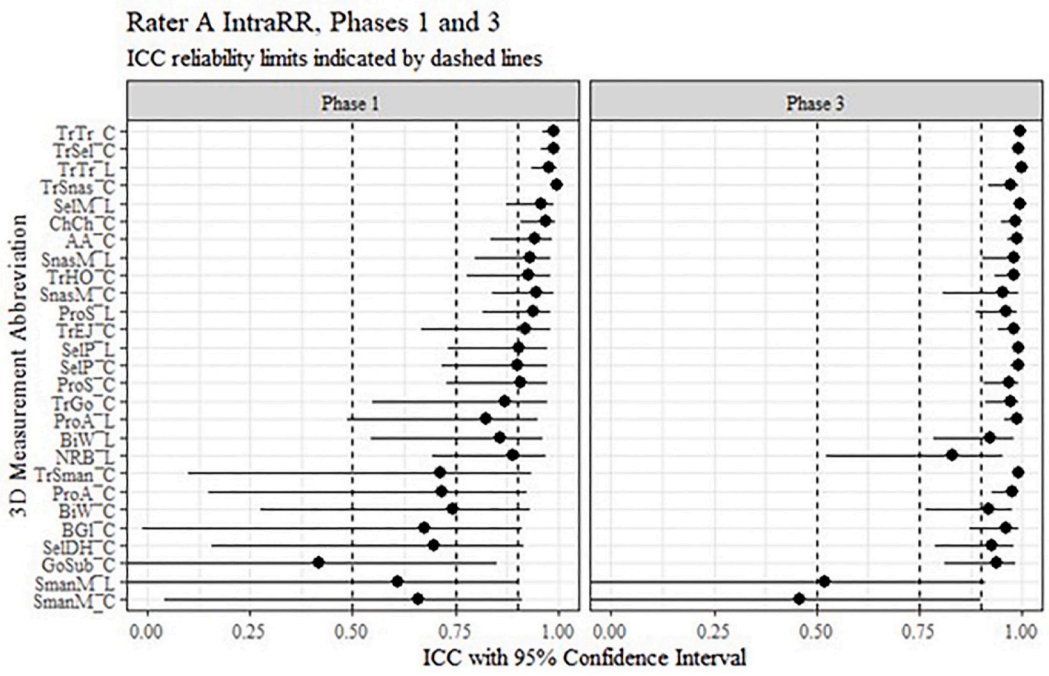


Fig. 8. Phases 1 and 3 ICC statistics and 95% confidence interval for Rater A. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

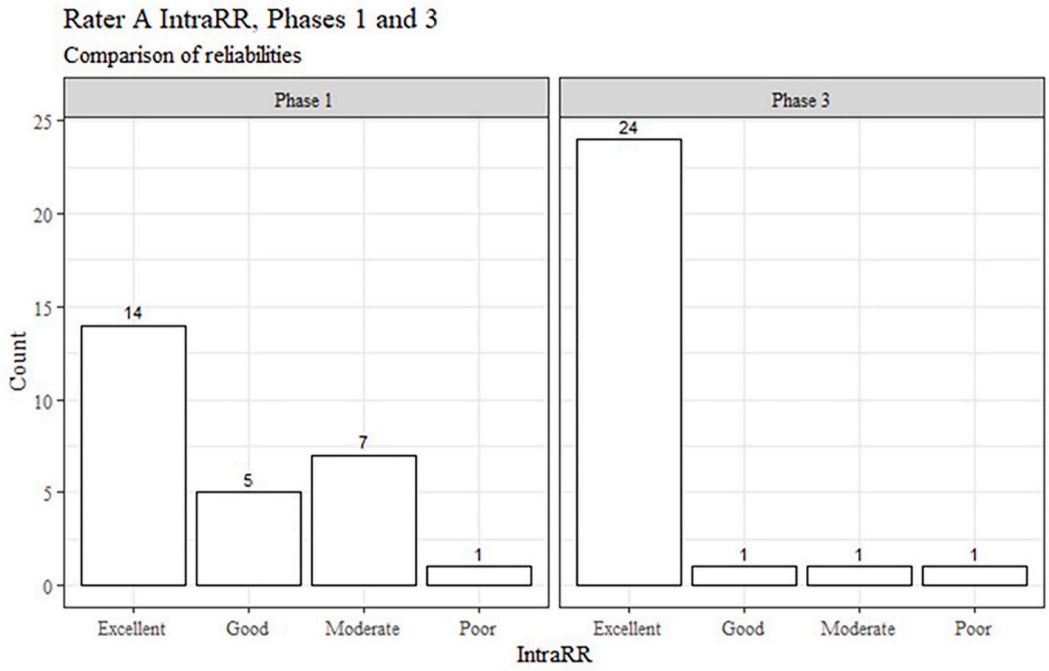


Fig. 9. Comparison of intraRR at Phase 1 and Phase 3 for Rater A. IntraRR was improved for 10 (37.04%), remained constant for 16 (59.26%), and decreased for 1 (3.70%) of 27 3D measurement locations.

collectors of landmark-based 3D measurement data to simply place 3D landmarks on top of the visible manually-indicated landmarks. Modabber et al. (2016) and Ayaz et al. (2020) used manual landmarking in their reliability studies of 3D facial measurements, which they posit contributed to the increased accuracy at the end of their studies.

4.5. Limitations

In the present study, scan subjects did not have landmarks placed manually prior to 3D scanning, thus the researchers do not have information regarding the increase or decrease of reliability as the result of 3D vs. manual landmarking. However, researchers have acknowledged that additional research is needed to assess 3D landmarking-based measurement reliability, as manual landmarking on the face requires

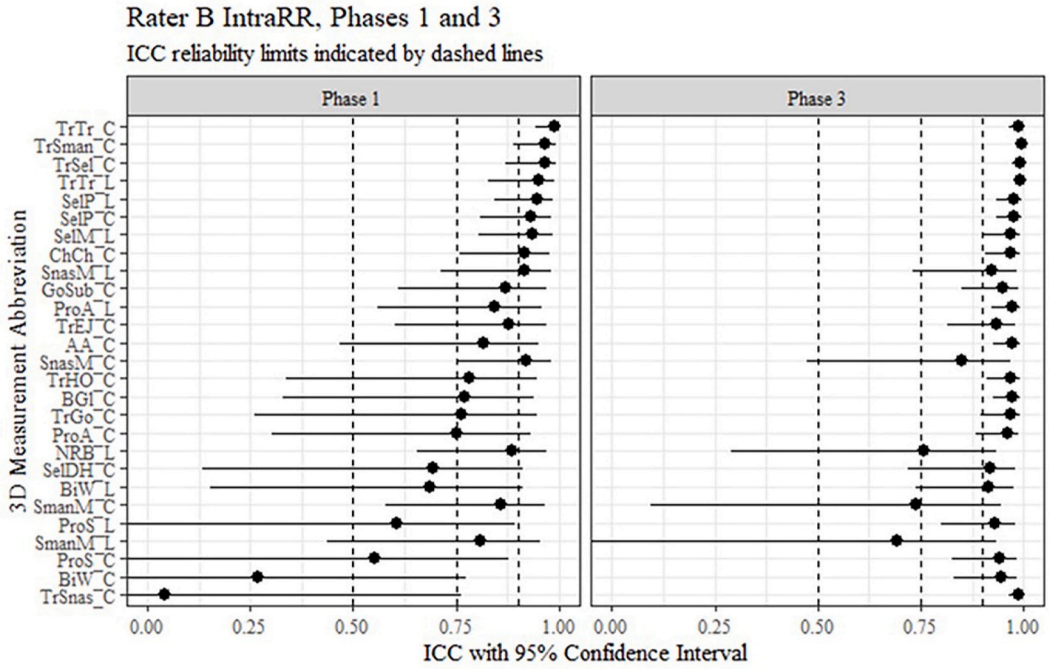


Fig. 10. Phases 1 and 3 ICC statistics and 95% confidence interval for Rater B. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

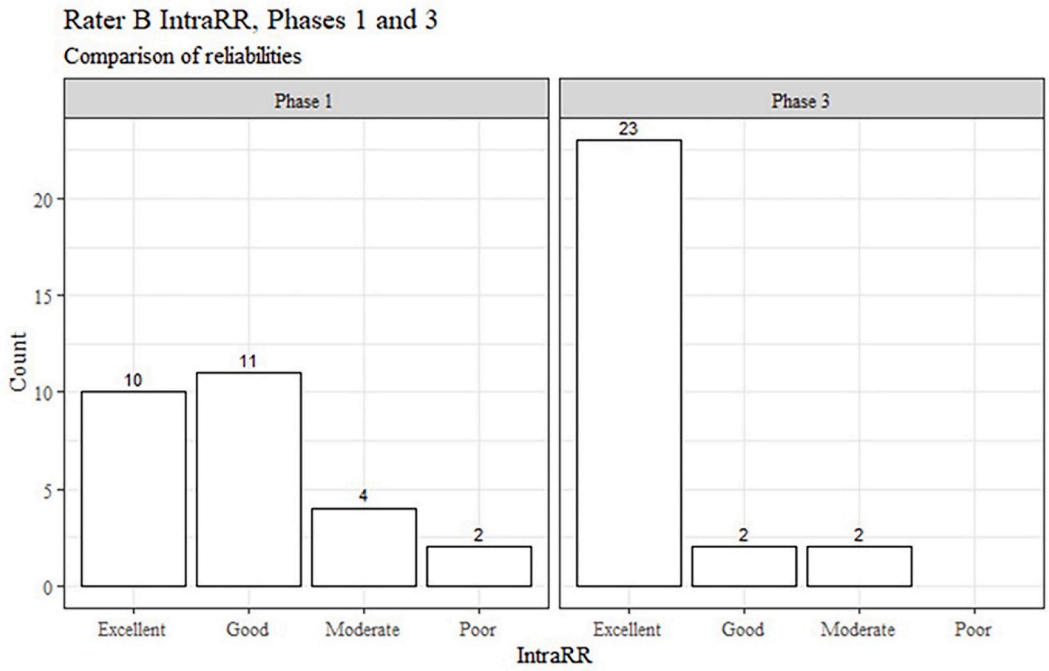


Fig. 11. Comparison of intraRR at Phase 1 and Phase 3 for Rater B. IntraRR was improved for 16 (59.26%), remained constant for 10 (37.04%), and decreased for 1 (3.70%) of 27 3D measurement locations.

more time and closer contact with scan participants than 3D landmarking (Bragança et al., 2016; Franco de Sá Gomes et al., 2019; Gibelli et al., 2020).

It is possible that the reliability assessments of 3D landmark placement (in X, Y, Z axis coordinate measurements) may provide a better understanding of overall 3D measurement reliability. The researchers of the present study used a custom-made measurement wizard in Anthroscan, developed by Human Solutions specifically for this research, which inherently limits the generalizability of the researchers' findings. However, measurement wizards such as the one used in the present study may be necessary for future work that uses 3D landmark placement in place of manual landmarking. Therefore, future research should continually assess the reliability of 3D landmark placement on 3D scans using custom 3D measurement wizards.

Although the 3D measurement data were collected for this research and are therefore original and empirical data, the 3D scans (from which

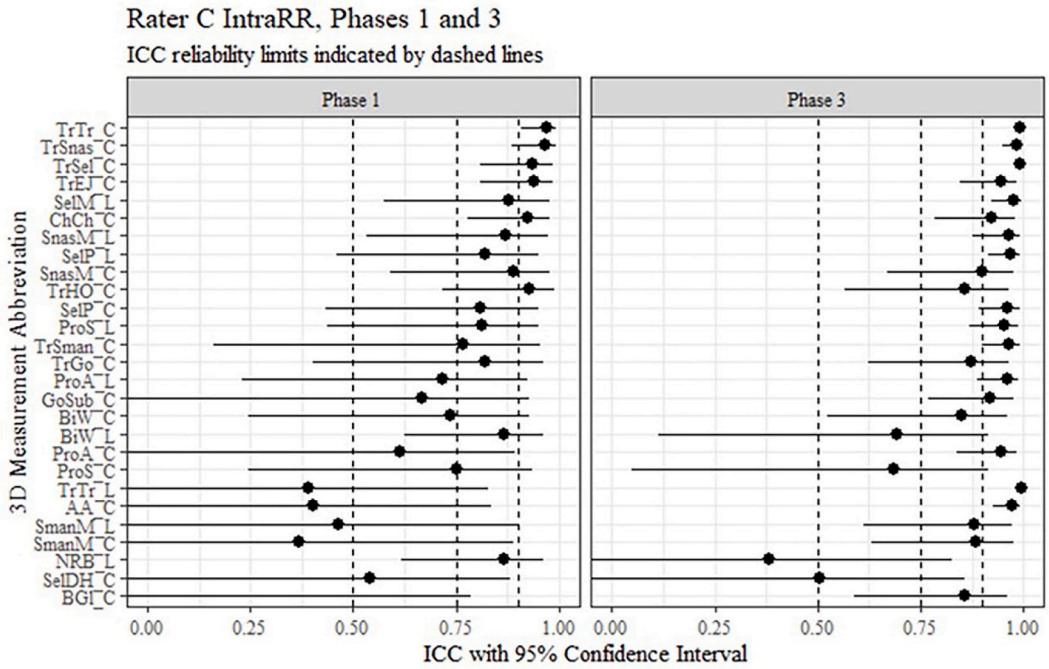


Fig. 12. Phases 1 and 3 ICC statistics and 95% confidence interval for Rater C. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

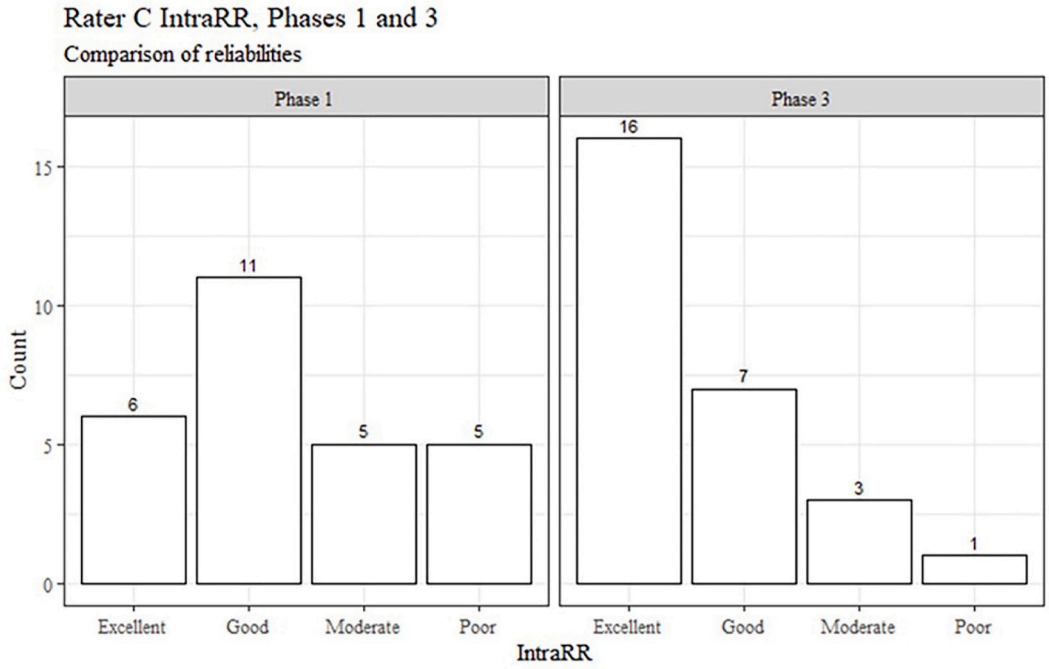


Fig. 13. Comparison of intraRR at Phase 1 and Phase 3 for Rater C. IntraRR was improved for 16 (59.26%), remained constant for 8 (29.63%), and decreased for 3 (11.11%) of 27 3D measurement locations.

3D measurement data were collected) were themselves collected by and purchased from Human Solutions. Therefore, the researchers did not have any activity in assuring the quality of each scan. For example, researchers could not be sure that 3D landmark and/or 3D measurement locations were not occluded, that the scan was oriented properly, and/or that the scan participant had a neutral facial expression. Future researchers should consider collecting 3D scans using staff that they have trained, to allow for better scan quality control.

An additional limitation in the current study was the use of inexperienced raters in placing 3D landmarks on the 3D scans. This limitation was addressed by training the raters and evaluating their performance in each phase of the study. It is recommended that future research incorporate individual and group rater training to improve the accuracy of, and quantify, rater performance.

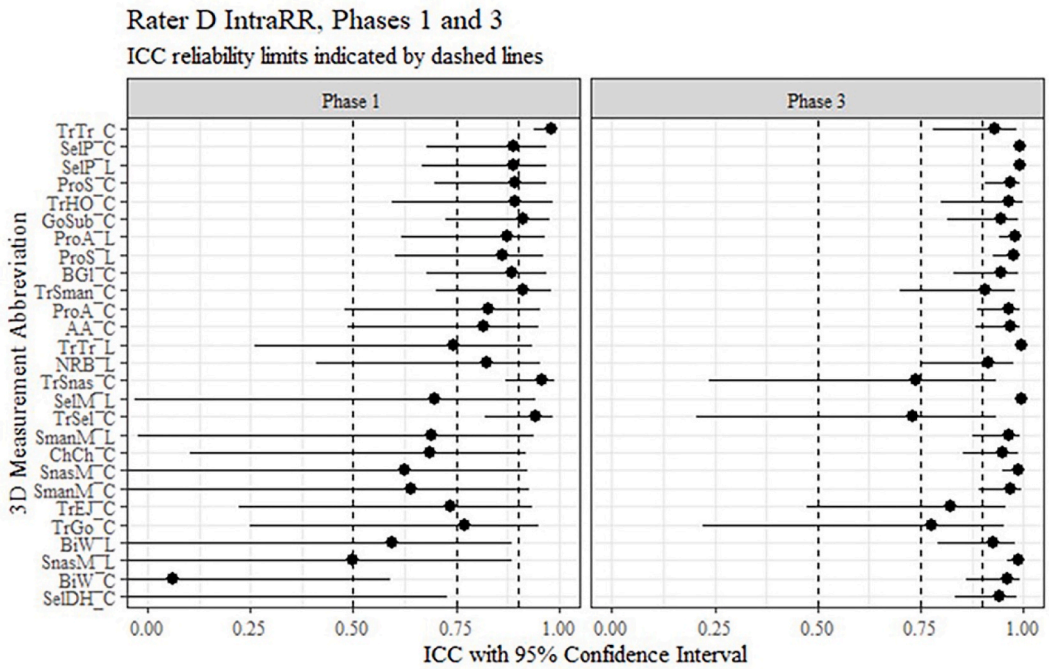


Fig. 14. Phases 1 and 3 ICC statistics and 95% confidence interval for Rater D. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

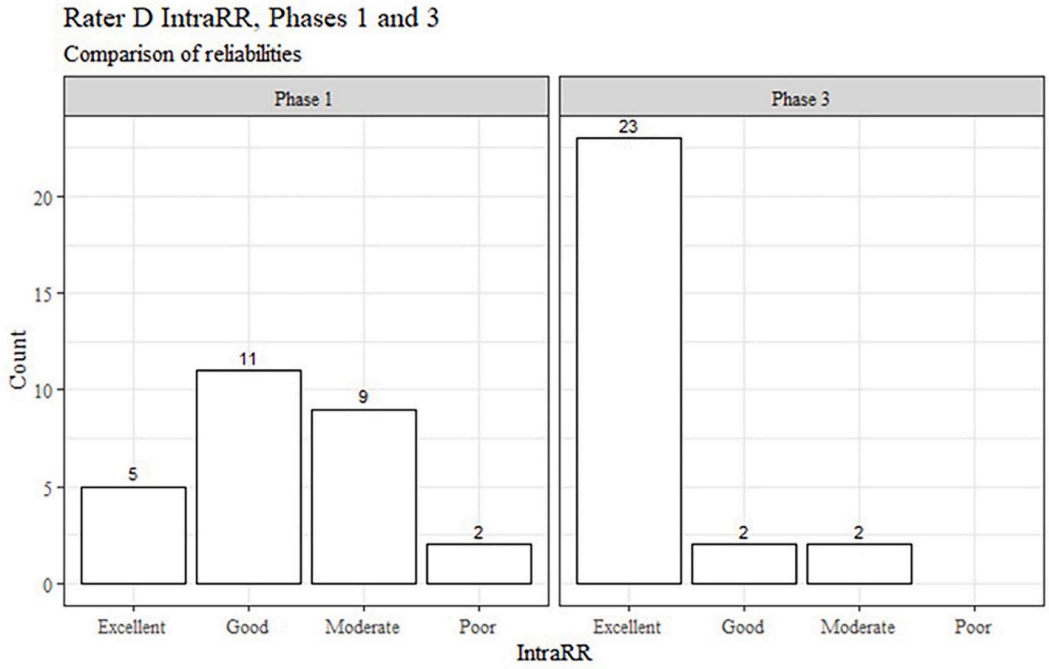


Fig. 15. Comparison of intraRR at Phase 1 and Phase 3 for Rater D. IntraRR was improved for 21 (77.78%), remained constant for 4 (14.81%), and decreased for 2 (7.41%) of 27 3D measurement locations.

5. Conclusions

3D scanning has become popular for use in anthropometric research and industry. Reliable anthropometric data is essential for the ergonomic design of wearable products. As designers, researchers, and practitioners implement 3D scanning in their anthropometric data collection, assessments of reliability allow for a better understanding of the strengths and limitations of 3D scanning tools prior to implementation. In addressing the aim of this study, the researchers assessed intra-rater and inter-rater reliabilities in a three-phase data collection process, where 3D scans were 3D landmarked to collect 27 3D measurements. To date, no other researchers have assessed the reliabilities of multiple raters of 3D measurement data by using 3D landmarking only. Based on the results of this study, the researchers found that the collection of 3D measurement data, by multiple raters and using 3D landmarking methods, yielded a high percentage of ICC statistics in

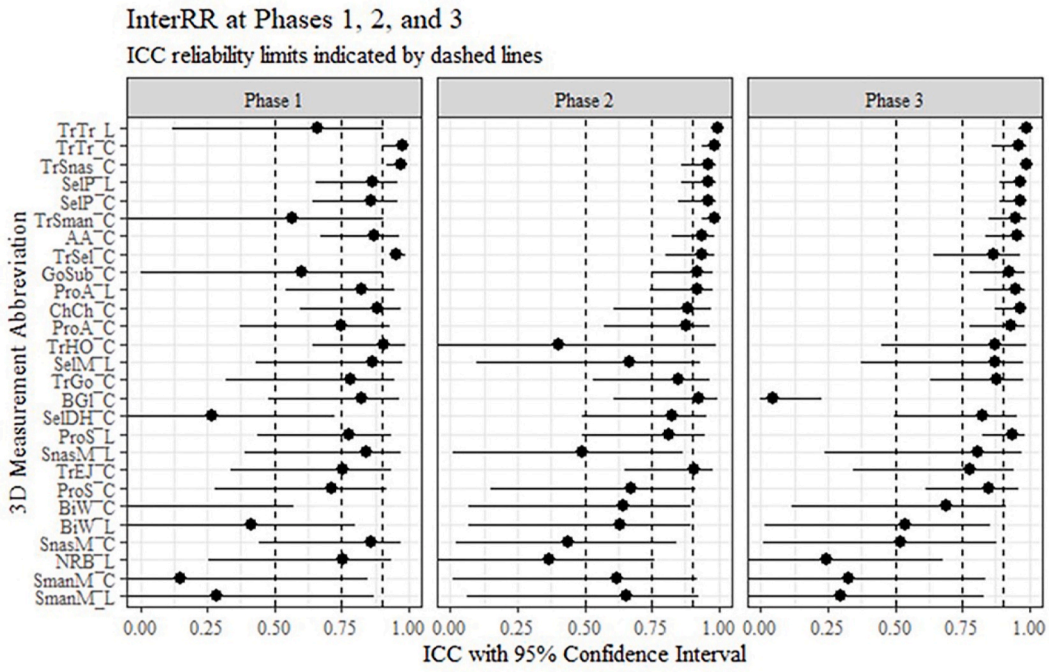


Fig. 16. Phases 1, 2, and 3 ICC statistics and 95% confidence interval for interRR across all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor <0.50, moderate 0.5–0.75, good 0.75–0.90, and excellent >0.90).

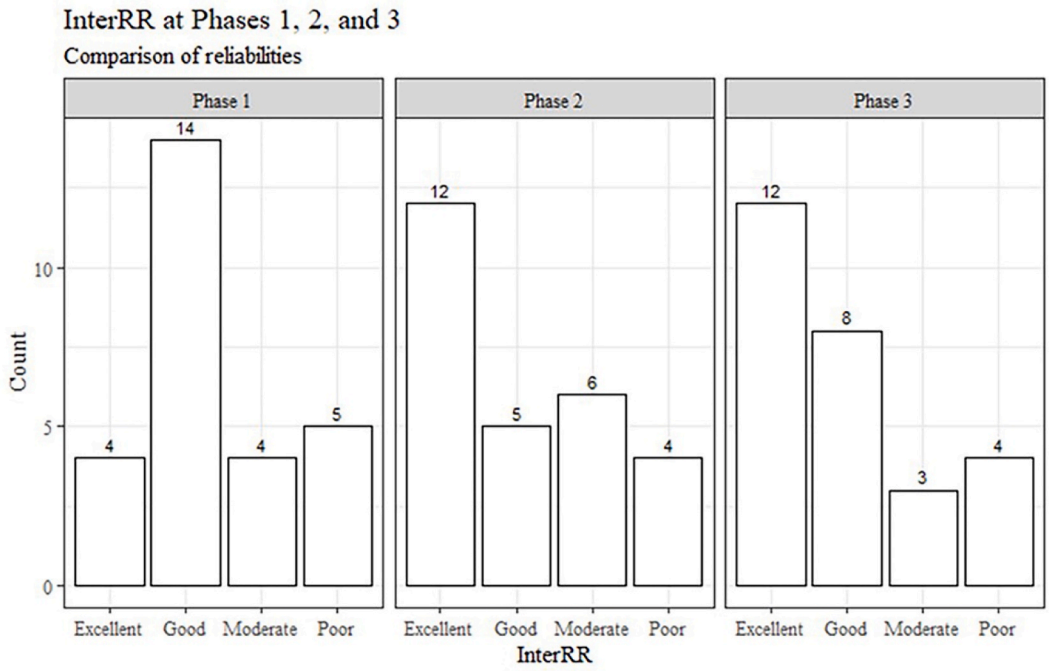


Fig. 17. Comparison of interRR at Phases 1, 2, and 3 across all raters.

the good to excellent (>0.75 ICC) reliability range. Rater training and experience were important considerations in improving intra- and inter-rater reliabilities. Future research is needed to continually assess the reliability of 3D landmarking, 3D measurement data collection, and 3D scanning tools for anthropometric surveying.

Funding

This work was supported by the NIOSH Mountain and Plains Education and Research Center, Centers for Disease Control and Prevention, Grant T42OH009229.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly in order to check the grammar of writing. After using this tool/service, the authors reviewed and edited the content as needed and take full

responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Androutsos, O., Anastasiou, C., Lambrinou, C.-P., Mavrogianni, C., Cardon, G., Van Stappen, V., Kivelä, J., Wikström, K., Moreno, L.A., Iotova, V., Tsochev, K., Chakarova, N., Ungvári, T., Jancso, Z., Makrilakis, K., Manios, Y., 2020. Intra- and inter- observer reliability of anthropometric measurements and blood pressure in primary schoolchildren and adults: the Feel4Diabetes-study. BMC Endocr. Disord. 20 (S1), 27. https://doi.org/10.1186/s12902-020-0501-1.

Artec3D. (n.d.). Artec Eva. Retrieved April 3, 2022, from https://www.artec3d.com/portable-3d-scanners/artec-eva.

Ayaz, I., Shaheen, E., Aly, M., Shujaat, S., Gallo, G., Coucke, W., Politis, C., Jacobs, R., 2020. Accuracy and reliability of 2-dimensional photography versus 3-dimensional soft tissue imaging. Imaging Science in Dentistry 50 (1), 15. https://doi.org/10.5624/isd.2020.50.1.15.

Aynechi, N., Larson, B.E., Leon-Salazar, V., Beiraghi, S., 2011. Accuracy and precision of a 3D anthropometric facial analysis with and without landmark labeling before image acquisition. Angle Orthod. 81 (2), 245–252. https://doi.org/10.2319/041810-210.1.

Bailar, J.C., Meyer, E.A., Pool, R., 2007. Assessment of the NIOSH head-and-face anthropometric survey of U.S. respirator users. In: Assessment of the NIOSH Head-And-Face Anthropometric Survey of U.S. Respirator Users. National Academies Press. https://doi.org/10.17226/11815.

Ban, K., Jung, E.S., 2020. Ear shape categorization for ergonomic product design. Int. J. Ind. Ergon. 80, 102962 https://doi.org/10.1016/j.ergon.2020.102962.

benAzouz, Z., Rioux, M., Shu, C., Lepage, R., 2006. Characterizing human shape variation using 3D anthropometric data. Vis. Comput. 22 (5), 302–314. https://doi.org/10.1007/S00371-006-0006-6, 2006 22:5.

Berkelaar, M., 2022. lpSolve: Interface to Lp_solve V. 5.5 to Solve Linear/Integer Programs. https://github.com/gaborcsardi/lpSolve.

Bonin, D., Radke, D., Wischniewski, S., 2019. Gathering 3D body surface scans and anthropometric data as part of an epidemiological health study – method and results. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (Eds.), Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Springer International Publishing, pp. 128–140.

Bragança, S., Arezes, P., Carvalho, M., Ashdown, S.P., 2016. Current state of the art and enduring issues in anthropometric data collection/Estado actual de la técnica y cuestiones perdurables en la recogida de datos antropométricos. Dyna 83 (197), 22–30. https://doi.org/10.15446/dyna.v83n197.57586.

Clauser, C., Tebbetts, I., Bredtmiller, B., McConville, J., Gordon, C.C., 1988. Measurer's Handbook: U.S. Army Anthropometric Survey 1987-1988.

Coblentz, A., Mollard, R., Ignazi, G., 1991. Three-dimensional face shape analysis of French adults, and its application to the design of protective equipment. Ergonomics 34 (4), 497–517. https://doi.org/10.1080/00140139108967332.

Coward, T.J., Watson, R.M., Scott, B.J.J., 1997. Laser scanning for the identification of repeatable landmarks of the ears and face. Br. J. Plast. Surg. 50 (5), 308–314. https://doi.org/10.1016/S0007-1226(97)90538-5.

de Miguel-Etayo, P., Gracia-Marco, L., Ortega, F.B., Intemann, T., Foraita, R., Lissner, L., Oja, L., Barba, G., Michels, N., Tornaritis, M., Molnár, D., Pitsiladis, Y., Ahrens, W., Moreno, L.A., 2014. Physical fitness reference standards in European children: the IDEFICS study. Int. J. Obes. 38 (S2), S57–S66. https://doi.org/10.1038/ijo.2014.136.

Düppe, K., Becker, M., Schönmeyr, B., 2018. Evaluation of facial anthropometry using three-dimensional photogrammetry and direct measuring techniques. J. Craniofac. Surg. 29 (5), 1245–1251. https://doi.org/10.1097/SCS.0000000000004580.

Fagertun, J., Harder, S., Rosengren, A., Moeller, C., Werge, T., Paulsen, R.R., Hansen, T. F., 2014. 3D facial landmarks: inter-operator variability of manual annotation. BMC Med. Imag. 14. http://www.biomedcentral.com/1471-2342/14/35.

Franco de Sá Gomes, C., Libdy, M.R., Normando, D., 2019. Scan time, reliability and accuracy of craniofacial measurements using a 3D light scanner. Journal of Oral Biology and Craniofacial Research 9 (4), 331–335. https://doi.org/10.1016/j.jobcr.2019.07.001.

Gamer, M., Lemon, J., Ian Fellows Puspendra Singh, 2019. Irr: Various Coefficients of Interrater Reliability and Agreement. https://www.r-project.org.

Gibelli, D., Dolci, C., Cappella, A., Sforza, C., 2020. Reliability of optical devices for three-dimensional facial anatomy description: a systematic review and meta-analysis. Int. J. Oral Maxillofac. Surg. 49 (8), 1092–1106. https://doi.org/10.1016/j.ijom.2019.10.019.

Gordon, C.C., Blackwell, C.L., Bradtmiller, B., Parham, J.L., Barrientos, P., Paquette, S.P., Corner, B.D., Carson, J.M., Venezia, J.C., Rockwell, B.M., Mucher, M., Kristensen, S., 2014. 2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics.

Gordon, C.C., Churchill, T., Clauser, C.E., Bradtmiller, B., McConville, J.T., Tebbetts, I., Walker, R.A., 1989. Anthropometric Survey of US Army Personnel: Methods and Summary Statistics 1988.

Goto, L., Lee, W., Molenbroek, J.F.M., Cabo, A.J., Goossens, R.H.M., 2019. Traditional and 3D scan extracted measurements of the heads and faces of Dutch children. Int. J. Ind. Ergon. 73, 102828 https://doi.org/10.1016/j.ergon.2019.102828.

Human Solutions. About human Solutions. https://www.human-solutions.com/en/about-human-solutions/about-us/index.html. (Accessed 8 May 2022).

Kim, A.J., Gu, D., Chandiramani, R., Linjawi, I., Deutsch, I.C.K., Allareddy, V., Masoud, M.I., 2018. Accuracy and reliability of digital craniofacial measurements using a small-format, handheld 3D camera. Orthod. Craniofac. Res. 21 (3), 132–139. https://doi.org/10.1111/ocr.12228.

Koo, T.K., Li, M.Y., 2016. Cracking the code: providing insight into the fundamentals of research and evidence-based practice A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kouchi, M., Mochimaru, M., 2004. Analysis of 3D face forms for proper sizing and CAD of spectacle frames. Ergonomics 47 (14), 1499–1516. https://doi.org/10.1080/00140130412331290907.

Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A., Scholz, M., 2016. Reliability of 3D laser-based anthropometry and comparison with classical anthropometry. Sci. Rep. 6 https://doi.org/10.1038/srep26672.

Lacko, D., Huysmans, T., Parizel, P.M., De Bruyne, G., Verwulgen, S., Van Hulle, M.M., Sijbers, J., 2015. Evaluation of an anthropometric shape model of the human scalp. Appl. Ergon. 48, 70–85. https://doi.org/10.1016/j.apergo.2014.11.008.

Lacko, D., Vleugels, J., Fransen, E., Huysmans, T., De Bruyne, G., Van Hulle, M.M., Sijbers, J., Verwulgen, S., 2017. Ergonomic design of an EEG headset using 3D anthropometry. Appl. Ergon. 58, 128–136. https://doi.org/10.1016/j.apergo.2016.06.002.

Lee, W., Jeong, J., Park, J., Jeon, E., Kim, H., Jung, D., Park, S., You, H., 2013. Analysis of the facial measurements of Korean Air Force pilots for oxygen mask design. Ergonomics 56 (9), 1451–1464. https://doi.org/10.1080/00140139.2013.816376.

Lee, W., Yang, X., Jung, H., Bok, I., Kim, C., Kwon, O., You, H., 2018. Anthropometric analysis of 3D ear scans of Koreans and Caucasians for ear product design. Ergonomics 61 (11), 1480–1495. https://doi.org/10.1080/00140139.2018.1493150.

Liljequist, D., Elfving, B., Roaldsen, K.S., 2019. Intraclass correlation – a discussion and demonstration of basic features. PLoS One 14 (7). https://doi.org/10.1371/JOURNAL.PONE.0219854.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

Modabber, A., Peters, F., Kniha, K., Goloborodko, E., Ghassemi, A., Lethaus, B., Christian, S., 2016. Evaluation of the Accuracy of a Mobile and a Stationary System for Three-Dimensional Facial Scanning. https://doi.org/10.1016/j.jcms.2016.08.008.

Pang, T.Y., Lo, T.S.T., Ellena, T., Mustafa, H., Babalija, J., Subic, A., 2018. Fit, stability and comfort assessment of custom-fitted bicycle helmet inner liner designs, based on 3D anthropometric data. Appl. Ergon. 68, 240–248. https://doi.org/10.1016/j.apergo.2017.12.002.

Park, B.-K.D., Corner, B.D., Hudson, J.A., Whitestone, J., Mullenger, C.R., Reed, M.P., 2021. A three-dimensional parametric adult head model with representation of scalp shape variability under hair. Appl. Ergon. 90, 103239 https://doi.org/10.1016/j.apergo.2020.103239.

R Core Team, 2022. R: the R Project for Statistical Computing. https://www.r-project.org/.

Skals, S., Ellena, T., Subic, A., Mustafa, H., Pang, T.Y., 2016. Improving fit of bicycle helmet liners using 3D anthropometric data. Int. J. Ind. Ergon. 55, 86–95. https://doi.org/10.1016/j.ergon.2016.08.009.

Srinivasan, A., Balamurugan, V., 2014. Occlusion detection and image restoration in 3D face image. In: TENCON 2014 - 2014 IEEE Region 10 Conference, pp. 1–6. https://doi.org/10.1109/TENCON.2014.7022477.

Taylor, P.J., 2009. 1 an Introduction to Intraclass Correlation that Resolves Some Common Confusions.

Traumann, A., Peets, T., Dabolina, I., Lapkovska, E., 2019. Analysis of 3-d body measurements to determine trousers sizes of military combat clothing. Textile and Leather Review 2 (1), 6–14. https://doi.org/10.31881/TLR.2019.2.

Viviani, C., Arezes, P.M., Bragança, S., Molenbroek, J., Dianat, I., Castellucci, H.I., 2018. Accuracy, precision and reliability in anthropometric surveys for ergonomics purposes in adult working populations: a literature review. Int. J. Ind. Ergon. 65, 1–16. https://doi.org/10.1016/j.ergon.2018.01.012.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., et al., 2019. Welcome to the tidyverse. J. Open Source Softw. 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Wong, J.Y., Oh, A.K., Ohta, E., Hunt, A.T., Rogers, G.F., Mulliken, J.B., Deutsch, C.K., 2008. Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. Cleft Palate-Craniofacial J. 45 (3), 232–239. https://doi.org/10.1597/06-175.

Zhuang, Z., Bradtmiller, B., 2005. Head-and-Face anthropometric survey of U.S. Respirator users. J. Occup. Environ. Hyg. 2 (11), 567–576. https://doi.org/10.1080/15459620500324727.

Zhuang, Z., Bradtmiller, B., Shaffer, R.E., 2007. New respirator fit test panels representing the current U.S. civilian work force. J. Occup. Environ. Hyg. 4 (9), 647–659. https://doi.org/10.1080/15459620701497538.

Zhuang, Z., Coffey, C.C., Berry Ann, R., 2005. The effect of subject characteristics and respirator features on respirator fit. J. Occup. Environ. Hyg. 2, 641–649. https://doi.org/10.1080/15459620500391668.

Zhuang, Z., Groce, D., Ahlers, H.W., Iskander, W., Landsittel, D., Guffey, S., Benson, S., Viscusi, D., Shaffer, R.E., 2008. Correlation between Respirator Fit and Respirator Fit Test Panel Cells by Respirator Size, vol. 5, pp. 617–628. https://doi.org/10.1080/15459620802293810, 10.

Zhuang, Z., Landsittel, D., Benson, S., Roberge, R., Shaffer, R., 2010. Facial anthropometric differences among gender, ethnicity, and age groups. Ann. Occup. Hyg. 54 (4), 391–402. https://doi.org/10.1093/annhyg/meq007.