THESIS

INTRA-RATER AND INTER-RATER RELIABILITY OF 3D FACIAL MEASUREMENTS

Submitted by

Isabel Rosalene Olmedo-Nockideneh

Department of Environmental Health and Radiological Sciences

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2023

Master's Committee:

Advisor: John Rosecrance

William Brazile
Margaret Gutilla

ABSTRACT

INTRA-RATER AND INTER-RATER RELIABILITY OF 3D FACIAL MEASUREMENTS

The purpose of this study was to assess the intra- and inter-rater reliability of a three-dimensional (3D) measurement system for determining the facial anthropometrics of 2,000 individuals. Intra-rater reliability is the degree of agreement among repeated administrations of a measurement system performed by a single rater and inter-rater reliability is the degree of agreement among independent raters who rate, code, or assess the same phenomenon using the same measurement system. Research studies that require the use of novel measurement systems by several raters must first establish that the phenomenon being measured have acceptable levels of both intra- and inter-rater reliability. Assessments of novel measurement systems are useful in refining the measurement tools given to raters by determining if a particular scale is appropriate for measuring a specific phenomenon.

The present study was one facet of a larger anthropometric study of 2,000 facial scans, which investigates the demographic variables that may account for differences in facial size and shape. For this reliability study, a random sample of 30 facial scans was hand-digitized by four coders. The randomized sample was used to assess the intra- and inter-rater reliability of 28 facial anthropometric landmarks. The intraclass correlation coefficient was used to assess rater reliability. The results of the study indicated generally good inter-rater reliability and a steady improvement in both intra- and inter-rater reliability with greater experience.

There are no conflicts of interest or a current funding source regarding this study. This study will impact environmental and occupational health by contributing a reliability study to my colleagues.

TABLE OF CONTENTS

CHAPTER 1 – INTRODUCTION

Background

Anthropometric research requires time, personnel, and other resources to conduct manual measurements of body and facial dimensions. Anthropometrics is a field that has many applications which expand with the development of novel indirect anthropometric measurement systems, such as the Anthroscan software developed by Human Solutions (Anthroscan for Human Solutions, n.d.). There is a consensus that these new technologies need to be evaluated for their level of intra-rater and inter-rater reliability, but researchers must be willing to collect facial anthropometric data using indirect methods to advance the field.

This reliability study was the first phase of a larger study on three-dimensional (3D) facial anthropometrics. Thus, the purpose of this phase of the study was to assess the intra-rater (intraRR) and inter-rater (interRR) reliability of coders who will digitize 3D facial scans for determining the facial anthropometric measures of 2,000 individuals. The results of the study determined the level of agreement within and between coders that digitize facial anthropometrics. This reliability study is the first phase of a larger research program comparing facial anthropometrics among workers of different ethnicity, race, sex, and age.

Specific Aims

Specific Aim 1

To assess the intraRR of coders that measure 3D facial anthropometric variables.

Objectives:

1. Quantify the degree of agreement for intraRR among coders that digitize facial anthropometrics immediately after the coder training (Time 1) and after each coder has digitized 100 scans (Time 3). The intraclass correlation coefficient statistic (ICC) was used to assess agreement. I hypothesized that the ICC would fall in the moderate range of Koo's ICC reliability categories (Koo, TK. & Li, MY., 2016).
2. If the ICCs for intraRR are below the moderate range, I identified and corrected deficiencies in the training process and the measurement techniques that may have contributed to the lack of agreement between coders.

Specific Aim 2

To assess the interRR of coders that measure 3D facial anthropometric variables.

Objectives:

1. Quantify the agreement for interRR among coders immediately after coder training (Time 1), after 30 facial scans (Time 2), and after 100 facial scans (Time 3). The ICC statistic was used to assess levels of agreement. I hypothesized that the ICC would remain the same or increase slightly with experience.

CHAPTER 2 – LITERATURE REVIEW

3D body scanning is becoming more appealing as its applications expand across different industries. Anthropometrics became a recognized discipline in the 1950s after hundreds of years of curiosity surrounding the difference in measurable characteristics from person to person (Simmons, KP. & Istook, CL., 2003). Anthropometrics, the dimensions that describe the human body, has applications in human factors engineering, ergonomics, vehicle and apparel design, and medicine. A 3D body scanner allows body and facial measurements to be taken without coming into contact with the participant and numerous data types can be determined with 3D laser-based optical triangulation to create a “virtual twin” (Kuehnapfel, A., Ahnert, P., Loeffler, M., & Scholz, M., 2017). 3D scanning technology is capable of linear measurements, body shapes and angles, and relational data points, allowing something as variable as the face to be subject to customized apparel and equipment. With such wide applications, 3D scanners must be assessed for reliability and consistency as intra- and inter-rater reliability of 3D facial anthropometric measurements has not been previously assessed (Simmons, KP. & Istook, CL., 2003). Any measure or instrument used in research must exhibit some degree of reliability. Reliability allows subjects or events to be categorized regardless of measurement errors present (Wolan-Nieroda, A., Guzik, A., Mocur, P., Drużbicki, M., & Maciejczak, A., 2020) and should be tested when a new measurement scale or instrument arises. When reliability is tested repeatedly amongst different coders, a scale’s stability and consistency can be evaluated across time and with different samples, ensuring the data collected is significant and accurate (Yang, Y., Wang, Y., Zhou, Y., Chen, C., & Xing, D., 2016). Research studies that require the use of novel measurement systems by several raters must first establish that the phenomenon being measured has acceptable levels of both intraRR and interRR. Assessments of reliability are useful for refining the measurement tools given to raters by determining if a particular scale is appropriate for measuring a particular variable.

Intra-rater reliability (intraRR) is the degree of agreement among repeated administrations of a measurement system performed by a single rater. Inter-rater reliability (interRR) is the degree of agreement among independent raters who rate, code, or assess the same phenomenon using the same measurement system. An assessment of interRR allows researchers to quantify the degree of agreement among two or more coders who make individual ratings about characteristics within a sample population. The true score of any scale or instrument is the variance of observed scores (VarX) and measurement errors (VarE) made by coders. Measurement errors (E) that affect the reliability in any study may arise due to inaccuracy, imprecision, poor item scaling, instability of the measurement over time, and instability of the measurements made between coders. If errors are present, they can negatively affect ICC estimates. Training protocols can be rectified or redeveloped to enable researchers to reduce the frequency of measurement deficiencies that may be affecting the coder, and this could increase the level of agreement within and between coders by reducing measurement errors (E) (Hallgren, KA., 2012).

Craniofacial anthropometry was performed using 3D digital photogrammetry (3DDP) and evaluated for reliability in 2008 (Wong, et al., 2008). 3DDP, or 3D stereo photogrammetry (3DSP), has an advantage over other 3D scanning methods, such as laser scanning and computer-assisted facial tomography. It was a novel method in 2008 and utilized the synchronization of digital cameras to procure multiple angles that would create a 3D image. Craniofacial anthropometry is commonly used for surgical treatment planning, monitoring surgical outcomes, and assessing change in the face over time. 3DDP is often used instead of a direct anthropometric approach because taking a direct facial measurement requires the patient to stay still for the duration of data collection, the training process is long, and there is a lack of ability to archive surface morphology of the craniofacial region. 3DDP is the favored method of data collection with craniofacial disorders that change over time as they are noninvasive and no longer require patients to stay still for longer than 30 seconds. Data points were collected by coders placing landmarks on the 3D facial scan. A setback for Wong et al.'s study and this study is the difficulty placing landmarks on bony facial structures, such as the zygomatic arch. To assess the performance of digital anthropometry, Wong and their colleagues set out to evaluate the reliability of 3DDP obtained with the 3dMDface system, with a study population of 20 adults. For the study, a single rater was trained to take direct facial anthropometric measurements, and a second single rater placed digital landmarks on the 3D facial scan. Their analysis was performed using SAS software and nine statistical tests for normality and nonnormality. IntraRR and interRR were assessed using the Shapiro-Wilk W test for normality, Pearson's product-moment correlation coefficient for normal data, and Spearman's correlation coefficient was used for nonnormative data. The correlation coefficient tests were found to be statistically significant, with the highest and lowest degree of reliability coming out to 0.99 and 0.66

respectively. The poor reliability could have been due to inconsistent landmark placement on subjects with facial hair and obscured facial scans. The analysis of the 3DDP measurement system showed that digitizing 3D facial scans was reliable and mostly unbiased in relation to the direct facial measurement ratings (Wong, et al., 2008).

Europe’s ToyBox-study (2014) set out to implement an elementary school program to prevent obesity that was based on the collection of direct anthropometric measurements. Direct anthropometrics, such as waist circumference, weight, and height, are generally non-invasive and utilize portable equipment to collect data. The ToyBox study was designed to minimize measurement errors that could arise with trained personnel, or fieldworkers. One fieldworker from each country participating in the ToyBox-study were given standardized anthropometric training and asked to directly measure at least ten children, resulting in 60 samples, or ten from the six participating countries. Intra-rater reliability was assessed using technical error of measurement (TEM) during the first training session for each fieldworker. Inter-rater reliability was assessed using TEM during the second training session for fieldworkers who needed to repeat measurements they had taken during the first training session. Waist circumference had poor intraRR, which was attributed to poor training. Fieldworkers from five of the six participating countries had to attend further training and repeat the measurements they collected during initial data collection, which is where the ToyBox researchers got their interRR data. Higher intraRR was achieved and interRR was assessed during the repeat training session, resulting in high interRR. For the statistical analysis of reliability, the technical error of measurement (TEM) was calculated across the three anthropometric measurements taken. IntraRR was above 0.99 for weight and height and 0.89 for waist circumference with interRR being above 0.98 for all three measurements. Having great agreement for intraRR and interRR, ToyBox researchers were satisfied with the standardization of study protocols they achieved after two training sessions. Their results also showed that TEM can be greatly reduced with protocol standardization and multiple training sessions (De Miguel-Etayo, P., et al., 2014).

The Feel4Diabetes study was a school and community-based intervention program that was designed to prevent type 2 diabetes in European families. The same anthropometric measurements from the ToyBox study and blood pressure data were gathered for the purposes of assessing intraRR and interRR amongst examiners participating in the Feel4Diabetes study. Researchers felt it was important to standardize direct anthropometric and blood pressure measurement protocols and ensure the study was reliable and without bias or other factors that would contribute to increased measurement error. A central training workshop on how to take the four measurements needed for the Feel4Diabetes study was held for six examiners before they were able to take baseline anthropometric measurements and two blood pressure measurements from study participants. The study protocols were the same for the follow-up 1 and follow-up 2 measurements, each examiner took three anthropometric measurements and two blood pressure measurements using the same electronic scale, stadiometer, measurement tape, and electronic blood pressure monitor for all three measurement phases. Reliability was calculated to assess anthropometric measurement intraRR for the three measurement phases of adults and children, which yielded an intraRR of above 0.99 for both study populations. The interRR of anthropometric measurements for children and adults ranged from 0.95-0.99, and the interRR of blood pressure measurements in adults were 0.76 for systolic measurements and 0.91 for diastolic measurements. Standardization of measurement equipment and protocols seemed to have enabled the six examiners to achieve excellent anthropometric reliability, and further reliability studies are needed to understand why there was a wide range of reliability for blood pressure measurements (Androutsos, A. et al., 2020).

Cosmetic surgery often requires surgeons to understand aesthetic facial anthropometrics and how to achieve aesthetically pleasing facial changes without disrupting harmony of the features. In aesthetic rhinoplasty, achieving an aesthetic result can be challenging and because of this, plastic surgeons have developed “Nasofacial Analysis” to plan and assess a patient’s face for the best possible rhinoplasty. Pre- and post-operational desired and real facial anthropometrics are obtained directly or indirectly. Indirect anthropometrics using photogrammetric facial analysis has been recognized as cost-effective, but previous indirect methods have been time consuming or difficult to perform. Meruane et al. (2015) sought to compare indirect anthropometrics to direct anthropometrics using a software developed in 2009 by the Apaydin research group. *Rhinobase®* is a comprehensive rhinoplasty software that allows for the storage and retrieval of patient nasofacial anthropometrics, the navigation of the creative rhinoplasty process, and the ability to run a facial analysis. By replicating Apaydin’s research protocol and comparing indirect and direct anthropometrics, Meruane et al. (2015) were able to assess the intraRR and interRR of each method’s pre- and post-operative nasofacial analysis. Patients involved in the study had their indirect and direct facial anthropometric measurements taken before the rhinoplasty occurred and two surgeons independently placed ten nasofacial landmarks and placed the same landmarks after a 30-day period had passed. This process was repeated six months post-operation. The reliability scores were calculated using the ICC. The ICC estimate for intraRR was greater than 0.90 for 8/10 measurements, 0.79 for intercantal width, and 0.88 for tip projection. The interRR ICC estimate was 0.90 for 8/10 measurements, 0.57 for intercantal width, and 0.81 for tip projection. IntraRR was slightly better than interRR amongst surgeons, but the researchers guessed this could have been due to observer bias, inadequate patient photos, or poor landmark placement in the *Rhinobase* ® software. The study proved that indirect anthropometric methods are reliable for facial analysis and could validate other 3D facial scan measurement systems.

In 2007, a committee from the Institute of Medicine (IOM) came together to assess the National Institute for Occupational Safety and Health's (NIOSH) anthropometric survey for respirator users in the United States (US). This work is important because millions of US workers use respirators as part of their personal protective equipment (PPE) at work. Respirators protect wearers from a slew of respiratory hazards such as airborne pathogens and silica, so it is important that a respirator properly fits and seals on the users face. To assess respirator performance, NIOSH conducted fit-test panels that were comprised of an anthropometric survey on a study population that represented the workforce that wears respirators. NIOSH contracted Anthrotech, Inc. to collect direct facial anthropometric measurements with calipers and measurement tapes. They collected a total of 18 head and facial measurements from over 4,000 subjects. Although Anthrotech made an effort to reduce human error, the IOM committee recommended that additional measurement error analyses should be conducted to assess agreement within and between measurement observers, or the field technicians gathering direct anthropometric measurements. Anthrotech researchers also gathered approximately 1,000 3D facials scans to create an indirect measurement method dataset, but NIOSH stated that there were discrepancies with the 3D scans that excluded them from the fit-test panel. The IOM committee recommended that NIOSH should collect and utilize 3D facial scan data in conjunction with direct anthropometric methods. This would allow for the assessment of agreement between observers creating indirect/direct datasets (Institute of Medicine, 2007).

In a study done by de Sá Gomes et al. (2019), the researchers sought to understand the reliability of a 3D light scanner compared to direct craniofacial anthropometry. Six females and nine males each had their faces scanned with the Artec Eva 3D light scanner. Facial scans were collected with and without direct craniofacial reference points, and direct measurements were taken using a digital caliper. Eleven linear measurements were collected from each measurement method, and intra- and inter-rater reliability was assessed using the intraclass correlation coefficient (ICC). The reliability range used in this study is as follows: poor: ICC < 0.4, medium to good: 0.4 ≤ ICC < 0.75, and excellent: ICC ≥ 0.75. Results from this study demonstrated 72% of the linear measurements falling into the excellent range for the Artec Eva 3D measurement system, which was the same percent agreement for the intra-rater reliability of direct measurements. For inter-rater reliability between Artec Eva 3D facial scans collected with and without points, 55% of the linear measurements had excellent reliability linear measurements collected without a reference point. 100% of the linear measurements collected with a reference point fell into the excellent reliability range. The study suggested that the Artec Eva is a reliable 3D measurement system and that accuracy of 3D data collection increases with the presence of craniofacial reference points on the subject to be scanned.

Ayaz et al. (2020) set out to evaluate the reliability and accuracy of 2D facial images and two types of 3D measurement systems. To achieve this, 2D facial images were taken with a professional camera and 3D scans were collected via laser scanning (Planmeca ProFace) and stereophotogrammetry (3D Vectra H1) on 50 Caucasian subjects. Two examiners placed 22 facial anthropometric landmarks on the subjects' faces and collected linear measurements at two different times, and the same was done to collect linear and angular measurements from the 3D facial scans. Intra- and inter-rater reliability was assessed using the intraclass correlation coefficient, and its range of reliability was <0.40 for poor reliability, 0.40-0.59 for fair, 0.60-0.74 for good, and 0.75-1.00 for excellent reliability. Inter-rater reliability was 0.96-0.99 (excellent reliability) for both direct and indirect measurements. Intra-rater reliability was above 0.99 for both measurement methods as well. Measurements collected via 2D facial imaging had the highest total combined error compared to the linear measurements collected from direct anthropometry. Of the three indirect measurement systems conducted, stereophotogrammetry had the highest accuracy. Facial scans collected with laser scanning were distorted, which could have been due to the subjects blinking and moving their head while being scanned. Overall, stereophotogrammetry had the highest amount of precision and reliability and was the most comparable to the linear measurements collected from direct anthropometric landmark placement.

In 2016, researchers voiced their concerns about the accuracy of 3D scanners when scanning a geometric subject, such as the human face. 3D scanners have applications in cosmetic surgery, where high reliability and accuracy is needed. To assess the accuracy of 3D scanners with applications in healthcare, Modabber et al. (2016) wanted to compare FaceScan3D and Artec Eva. Forty-one subjects were scanned by two examiners, one with Artec Eva and the other with FaceScan3D, with legos on their forehead and right cheek for measurement error assessment. After processing the data, Artec Eva had a mean error of 0.228 for the forehead lego and 0.241 for the right cheek lego. FaceScan3D had a mean error of 0.523 for the forehead lego and 0.630 for the right cheek lego. Overall mean error and measured deviations were much lower with Artec Eva. Although Artec Eva had significantly more accuracy than FaceScan3D, Artec Eva is a mobile scanner, and the accuracy of the data may have been lost due to the fact that the mobile scanner collects many 3D pictures to produce a single 3D facial scan.

Traditional anthropometric research studies require significant resources, which were primarily associated with manual measurements of body dimensions. With the development and application of novel indirect anthropometric measurement systems (Simmons, KP. & Istook, CL., 2003), researchers understood the need for reliability assessments. Without the evaluation of the level of intra- and inter-rater reliability of new measurement systems, researchers cannot be confident that their data represents the phenomenon that is being measured. Thus, it is imperative that intra- and inter-rater variability be assessed with measurement systems that employ raters to assess measurements.

CHAPTER 3 – METHODS

Data Source

This study utilized a secondary dataset of approximately 2,000 3D facial scans that was acquired from Human Solutions (Human Solutions, n.d.), a company based out of North Carolina. A randomized sample of 30 facial scans was pulled from the entire dataset for Anthroscan coders to digitize over the duration of this study. Of the 30 facial scans, 53% of the dataset was female, and 47% self-reported as male. The female population was 7% Asian/Asian American, 23% Black, African, or African American, and 23% White/Caucasian. The male population was 3% Asian/Asian American, 20% Black, African, or African American, and 20% White/Causian. Of the 14 males that were apart of the smaller dataset utilized in this study, one participant did not self-report their race/ethnicity. Members of the Latin/Hispanic, American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander, and “Other” race/ethnicities are not represented by this dataset, however, they are included in the original 2,000 dataset from Human Solutions.

Coder Training

The study design was based on a dataset of 2,000 3D facial scans from Human Solutions (Anthroscan for Human Solutions, n.d.). The facial scans were made available as a 3D file that could be opened in the Anthroscan software (Version 3.6.1). Four coders were instructed to digitize specific facial features that correspond to 25 facial landmarks, which are listed in Table 1 and illustrated in Figures 1 and 2.

Table 1. Description of the 25 facial landmarks.

Face Landmarks	**Definition**
Above the Ear	The area found slightly above the tip of the helix.
Alare	The lateral point on the flare of the nose on the nostril.
Back of the Head	Also called the opisthocranion, the protrusion on the posterior side of the head parallel to the sellion.
Back of the Helix	The back point of the ear.
Base of the Intertragal Notch	The "pit" below the tragus.
Center of the Eye	The center of the pupil.
Cheilion	The lateral point of the juncture of the fleshy tissue of the lips and the facial skin at the corner of the mouth.
Cheilion Center	The point of intersection between the upper and lower lip in the midsagittal plane when the mouth is closed.
Concha Cymba	The "pit" above the tragus.
Dorsal Hump	The bony protrusion on the nose.
Earlobe Juncture	The point where the earlobe meets the face.
Glabella	The most anterior point on the frontal bone midway between the bony brow ridges. Also found between the eyebrows on the browbone.
Gonion	The lateral point on the posterior angle of the mandible, the jawbone angle.
Gonion-Submandibular Midpoint	The midpoint between the gonion and submandibular.
Inion	The most prominent point at the back of the head, usually below the back of the head.
Menton	The inferior point of the mandible in the midsagittal plane (bottom of the chin).
Most Posterior Expansion of the Concha	The deepest part of the ear.
Nasal Root	The point between the nose bridge and inner eye, where the nose flattens to become the face.
Otobasion Superius	The point where the helix meets the head, found very close to the tragion.

Outer Corner of the Eye	The outer corner of the eye, where the eyelids meet.
Pronasale	The point of the anterior projection of the tip of the nose.
Sellion	The point of the deepest depression of the nasal bones at the top of the nose.
Submandibular	The juncture, in the midsagittal plane, of the lower jaw and the neck.
Subnasale	The point of intersection between the piltrum (groove of the upper lip) and the inferior surface of the nose in the midsagittal plane.
Tip of the Helix	The highest point of the helix, on top of the ear.
Tip of the Lobe	The lowest tip of the earlobe.
Top of the Head	The highest point on the head, often parallel to the tragus.
Tragion	The superior point on the juncture of the cartilaginous flap (tragus) of the ear with the head.
Tragus	The prominence on the inner side of the external ear, in front of and partly closing the passage to the organs of the ear.
Zygomatic Arch	The most protrusive part of the cheekbone.

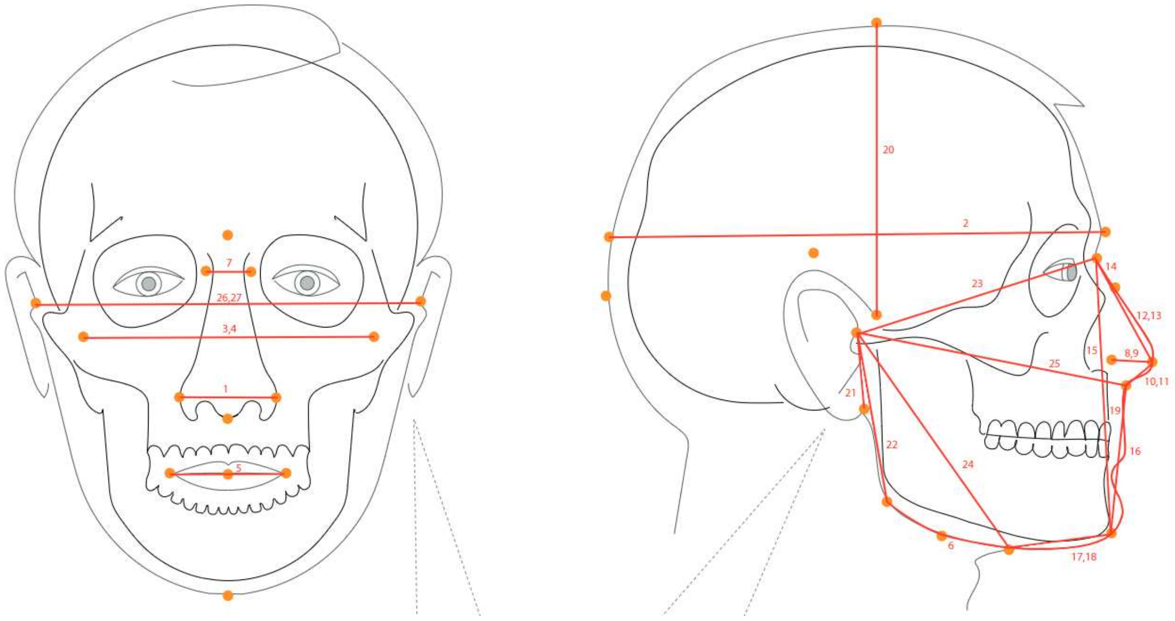
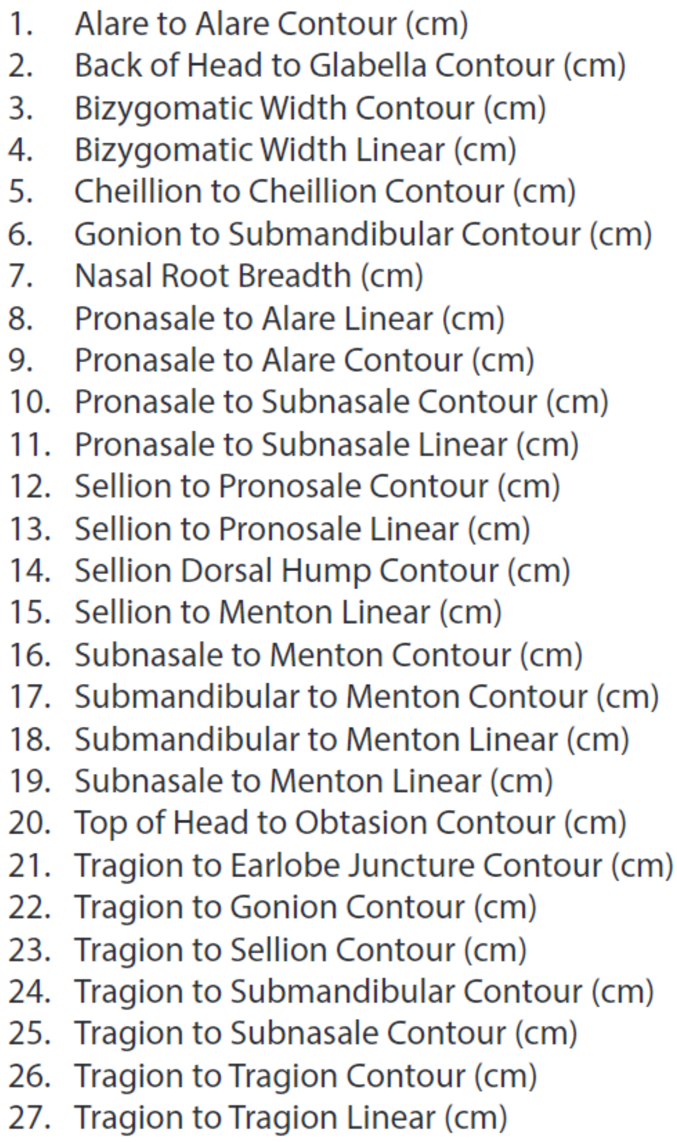


Figure 1. Anterior and side views of the 27 linear and contoured measurements created from digitization of 3D facial scans in the Anthroscan software.



1. Alare to Alare Contour (cm)
2. Back of Head to Glabella Contour (cm)
3. Bizygomatic Width Contour (cm)
4. Bizygomatic Width Linear (cm)
5. Cheillion to Cheillion Contour (cm)
6. Gonion to Submandibular Contour (cm)
7. Nasal Root Breadth (cm)
8. Pronasale to Alare Linear (cm)
9. Pronasale to Alare Contour (cm)
10. Pronasale to Subnasale Contour (cm)
11. Pronasale to Subnasale Linear (cm)
12. Sellion to Pronosale Contour (cm)
13. Sellion to Pronosale Linear (cm)
14. Sellion Dorsal Hump Contour (cm)
15. Sellion to Menton Linear (cm)
16. Subnasale to Menton Contour (cm)
17. Submandibular to Menton Contour (cm)
18. Submandibular to Menton Linear (cm)
19. Subnasale to Menton Linear (cm)
20. Top of Head to Obtasion Contour (cm)
21. Tragion to Earlobe Juncture Contour (cm)
22. Tragion to Gonion Contour (cm)
23. Tragion to Sellion Contour (cm)
24. Tragion to Submandibular Contour (cm)
25. Tragion to Subnasale Contour (cm)
26. Tragion to Tragion Contour (cm)
27. Tragion to Tragion Linear (cm)

Figure 2. The numbered 27 linear and contoured measurements that are shown in Figure 1.

Before coders digitized facial scans for the Time 1 intraRR and interRR phase, they were provided with a guide of the 25 facial landmarks they had to be familiar with and a step-by-step guide of the landmark placement process. After reviewing the documents, coders were asked by the primary investigator to watch a video that explains the importance of facial anthropometry and a video tutorial of the primary investigator placing landmarks in Anthroscan. In the video tutorial, the primary investigator explains exactly what the coders will be doing in Anthroscan, such as how to use the mouse to place a landmark or zoom in and what keys allow the coders to move the subject to get the best view of a landmark. Once coders felt comfortable with their knowledge of the 25 facial landmarks, they were asked to take a quiz testing their ability to recognize the landmarks. If coders received full credit on the landmark quiz, they were able to begin digitizing facial scans meant for training, and if they did not receive full credit, they reviewed the training materials and re-attempted the quiz. This part of the training process took a week and a half to complete.

Time 1 IntraRR and InterRR Phase

After passing the landmark quiz, each coder was instructed to digitize a sample of ten 3D facial scans that were selected by the primary investigator for training purposes. Each of the ten facial scans were digitized three times and coders did one set of ten right after the other at their own pace. It was expected that coders would take one week to digitize the ten facial scans three times, completing the entirety of their training within a three-week timespan. After landmark digitization was complete, coders checked the landmarks and resulting measurements for errors and the facial scan data was saved as a comma-separated value (CSV) file. Coders opened the CSV file, copied 28 landmark measurements (Table 2), and pasted the data into a Microsoft Excel spreadsheet for each record. All samples were denoted on a numeric system such that coders would not have access to any personal identifiers or demographic information of the subjects represented by the 3D facial scan. Once all training materials and protocols were complete, the training data was reviewed for each coder and any areas of concern were addressed in a remote one-on-one meeting with the coders. Training materials and protocols were assessed for proficiency with each coder and deficiencies in the training protocols were remedied.

Table 2. Anthroscan measurement identification numbers and IDs.

Facial Measurement Number	Facial Measurement ID	Facial Measurement Name
1	Alare_Contour	Alare to Alare Contour (cm)
2	BckHD_Glab	Back of Head to Glabella Contour (cm)
3	Bizyg_Width	Bizygomatic Width Contour (cm)
4	Bizyg_Width_Linear	Bizygomatic Width Linear (cm)
5	Cheill_Contour	Cheillion to Cheillion Contour (cm)
6	DUMMY1	Dummy (cm)
7	Gonion_Subman	Gonion to Submandibular Contour (cm)
8	Nas_Root_Brdth	Nasal Root Breadth (cm)
9	ProNas_AL_Linear	Pronasale to Alare Linear (cm)
10	ProNas_Alare	Pronasale to Alare Contour (cm)
11	ProNas_SubNas	Pronasale to Subnasale Contour (cm)
12	ProNas_SubNas_Linear	Pronasale to Subnasale Linear (cm)
13	Sel_Pronasale	Sellion to Pronosale Contour (cm)
14	Sel_Pronasale_Linear	Sellion to Pronosale Linear (cm)
15	Sell_Dorsal	Sellion Dorsal Hump Contour (cm)
16	Sellion_Ment	Sellion to Menton Linear (cm)
17	SubNas_Ment	Subnasale to Menton Contour (cm)
18	Subman_Ment	Submandibular to Menton Contour (cm)
19	Subman_Ment_Linear	Submandibular to Menton Linear (cm)
20	Subnas_Ment_Linear	Subnasale to Menton Linear (cm)
21	TopHD_Obt	Top of Head to Tragion Contour (cm)
22	Trag_Earlobe	Tragion to Earlobe Juncture Contour (cm)
23	Trag_Gonion	Tragion to Gonion Contour (cm)

24	Trag_Sel	Tragion to Sellion Contour (cm)
25	Trag_Subman	Tragion to Submandibular Contour (cm)
26	Trag_Subnas	Tragion to Subnasale Contour (cm)
27	TragtoTrag_Contour	Tragion to Tragion Contour (cm)
28	TragtoTrag_Linear	Tragion to Tragion Linear (cm)

The ICC statistic was used to analyze agreement within and between coders. The researcher chose ICC characteristics based on Koo guidelines (Table 3): the two-way mixed effects model was chosen because the four coders were the only raters of interest due to time and resources allotted for this study. The ICC type, the “mean of the *k* raters,” was chosen because the reliability study design uses data from four coders. The ICC definition of “absolute agreement” was chosen to analyze whether coders assigned the same scores to the same subjects, and this definition allowed the researcher to identify broad measurement errors due to coder training or other factors that potentially affected the selected raters as a group versus individually.

Table 3. Definition of Koo’s ICC Characteristics (Recreated table from Koo & Li (2016)).

Model	
One-way random-effects	Each subject is rated by a different set of raters who were randomly chosen from a larger population of possible raters.
Two-way random-effects	Raters are randomly selected from a larger population of raters with similar characteristics.
Two-way mixed-effects	Selected raters are the only raters of interest, cannot be generalized to other raters.
Type	
"Single rater"	Using the value from a single rater as the basis for the actual measurement.
The "mean of *k* raters"	Using the value of 4/5 raters as an assessment basis, the reliability experiment design will involve 4/5 raters.

Definition	
Absolute agreement	If different raters assign the same score to the same subject.
Consistency	If raters' scores to the same group of subjects are correlated in an additive manner.

Time 2 InterRR Phase

The four coder’s level of interRR was assessed after they completed all Time 1 training materials and protocols and reached the 30-scan threshold. Each coder digitized the same novel sample of ten 3D facial scans once, and the ICC statistic was used to assess agreement between coders. ICC values that fell in the moderate range (Table 4) were considered an acceptable level of agreement to continue to the Time 3 intraRR and interRR phase of the study.

Table 4. Koo’s ICC Agreement Range (Recreated table from Koo & Li (2016)).

ICC Values	Interpretation
< 0.500	Poor Reliability
0.500 - 0.750	Moderate Reliability
0.750-0.900	Good Reliability
> 0.900	Excellent Reliability

Time 3 IntraRR and InterRR Phase

After coders digitized 100 facial scans, they were expected to be fully competent in the digitization process of facial landmarks using the Anthroscan software. Coders were required to make judgements regarding the placement or non-placement of landmarks on a wide variety of face shapes. The four experienced coder’s level of intraRR and interRR was assessed in the final phase of the study, which was denoted as Time 3. Each coder digitized an identical novel sample of ten 3D facial scans three times. The first set of ten scans were digitized once a coder reached the 100-scan threshold, the same set of facial scans were digitized a second time after one week had passed, and the third digitization of the ten facial scans occurred the following week (Table 5). Once data collection was complete, the ICC statistic was used to assess agreement within and between coders. The ICC characteristics defining an acceptable level of agreement in the Time 2 reliability phase were used for Time 3.

Table 5. Post-100 facial scan schedule for coders.

Coder	Week 1	Week 2	Week 3
A	1 - 10	11 - 20	21 - 30
B	1 - 10	11 - 20	21 - 30
C	1 - 10	11 - 20	21 - 30
D	1 - 10	11 - 20	21 - 30

Statistical Analysis

The facial scans were digitized by the same set of coders throughout the entirety of the study, making it a fully crossed research design. This allows the researcher to control for systematic bias between coders and achieve higher ICC estimates. A fully crossed research design would also give a better estimate of true reliability and negates the need for complicated statistical analyses (Hallgren, K., 2012). The ICC statistic was used to analyze agreement within and between coders. The researcher chose ICC characteristics based on Koo guidelines (Table 3).

Times 1, 2, and 3 intraRR and interRR phase analysis was conducted using RStudio (R Core Team, 2020) and the irr (Gamer, M., Lemon, J., & Fellows Puspendra

Singh, I., 2019) and lpSolve (Berkelaar, M. et al., 2020) packages. The irr package was chosen over other available packages that analyze ICC estimates because it allowed the researcher to denote ICC characteristics, such as the “two-way” model and the “agreement” definition, in the code to calculate a single ICC value. Other RStudio packages available to calculate the ICC statistic calculate the values of all ICC characteristics, yielding a larger dataset that is not consistent with a given reliability study’s chosen ICC characteristics (DataNovia, n.d.). The ICC statistic was calculated for all 28 landmark measurements between coders (interRR) and within coders (intraRR).

CHAPTER 4 – RESULTS

ICC estimates were chosen based on Koo guidelines (Table 3) and calculated using RStudio software with the irr (Gamer, M., Lemon, J., & Fellows Puspendra Singh, I., 2019) and lpSolve (Berkelaar, M. et al., 2020) packages based on a mean rating (k = 4), absolute agreement, and the two-way mixed-effects model for each facial measurement. Training deficiencies were addressed during remote meetings with each coder after their data from Time 1 was reviewed.

Specific Aim 1

Time 1 IntraRR

Coder A had one facial measurement, the Gonion_Subman, that fell into the poor reliability range (ICC<0.500), but this measurement was not statistically significant (p>0.05). 15 of 28 (54%) measurements at Time 1 had excellent intraRR agreement (ICC>0.900). Coder B had three facial measurements that had poor reliability, which were all not statistically significant (p>0.05) and 10 of the 28 (36%) facial measurements had excellent agreement. Coder B's only other insignificant measurement fell into the moderate reliability range. Coder C had five poor facial measurements, which were all statistically insignificant (p>0.05), and 7 of 28 (25%) measurements had excellent agreement. Coder D had two measurements that showed poor reliability, both of which were not statistically significant (p>0.05), the Bizyg_Width and Sell_Dorsal. Coder D had 6 of the 28 (21%) measurements with excellent agreement. Only one facial anthropometric measurement, Bizyg_Width, had poor reliability overlap between Coders B and D, the other poor measurements were all unique to each coder.

Table 6. Time 1 intra-rater reliability percent agreement for all coders.

	Intra RR Percent Agreement (%)			
Coder	Poor	Moderate	Good	Excellent
A	3	25	18	54
B	11	14	39	36
C	18	18	39	25
D	7	32	39	21

Training Reassessment

Coders were asked to fill out a post-training questionnaire after they had completed the digitization process in Anthroscan for Time 1. The results of the questionnaire can be found in the appendix. Each coder provided thoughtful responses that allowed the researcher to understand what training and/or software deficiencies were present. Coders had difficulty with the platform Anthroscan was operating on and the quality of certain scans. Most, if not all, of the coders agreed that the zygomatic arches, the gonion, and the submandibular were the hardest landmarks to place.

Time 3 IntraRR

Coder A had one facial measurement, the Subman_Ment, that fell into the poor reliability range and was statistically insignificant (ICC<0.500 and p>0.05) and 25 of the 28 (89%) facial measurements showed excellent intraRR agreement (ICC>0.900). Coder B had no facial measurements that fell into the poor reliability range and had 24 of the 28 (86%) measurements with excellent agreement. All of Coder B’s data was statistically significant (p<0.05). Coder C had one poor statistically insignificant measurement, the Nas_Root_Brdth, and 17 of the 28 (61%) facial measurements had excellent agreement. Coder D had no measurements that were statistically insignificant or fell below the moderate (ICC=0.500-0.750) range and 24 of 28 (86%) facial measurements had excellent agreement. There was excellent intraRR agreement in the majority of the facial anthropometric measurements assessed at Time 3 for each coder, showing great improvement for each coder from Time 1.

Table 7. Time 3 intra-rater reliability percent agreement for all coders.

	Percent Agreement (%)			
Coder	Poor	Moderate	Good	Excellent
A	3	3	3	89
B	0	7	7	86
C	3	11	25	61
D	0	7	7	86

Coder ICC Comparisons

Coders A, B, and D demonstrated great intraRR improvement from Time 1 to Time 3. Each coder reduced their overall number of facial measurements with poor reliability (ICC<0.500) by the end of the intraRR study.

Table 8. Comparison of Time 1 and Time 3 intra-rater reliability in percent agreement for all coders.

Time 1 Percent Agreement (%)				
Coder	Poor	Moderate	Good	Excellent
A	4	25	18	54
B	11	14	39	36
C	18	18	39	25
D	7	32	39	21
Time 3 Percent Agreement (%)				
Coder	Poor	Moderate	Good	Excellent
A	4	4	4	89
B	0	7	7	86

C	0	11	25	61
D	0	7	7	86

Specific Aim 2

Time 1 InterRR

The ICC statistic at Time 1 for interRR indicated that 6 of the 28 (21%) measurements fell into the poor reliability range (ICC<0.500) and 4 of the 28 (14%) facial measurements indicated excellent inter-rater agreement (ICC>0.900).

Time 2 InterRR

The ICC statistic at Time 2 for interRR indicated that four facial measurements fell in the poor reliability range (ICC<0.500). The poor facial measurements were the Nas_Root_Brdth, SubNas_Ment, Subnas_Ment_Linear, and TopHD_Obt.Thirteen of the 28 (46%) facial measurements at Time 2 had excellent interRR agreement (ICC>0.900).

Time 3 InterRR

After the ICC statistic was computed in RStudio for the Time 3 interRR phase, 4 of the 28 (14%) measurements fell into the poor reliability range (ICC<0.500) and 13 of the 28 (46%) facial measurements showed excellent inter-rater agreement (ICC>0.900).

InterRR ICC Comparisons

Time 1 had 6 of the 28 (21%) measurements that had poor reliability (ICC<0.500) and 14% of the measurements had excellent reliability (ICC>0.900). Times 2 and 3 had 4 of the 28 (14%) measurements that had poor reliability and 46% of the measurements for Times 2 and 3 had excellent reliability. Figure 3 shows a visual comparison of the number of poor, moderate, good, and excellent ICC values during each time of the reliability study and shows a stable increase in the number of excellent ICC values and a decrease in the amount of poor facial anthropometric measurements.

Table 9. Percent agreement of inter-rater reliability during Times 1, 2, and 3.

	Percent Agreement (%)			
Time	Poor	Moderate	Good	Excellent
1	21	18	46	14
2	14	21	48	46
3	14	11	29	46

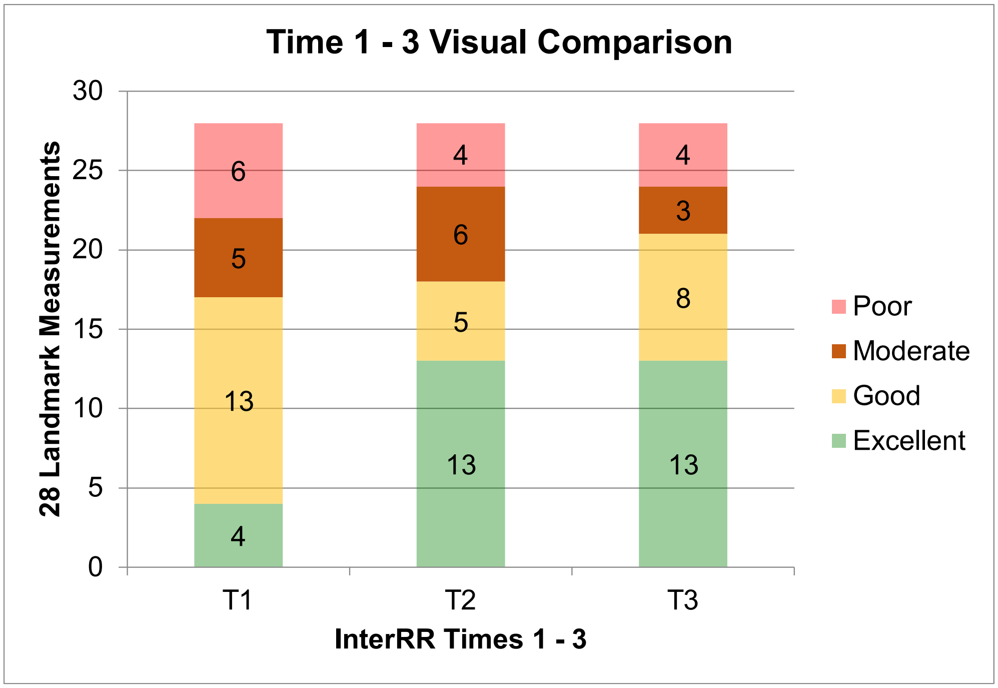


Figure 3. Visual comparison of the number of poor, moderate, good, and excellent ICC values for the 28 facial anthropometric measurements.

CHAPTER 5 – DISCUSSION

This study evaluated intra- and inter-rater reliability of 3D facial measurements being derived from four coders digitizing 25 facial landmarks with the Anthroscan software. The digital identification of facial landmarks resulted in 28 facial anthropometric measurements. IntraRR was generally moderate to excellent for most coders, with two coders having a single facial measurement that had poor reliability (ICC<0.500). The high degree of agreement within coders indicates that the coders digitized the same facial scans similarly. InterRR of the coders' 28 facial measurements suggests moderate to good reliability (ICC=0.500-0.900).

Intra-Rater Reliability

The intraRR analysis revealed that the strength of agreement between Time 1 and Time 3 varied widely between each coder. For instance, during Time 1, each coder had a number of facial measurements that fell into poor reliability, but the specific facial measurements with poor agreement were primarily unique to each coder. Coders were asked to fill out a post-training questionnaire after they digitized scans for Time 1, and one of these questions asked the coders to name the hardest landmarks to place. The coders mentioned three facial landmarks, the gonion, the zygomatic arches, and the submandibular, that were hard to place. The gonion and submandibular point are prominent on the jaw and coders mentioned that some people did not have a defined jawline compared to others with gonions and submandibular points that were easy to place. The zygomatic arches were also noted as being difficult to place because it was hard to place points where the cheekbones would be if you cannot palpate the face and had to guess where the cheekbones would actually be found on a face. After Time 3 ICC estimates were calculated for each coder, Coders A and C were only ones with a facial measurement that had poor reliability. Coder A's facial measurement with poor reliability was the Subman_Ment, or the submandibular to menton contoured measurement. Coder C's facial measurement with poor reliability was the Nas_Root_Brdth, the linear distance between two nasal root landmarks. Coders A and D had the greatest number of facial measurements that showed excellent intraRR.

Inter-Rater Reliability

InterRR during Times 2 and 3 of facial scan digitization had greater agreement than the facial scans digitized during Time 1 (14%) at the beginning of this study. Coder agreement was strongest during Times 2 and 3 (46%), meaning excellent reliability increased by 32%. At Time 1, 21% of the facial measurements had poor reliability (ICC<0.500), and the number of poor facial measurements decreased by 7% at the conclusion of this study. This suggests that identifying and correcting training deficiencies after Time 1 had a positive impact on reliability.

Comparison to Other Studies

Previous studies reported intra- or inter-rater agreement on facial measurements that were collected using either direct (hands-on) or indirect (3D) facial anthropometric measurement methods/systems. Regardless of the anthropometric methods used, training researchers, specifically coders, involved in anthropometric studies is necessary for reliable results.

In the ToyBox study (2014), six fieldworkers were asked to collect three direct anthropometric measurements. The ToyBox researchers attributed their excellent intra- and interRR to multiple training sessions for each fieldworker over the duration of the study and the initial identification of measurement error due to systematic observer bias, personal skill level, and lack of ability to adhere to study protocols (De Miguel-Etayo, P., et al., 2014). Intra- and interRR was similarly assessed in the Feel4Diabetes study (2020). Before errors could be accounted for in a reliability assessment, the Feel4Diabetes researchers ensured a standardized study protocol was in place before examiners could collect direct measurements. The Feel4Diabetes researchers provided a training workshop that all examiners involved in the study could attend and provided the same training materials during each phase of the study. IntraRR and interRR for the three anthropometric measurements collected from the adult and children populations ranged from 0.950-0.990 (Androutsos, A., et al., 2020). The high degree of agreement amongst and between examiners could have been due to the implementation of a standardized training plan and study protocol.

Standardization in study protocols and the identification of measurement errors due to human error seems to have a positive impact on the overall reliability of measurement methods. In the present study, the effect of re-training and the identification of coder-specific measurement errors after Time 1 improved interRR for Times 2 and 3 by 32%. This was similar to the effect that multiple training sessions had on intra- and inter-RR for the ToyBox researchers. Intra- and interRR was excellent at the conclusion of the ToyBox study, with 100% of their anthropometric measurements achieving ICCs in the excellent range (ICC>0.900). At the conclusion of the present study, 46% of the 28 facial anthropometric measurements fell into the excellent range. Higher agreement could potentially be achieved at Time 1 if the coders were to attend training sessions where they are able to ask questions and practice digitizing facial scans before they begin data collection like the Feel4Diabetes researchers had their examiners do. Developing and implementing a standardized training plan like the Feel4Diabetes researchers and having the Anthroscan coders attend the same training sessions before data collection occurs could lead to a higher percentage of excellent ICCs at Time 1 of future studies. Reliability may also increase in future phases of a reliability study if multiple training sessions are implemented for the Anthroscan coders as they were for the ToyBox study fieldworkers.

In 2007, NIOSH contracted Anthrotech, Inc. to collect 18 direct head and facial anthropometric measurements. Anthrotech researchers collected measurements from over 4,000 people, however, measurement error analyses, such as intra- and interRR were not conducted in this study. Comparing the Anthrotech study to the present study, Anthrotech's sample size was 4,000 subjects, which is larger than the sample size used in this study, which was 30 facial scans, or subjects. The present study collected ten more facial anthropometric measurements, but comparisons between the two studies is difficult to make as there was no error analyses for direct measurements in the Anthrotech study. Furthermore, with the indirect measurements that were collected from 1,000 individuals were not assessed for agreement (Institute of Medicine, 2007).

Meruane et al. (2015) sought to compare an indirect anthropometric measurement system to direct methods by using the *Rhinobase* ® software developed in 2009. Like Anthroscan, *Rhinobase* ® can utilize 3D facial scanning to derive nasofacial anthropometric measurements. To assess reliability, Meruane et al. took nasofacial measurements directly and indirectly of rhinoplasty patients before and after surgery using the ICC statistic. Two surgeons placed ten nasofacial landmarks on patients and repeated the process after six months had passed. Both intra- and interRR was above 0.900 for 80% of the nasofacial measurements, and in the present study, intra- and interRR was above 0.900 for 46% of the facial measurements. Researchers of the *Rhinobase* ® software guessed that poor to moderate reliability could have been due to poor patient photos and landmark placement. This is a problem that occurred in the present study and with the Anthroscan software.

A facial anthropometric study conducted by de Sá Gomes et al. (2019) also assessed reliability of facial images. The de Sá Gomes et al. study collected eleven direct craniofacial anthropometric measurements; eleven measurements without craniofacial reference points, and eleven measurements with craniofacial measurement points. For intra- and inter-rater reliability in the de Sá Gomes et al. study, 72% and 55% of their linear facial measurements, respectively achieved excellent reliability when measurements were collected without a reference point like the present study. For intra- and inter-rater reliability of the present study, the coders averaged 81% and 46% respectively. Considering that the Anthroscan coders digitized 25 facial landmarks, which resulted in 28 linear and contoured measurements, versus the eleven linear measurements collected in the de Sá Gomes et al. study, the Anthroscan coders exhibited great reliability.

Modabber et al. (2016) utilized reference points in a study of facial anthropometrics. The presence of reference points, which were placed during data collection, contributed to the increased accuracy at the end of each study. Modabber et al. achieved mean errors of 0.288 and 0.241 for the forehead and right cheek of Artec Eva and 0.523 and 0.630 for FaceScan3D. The present study conducted facial scans without any type of facial anthropometric reference point, which might have increased accuracy as well as reliability if the facial landmarks had been drawn on as reference points prior to 3D facial scanning. The present study also did not calculate mean error for the 28 facial anthropometric measurements, but the percentage of facial measurements that exhibited poor reliability was 1% for average intra-rater reliability and 14% for inter-rater reliability at the end of Time 3.

Ayaz et al. (2020) conducted reliability and accuracy evaluations between direct and indirect anthropometric measurement systems. Twenty-two facial landmarks were placed directly on the subjects' faces, which was used as reference points when the researchers collected 2D images and 3D facial scans. Ayaz et al. achieved excellent reliability (96%+) for both intra- and inter-rater reliability for measurements derived from 3D facial scans. The present study achieved an average of 81% for intra-rater reliability and 46% for inter-rater reliability for facial measurements that fell into the excellent reliability range. If the present study had been able to assess the accuracy of landmark placement, or if Human Solutions had been able to place the 25 facial landmarks directly on the subjects' faces as reference points for 3D facial scan, followed by the digitization of each facial scan, the present study may have had reliability that mirrored Ayaz et al.'s results.

Limitations and Future Research

During the present study, the researchers did not assess reliability of facial anthropometrics using intentionally placed landmarks. Instead, they analyzed the reliability of facial measurements derived from a combination of 25 facial features identified by coders from 3D facial scans. Future research should assess the reliability of indirect landmark placement versus direct landmark placement and investigate the accuracy of each coder digitizing the 25 facial landmarks rather than the agreement of facial measurements that are derived from 2-3 specific landmarks.

Due to limited resources, the present study used a fully crossed research design with a sample of 30 facial scans with four coders being assessed for inter- and intra-rater reliability, however, with the original dataset including approximately 2,000 facial scans, future researchers should assess inter- and intra-rater reliability of a new set of coders digitizing the entire dataset. This would allow researchers to assess inter- and intraRR with a larger or full dataset and compare their results against the present study's generalized results. Lastly, coders expressed that they experienced visual and mental fatigue while digitizing the 3D facial scans. As coders spent more time using the Anthroscan software, it may have affected the quality of their facial landmark placement. Future research should assess the effect of visual and mental fatigue on landmark placement with 3D anthropometric software, however, this would be a longitudinal study that would follow coders working a set amount of hours each week for approximately one year.

CHAPTER 6 – CONCLUSION

3D scanning is becoming more popular for anthropometric research and applications due to its advancement and general ease of use for both researchers and subjects. Recent studies suggest that 3D anthropometric methods are valid, and have high levels of reliability, and may be more efficient and useful when measuring large cohorts. The present study indicated that interRR ranged between good to excellent following coder training and practice for 75% of the facial anthropometric measurements. Additionally, among the 4 coders, the mean ICCs for intraRR were in the excellent range for 81% of the facial anthropometrics. The interRR of the facial anthropometric measurements in the present study were relatively high. Other direct and indirect anthropometric studies reporting interRR have also reported generally good to excellent reliability. The specific differences in inter and intra- and interRR results between studies is likely due to differences in methodology, which include the measurement system employed, 2D versus 3D assessments, direct versus indirect measurements, variables assessed, and statistical methods for determining reliability. The results of the present study indicate that 3D facial scans digitized by several coders to obtain anthropometric measurements of the face yield result in good to excellent reliability. Lastly, rater (coder) training is an important first step in the process of obtaining high intra and inter-rater reliability of 3D facial anthropometric measurements.

REFERENCES

Koo, TK. & Li, MY. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*. 15: 155-163.

Kuehnapfel, A., Ahnert, P., Loeffler, M., & Scholz, M. (2017). Body surface assessment with 3D laser-based anthropometry: reliability, validation, and improvement of empirical surface formulae. *European Journal of Applied Physiology*. 117: 371-380. doi: 10.1007/s00421-016-3525-5

Simmons, KP. & Istook, CL. (2003). Body measurement techniques: Comparing 3D body-scanning and anthropometric methods for apparel applications. *Journal of Fashion Marketing and Management*. 7(3): 306-332. https://doi.org/10.1108/13612020310484852

Wolan-Nieroda, A., Guzik, A., Mocur, P., Drużbicki, M., & Maciejczak, A. (2020). Assessment of Interrater and Intrarater Reliability of Cervical Range of Motion (CROM) Goniometer. *Biomedical Research International*. 20. https://doi.org/10.1155/2020/8908035

Yang, Y., Wang, Y., Zhou, Y., Chen, C., & Xing, D. (2016). Reliability of functional gait assessment in patients with Parkinson disease: Interrater and intrarater reliability and internal consistency. *Medicine*.

Hallgren, K. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quantitative Methods Psychology*. 8(1): 23-34.

Wong, JY. et al. (2008). Validity and Reliability of Craniofacial Anthropometric Measurement of 3D Digital Photogrammetric Images. *The Cleft Palate-Craniofacial Journal.* 45(3): 232-239. https://doi.org/10.1597/06-175

De Miguel-Etayo, P., et al. (2014). Reliability of anthropometric measurements in European preschool children: the ToyBox-study. *Obesity Reviews*, *15*(S3), 67–73. https://doi.org/10.1111/obr.12181

Androutsos, A. et al. (2020). Intra- and inter- observer reliability of anthropometric measurements and blood pressure in primary schoolchildren and adults: the Feel4Diabetes-study. *BMC Endocrine Disorders*, *20*(Suppl 1), 27–27. https://doi.org/10.1186/s12902-020-0501-1

Meruane, Ayala, M. F., García-Huidobro, M. A., & Andrades, P. (2015). Reliability of Nasofacial Analysis Using Rhinobase® Software. *Aesthetic Plastic Surgery*, *40*(1), 149–156. https://doi.org/10.1007/s00266-015-0569-6

Institute of Medicine. (2007). Assessment of the NIOSH Head-and-Face Anthropometric Survey of U.S. Respirator Users. Washington, DC: The National Academies Press. https://doi.org/10.17226/11815.

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

Ayaz, 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), 15–22. https://doi.org/10.5624/isd.2020.50.1.15Modabber, 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

Human Solutions. (n.d.). About Human Solutions. Retrieved May 8, 2022, from https://www.human-solutions.com/en/about-human-solutions/about-us/index.html

Matthias Gamer, Jim Lemon and Ian Fellows Puspendra Singh (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Michel Berkelaar and others (2020). lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs. R package version 5.6.15. https://CRAN.R-project.org/package=lpSolve

DataNovia. How to choose the correct ICC forms. DataNovia. https://www.datanovia.com/en/lessons/intraclass-correlation-coefficient-in-r/

APPENDIX

Appendix I. Post-training questionnaire results for all coders for question 1.

Coders	Question 1: What did you find difficult about digitizing the landmarks?			
A	The dark blue background if subjects were a person of color	Certain landmarks/subjects required toggling between "True Color" and "Monochrome"	Using the zoom window sometimes placed a "permanent" landmark that you would have to get rid of	Poor lighting and skin color made placing some landmarks difficult
B	Moving the 3D object around in Anthroscan is a little clunky, the software could be updated to be more intuitive.	Sometimes hair is covering the landmarks, which makes it very difficult to locate them	Sometimes the scan's orientation is tilted, making it difficult to rotate the scan object as needed to accurately place the landmarks.	
C	Working to understand the level of variation that exists between gender and race	It took some time to get used to understanding if a landmark position was actually obstructed by facial hair or some other discrepancy		
D	Not having enough reference photos for a specific landmark	When a person didn't fit into the "standard" facial landmark example		

Appendix II. Post-training questionnaire results for all coders for question 2.

Coders	Question 2: What landmarks were the hardest to place?				
A	Zygomatic Arches	Lip landmarks if subject had dark skin	Center of the pupil if the eyes were brown/black	Dorsal hump (if not pronounced)	Gonion and submandibular

B	Zygomatic arches	Nasal root	Gonion (on people without a pronounced jaw)	Submandibular (on people with extra neck skin/mass)	Inner-ear landmarks (concha cymba, most posterior expansion of the concha, etc)
C	Landmarks in/around the ears	Rotated/skewed facial scans forced the coder to work slower to hit all the landmarks	Practice makes landmark placement easier		
D	Zygomatic arches	Jaw landmarks	Ear landmarks		

Appendix III. Post-training questionnaire results for all coders for question 3.

Coders	**Question 3: What was easy about landmark placement?**		
A	Using the arrow keys to move a subject around was fast	Double-clicking to place a landmark	"Easy" landmarks were fun to place
B	Finding easy landmarks (center of the pupil, tip of the nose, etc)	Using a mouse made the process easier	Repetition allowed for the process to be made easier
C	Landmark placement was easy when the person had "model" landmarks, i.e. they looked exactly like the reference guide	The double-click to place function made sure you didn't place the landmark somewhere else on accident during the navigation process	
D	How intuitive it was to place landmarks with the mouse controls		

Appendix IV. Post-training questionnaire results for all coders for question 4.

Coders	**Question 4: Is there anything that could be improved/made easier for future data collection?**		
A	The background color changing to a lighter color to see ALL persons would be nice	Prepping subjects for scanning should be done in a more controlled way	Perhaps marking some hard landmarks, such as the zygomatic arches, during scanning could help in the digitization process
B	Anthroscan software!	If the scans were all done correctly… ears are uncovered, hair pulled back behind ears, no wrinkles in swim cap.	
C	To create more consistency in data collection, it would be helpful to ensure that most or all of the head facial scan data comes into the software oriented the same (bust should be level and upright)		
D	Having multiple example photos for each landmark, not just a drawing		

Appendix V. Time 1 intra-rater reliability ICC values and F-Test for 28 facial measurements for all coders.

Coder A		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.944	0.835	0.985	16.30	9	18.2	5.62E-07
BckHD_Glab	0.675	-0.012	0.914	2.89	9	18.2	0.0261
Bizyg_Width	0.745	0.275	0.930	5.01	9	13.1	0.0046
Bizyg_Width_Linear	0.859	0.546	0.963	9.89	9	11.0	0.0004

Cheill_Contour	0.968	0.907	0.992	28.60	9	18.0	7.05E-09
DUMMY1	0.993	0.979	0.998	137.00	8	17.9	1.8E-14
Gonion_Subman	0.418	-0.596	0.850	1.78	8	17.9	0.1470
Nas_Root_Brdth	0.891	0.693	0.970	9.86	9	19.2	1.65E-05
ProNas_AL_Linear	0.824	0.486	0.952	5.43	9	19.4	0.0009
ProNas_Alare	0.715	0.150	0.924	3.35	9	19.1	0.0125
ProNas_SubNas	0.907	0.730	0.975	10.20	9	19.1	1.38E-05
ProNas_SubNas_Linear	0.938	0.817	0.983	14.80	9	18.5	1.02E-06
Sel_Pronasale	0.899	0.716	0.972	10.30	9	19.8	9.80E-06
Sel_Pronasale_Linear	0.906	0.733	0.975	11.90	9	18.3	5.92E-06
Sell_Dorsal	0.698	0.155	0.916	4.37	9	11.9	0.0104
Sellion_Ment	0.958	0.872	0.989	23.40	8	18.0	5.33E-08
SubNas_Ment	0.933	0.798	0.983	15.20	8	18.0	1.56E-06
Subman_Ment	0.661	0.041	0.914	3.21	8	17.0	0.0206
Subman_Ment_Linear	0.608	-0.086	0.899	2.79	8	16.8	0.0362
Subnas_Ment_Linear	0.947	0.838	0.987	17.90	8	17.6	5.27E-07
TopHD_Obt	0.926	0.778	0.982	13.80	8	17.9	3.10E-06
Trag_Earlobe	0.918	0.665	0.981	20.20	8	7.7	0.0002
Trag_Gonion	0.869	0.547	0.972	6.93	7	14.6	0.0010
Trag_Sel	0.987	0.960	0.997	93.90	8	15.3	1.40E-11
Trag_Subman	0.711	0.100	0.936	3.61	7	15.9	0.0161
Trag_Subnas	0.995	0.983	0.999	164.00	8	16.1	6.1E-14
TragtoTrag_Contour	0.988	0.963	0.997	79.20	8	17.9	2.08E-12
TragtoTrag_Linear	0.978	0.935	0.995	45.00	8	17.	2.75E-10

						9	

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder B		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.817	0.466	0.950	5.23	9	19.4	0.0011
BckHD_Glab	0.770	0.331	0.938	4.21	9	19.6	0.0038
Bizyg_Width	0.267	-0.612	0.774	1.47	9	18.7	0.2290
Bizyg_Width_Linear	0.687	0.153	0.912	3.80	9	15.4	0.0105
Cheill_Contour	0.916	0.758	0.977	11.40	9	19.5	5.00E-06
DUMMY1	0.181	-1.464	0.797	1.22	8	17.3	0.3430
Gonion_Subman	0.870	0.609	0.968	7.66	8	18.0	0.0002
Nas_Root_Brdth	0.886	0.655	0.969	11.10	9	14.0	5.79E-05
ProNas_AL_Linear	0.843	0.558	0.957	6.63	9	19.7	0.0002
ProNas_Alare	0.751	0.302	0.932	4.10	9	20.0	0.0041
ProNas_SubNas	0.550	-0.322	0.879	2.18	9	19.6	0.0714
ProNas_SubNas_Linear	0.607	-0.085	0.892	2.61	9	20.0	0.0358
Sel_Pronasale	0.932	0.807	0.982	15.90	9	19.1	4.30E-07
Sel_Pronasale_Linear	0.945	0.843	0.985	19.00	9	19.6	7.69E-08
Sell_Dorsal	0.692	0.134	0.914	4.36	9	11.3	0.0118
Sellion_Ment	0.934	0.803	0.984	15.70	8	17.9	1.25E-06
SubNas_Ment	0.917	0.713	0.980	16.70	8	10.9	4.34E-05
Subman_Ment	0.858	0.579	0.965	7.68	8	17.	0.0002

						1	
Subman_Ment_Linear	0.810	0.438	0.953	5.49	8	17.8	0.0014
Subnas_Ment_Linear	0.921	0.755	0.981	15.00	8	15.0	7.22E-06
TopHD_Obt	0.781	0.336	0.945	5.85	8	12.3	0.0031
Trag_Earlobe	0.876	0.601	0.970	7.31	8	16.2	0.0004
Trag_Gonion	0.762	0.259	0.947	4.49	7	15.6	0.0065
Trag_Sel	0.967	0.869	0.992	48.00	8	8.7	2.45E-06
Trag_Subman	0.965	0.888	0.992	29.30	7	15.9	5.84E-08
Trag_Subnas	0.041	-1.913	0.762	1.04	8	16.4	0.4450
TragtoTrag_Contour	0.988	0.944	0.997	154.00	8	7.0	3.41E-07
TragtoTrag_Linear	0.951	0.829	0.988	27.70	8	11.4	2.31E-06

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder C		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.402	-0.656	0.835	1.69	9	19.9	0.1580
BckHD_Glab	-0.201	-4.682	0.787	0.84	6	8.9	0.5720
Bizyg_Width	0.734	0.246	0.927	3.76	9	20.0	0.0065
Bizyg_Width_Linear	0.866	0.623	0.963	7.91	9	19.5	0.0001
Cheill_Contour	0.922	0.779	0.979	13.80	9	19.2	1.28E-06
DUMMY1	0.938	0.815	0.985	16.00	8	18.0	1.01E-06
Gonion_Subman	0.665	-0.071	0.926	3.00	7	16.0	0.0324
Nas_Root_Brdth	0.865	0.617	0.963	8.17	9	18.4	8.12E-05
ProNas_AL_Linear	0.718	0.229	0.922	3.83	9	19.	0.0066

						0	
ProNas_Alare	0.612	-0.048	0.892	2.71	9	19.6	0.0310
ProNas_SubNas	0.752	0.244	0.934	3.76	9	18.4	0.0077
ProNas_SubNas_Linear	0.814	0.435	0.950	4.95	9	18.2	0.0019
Sel_Pronasale	0.810	0.432	0.949	4.92	9	18.6	0.0018
Sel_Pronasale_Linear	0.820	0.460	0.952	5.17	9	18.5	0.0014
Sell_Dorsal	0.542	-0.463	0.880	2.08	9	18.3	0.0887
Sellion_Ment	0.879	0.576	0.977	9.63	6	12.4	0.0005
SubNas_Ment	0.870	0.533	0.975	9.78	6	10.8	0.0008
Subman_Ment	0.369	-2.046	0.890	1.51	6	12.3	0.2530
Subman_Ment_Linear	0.464	-1.507	0.906	1.75	6	12.3	0.1900
Subnas_Ment_Linear	0.889	0.591	0.979	11.60	6	10.5	0.0004
TopHD_Obt	0.928	0.717	0.989	16.20	5	10.9	0.0001
Trag_Earlobe	0.939	0.810	0.985	19.30	8	15.2	1.24E-06
Trag_Gonion	0.822	0.403	0.961	5.33	7	15.4	0.0030
Trag_Sel	0.937	0.810	0.984	16.40	8	17.8	9.42E-07
Trag_Subman	0.765	0.162	0.956	4.17	6	13.9	0.0132
Trag_Subnas	0.966	0.885	0.992	37.10	8	13.0	1.02E-07
TragtoTrag_Contour	0.971	0.910	0.993	38.50	8	16.4	3.67E-09
TragtoTrag_Linear	0.390	-0.402	0.828	1.93	8	13.7	0.1360

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder D		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.816	0.485	0.949	5.97	9	18.5	0.0006
BckHD_Glab	0.886	0.677	0.969	8.77	9	20.0	3.07E-05
Bizyg_Width	0.063	-0.370	0.592	1.15	9	18.5	0.3790
Bizyg_Width_Linear	0.594	-0.097	0.886	4.46	9	5.9	0.0430
Cheill_Contour	0.687	0.105	0.921	3.46	8	17.1	0.0149
DUMMY1	0.933	0.795	0.984	14.10	8	17.3	3.52E-06
Gonion_Subman	0.912	0.723	0.978	10.40	8	16.6	3.91E-05
Nas_Root_Brdth	0.824	0.409	0.954	8.70	9	8.9	0.0019
ProNas_AL_Linear	0.872	0.617	0.966	7.17	9	18.2	0.0002
ProNas_Alare	0.827	0.479	0.954	5.35	9	18.3	0.0012
ProNas_SubNas	0.893	0.697	0.971	9.93	9	19.4	1.50E-05
ProNas_SubNas_Linear	0.861	0.601	0.962	7.02	9	19.8	0.0002
Sel_Pronasale	0.890	0.679	0.970	8.53	9	18.9	5.13E-05
Sel_Pronasale_Linear	0.888	0.668	0.970	8.26	9	18.5	7.30E-05
Sell_Dorsal	-0.277	-1.040	0.727	0.58	9	0.3	0.8340
Sellion_Ment	0.697	-0.031	0.941	7.52	6	4.4	0.0296
SubNas_Ment	0.500	-0.136	0.884	5.33	6	3.5	0.0774
Subman_Ment	0.641	-0.092	0.927	6.10	6	4.4	0.0420
Subman_Ment_Linear	0.691	-0.022	0.939	5.92	6	5.6	0.0279
Subnas_Ment_Linear	0.626	-0.090	0.923	7.26	6	3.7	0.0456
TopHD_Obt	0.894	0.593	0.984	10.40	5	11.5	0.0006
Trag_Earlobe	0.736	0.222	0.934	3.91	8	17.9	0.0079
Trag_Gonion	0.769	0.248	0.949	4.27	7	15.9	0.0078
Trag_Sel	0.941	0.822	0.985	18.30	8	17.3	5.41E-07

Trag_Subman	0.911	0.701	0.981	10.40	7	15.2	8.76E-05
Trag_Subnas	0.958	0.869	0.990	21.60	8	16.5	2.47E-07
TragtoTrag_Contour	0.981	0.939	0.995	656.80	8	14.1	7.91E-10
TragtoTrag_Linear	0.744	0.260	0.935	4.42	8	16.0	0.0056

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Appendix VI. Time 3 intra-rater reliability ICC values and F-Test for 28 facial

measurements for all coders.

Coder A		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.987	0.963	0.996	74.20	9	19.6	3.28E-13
BckHD_Glab	0.960	0.871	0.991	24.70	7	16.0	2.01E-07
Bizyg_Width	0.917	0.765	0.977	12.00	9	20.0	2.82E-06
Bizyg_Width_Linear	0.924	0.784	0.979	13.10	9	20.0	1.41E-06
Cheill_Contour	0.984	0.949	0.997	61.80	7	15.9	2.42E-10
DUMMY1	0.983	0.949	0.996	57.00	8	17.8	3.89E-11
Gonion_Subman	0.937	0.811	0.984	16.20	8	17.9	9.35E-07
Nas_Root_Brdth	0.830	0.525	0.953	6.30	9	19.3	0.0004
ProNas_AL_Linear	0.986	0.957	0.996	79.40	9	17.4	2.51E-12
ProNas_Alare	0.975	0.927	0.993	43.60	9	18.6	1.20E-10
ProNas_SubNas	0.969	0.909	0.992	29.40	9	18.1	5.19E-09
ProNas_SubNas_Linear	0.961	0.889	0.989	25.00	9	19.8	6.24E-09
Sel_Pronasale	0.991	0.973	0.997	99.00	9	19.0	4.61E-14
Sel_Pronasale_Linear	0.991	0.975	0.998	107.00	9	19.2	1.71E-

							14
Sell_Dorsal	0.926	0.787	0.980	15.30	9	17.8	1.12E-06
Sellion_Ment	0.995	0.979	0.999	234.00	6	10.8	7.90E-11
SubNas_Ment	0.980	0.904	0.997	71.70	5	8.0	2.10E-06
Subman_Ment	0.457	-1.100	0.900	1.80	6	13.4	0.1750
Subman_Ment_Linear	0.522	-0.812	0.912	2.04	6	13.5	0.1300
Subnas_Ment_Linear	0.954	0.807	0.992	32.10	6	8.6	1.98E-05
TopHD_Obt	0.981	0.934	0.997	58.80	6	13.4	7.17E-09
Trag_Earlobe	0.979	0.940	0.994	54.40	9	18.0	3.31E-11
Trag_Gonion	0.971	0.913	0.993	36.20	8	17.8	1.86E-09
Trag_Sel	0.993	0.979	0.998	133.00	9	19.7	1.24E-15
Trag_Subman	0.993	0.978	0.998	130.00	8	17.6	4.44E-14
Trag_Subnas	0.971	0.917	0.992	33.60	9	19.9	3.96E-10
TragtoTrag_Contour	0.997	0.990	0.999	298.00	8	17.8	2.28E-17
TragtoTrag_Linear	0.998	0.995	1.000	530.00	8	16.0	6.56E-18

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder B		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.974	0.927	0.993	39.90	9	19.9	8.72E-11
BckHD_Glab	0.974	0.927	0.993	39.90	9	19.9	8.72E-11
Bizyg_Width	0.944	0.832	0.985	21.60	9	15.7	3.14E-07
Bizyg_Width_Linear	0.915	0.740	0.977	14.70	9	14.3	9.15E-06
Cheill_Contour	0.968	0.909	0.991	33.00	9	19.	6.52E-10

						5	
DUMMY1	0.996	0.988	0.999	234.00	8	16.4	2.23E-15
Gonion_Subman	0.950	0.850	0.988	20.70	8	17.8	1.51E-07
Nas_Root_Brdth	0.757	0.291	0.935	3.98	9	19.5	0.0052
ProNas_AL_Linear	0.974	0.924	0.993	35.40	9	18.8	6.35E-10
ProNas_Alare	0.960	0.885	0.989	23.50	9	19.0	1.84E-08
ProNas_SubNas	0.941	0.828	0.984	19.40	9	17.4	2.36E-07
ProNas_SubNas_Linear	0.931	0.799	0.981	16.90	9	17.0	8.34E-07
Sel_Pronasale	0.977	0.936	0.994	44.10	9	20.0	3.03E-11
Sel_Pronasale_Linear	0.978	0.936	0.994	49.80	9	18.4	4.55E-11
Sell_Dorsal	0.920	0.719	0.979	18.40	9	10.0	4.25E-05
Sellion_Ment	0.969	0.899	0.993	33.10	7	15.8	2.59E-08
SubNas_Ment	0.921	0.731	0.983	11.40	7	14.5	6.34E-05
Subman_Ment	0.740	0.093	0.944	3.60	7	15.0	0.0177
Subman_Ment_Linear	0.692	-0.089	0.934	3.05	7	14.9	0.0332
Subnas_Ment_Linear	0.849	0.474	0.967	6.02	7	14.6	0.0020
TopHD_Obt	0.970	0.911	0.993	33.60	8	18.0	2.85E-09
Trag_Earlobe	0.935	0.814	0.982	17.00	9	18.5	3.50E-07
Trag_Gonion	0.969	0.895	0.993	28.90	7	14.8	1.56E-07
Trag_Sel	0.990	0.972	0.997	97.00	9	19.5	3.01E-14
Trag_Subman	0.997	0.989	0.999	275.00	7	15.3	6.22E-15
Trag_Subnas	0.988	0.964	0.997	77.40	8	16.0	2.51E-11
TragtoTrag_Contour	0.989	0.965	0.997	79.20	8	16.5	1.17E-11
TragtoTrag_Linear	0.992	0.976	0.998	144.00	8	16.	7.46E-14

						6	

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder C		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.974	0.925	0.993	35.90	9	19.0	4.59E-10
BckHD_Glab	0.856	0.588	0.961	8.07	9	16.8	0.0001
Bizyg_Width	0.849	0.523	0.960	9.07	9	11.4	0.0005
Bizyg_Width_Linear	0.692	0.115	0.915	4.64	9	9.8	0.0132
Cheill_Contour	0.924	0.783	0.979	12.90	9	19.9	1.63E-06
DUMMY1	0.963	0.888	0.991	29.10	8	17.4	1.47E-08
Gonion_Subman	0.919	0.771	0.978	12.90	9	19.8	1.77E-06
Nas_Root_Brdth	0.384	-0.654	0.828	1.66	9	20.0	0.1660
ProNas_AL_Linear	0.961	0.889	0.989	25.20	9	19.9	5.44E-09
ProNas_Alare	0.945	0.839	0.985	16.80	9	18.9	3.18E-07
ProNas_SubNas	0.686	0.047	0.916	3.02	9	18.8	0.0206
ProNas_SubNas_Linear	0.955	0.868	0.988	20.40	9	18.9	6.42E-08
Sel_Pronasale	0.962	0.891	0.990	25.30	9	19.7	5.96E-09
Sel_Pronasale_Linear	0.970	0.914	0.992	32.00	9	19.6	8.27E-10
Sell_Dorsal	0.503	-0.278	0.859	2.16	9	19.1	0.0753
Sellion_Ment	0.977	0.924	0.995	46.00	7	15.4	3.46E-09
SubNas_Ment	0.963	0.876	0.992	31.50	7	13.9	1.74E-07
Subman_Ment	0.885	0.632	0.975	9.02	7	15.9	0.0001

Subman_Ment_Linear	0.879	0.613	0.973	8.71	7	15.7	0.0002
Subnas_Ment_Linear	0.899	0.669	0.978	11.80	7	13.6	7.92E-05
TopHD_Obt	0.857	0.567	0.965	6.86	8	17.9	0.0004
Trag_Earlobe	0.946	0.847	0.985	19.80	9	19.4	6.12E-08
Trag_Gonion	0.871	0.624	0.965	7.36	9	19.4	0.0001
Trag_Sel	0.993	0.980	0.998	130.00	9	18.6	7.17E-15
Trag_Subman	0.965	0.900	0.991	27.00	9	19.0	5.52E-09
Trag_Subnas	0.983	0.950	0.995	54.20	9	19.2	1.00E-11
TragtoTrag_Contour	0.993	0.979	0.998	146.00	8	17.9	1.05E-14
TragtoTrag_Linear	0.995	0.986	0.999	198.00	8	16.2	1.11E-14

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Coder D		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.969	0.885	0.992	47.50	9	10.0	4.71E-07
BckHD_Glab	0.947	0.832	0.987	22.90	8	14.4	6.86E-07
Bizyg_Width	0.962	0.861	0.990	39.20	9	10.0	1.28E-06
Bizyg_Width_Linear	0.926	0.791	0.980	14.50	9	19.3	8.35E-07
Cheill_Contour	0.949	0.855	0.986	21.40	9	19.0	4.04E-08
DUMMY1	0.995	0.984	0.999	200.00	8	17.2	2.05E-15
Gonion_Subman	0.944	0.817	0.988	16.80	7	15.3	4.06E-06
Nas_Root_Brdth	0.915	0.754	0.977	11.20	9	19.4	6.19E-06

ProNas_AL_Linear	0.982	0.943	0.995	71.70	9	13.9	4.18E-10
ProNas_Alare	0.964	0.887	0.990	35.10	9	14.2	3.78E-08
ProNas_SubNas	0.967	0.908	0.991	31.40	9	19.9	7.36E-10
ProNas_SubNas_Linear	0.975	0.927	0.993	36.90	9	19.0	3.78E-10
Sel_Pronasale	0.992	0.976	0.998	115.00	9	19.8	3.73E-15
Sel_Pronasale_Linear	0.992	0.975	0.998	134.00	9	17.9	1.49E-14
Sell_Dorsal	0.943	0.836	0.985	16.60	9	19.1	3.15E-07
Sellion_Ment	0.997	0.990	0.999	327.00	6	13.9	3.63E-14
SubNas_Ment	0.989	0.962	0.998	92.60	6	14.0	1.88E-10
Subman_Ment	0.970	0.891	0.994	36.80	6	13.1	1.90E-07
Subman_Ment_Linear	0.964	0.875	0.993	30.10	6	13.7	4.10E-07
Subnas_Ment_Linear	0.986	0.951	0.997	73.20	6	14.0	9.63E-10
TopHD_Obt	0.964	0.799	0.998	34.50	3	7.2	0.0001
Trag_Earlobe	0.823	0.475	0.956	6.27	8	16.5	0.0008
Trag_Gonion	0.777	0.220	0.952	4.15	7	14.8	0.0101
Trag_Sel	0.732	0.204	0.933	3.79	8	18.0	0.0090
Trag_Subman	0.907	0.700	0.980	10.60	7	16.0	5.71E-05
Trag_Subnas	0.740	0.235	0.935	3.98	8	17.9	0.0072
TragtoTrag_Contour	0.932	0.781	0.985	15.50	7	15.7	5.62E-06
TragtoTrag_Linear	0.995	0.985	0.999	225.00	7	15.5	1.97E-14

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Appendix VII. Intra-rater reliability ICC comparisons for all coders for Times 1 and 3.

Coder A	Time 1	Time 3
Landmarks	ICC	ICC
Alare_Contour	0.944	0.987
BckHD_Glab	0.675	0.960
Bizyg_Width	0.745	0.917
Bizyg_Width_Linear	0.859	0.924
Cheill_Contour	0.968	0.984
DUMMY1	0.993	0.983
Gonion_Subman	0.418	0.937
Nas_Root_Brdth	0.891	0.830
ProNas_AL_Linear	0.824	0.986
ProNas_Alare	0.715	0.975
ProNas_SubNas	0.907	0.969
ProNas_SubNas_Linear	0.938	0.961
Sel_Pronasale	0.899	0.991
Sel_Pronasale_Linear	0.906	0.991
Sell_Dorsal	0.698	0.926
Sellion_Ment	0.958	0.995
SubNas_Ment	0.933	0.980
Subman_Ment	0.661	0.457
Subman_Ment_Linear	0.608	0.522
Subnas_Ment_Linear	0.947	0.954
TopHD_Obt	0.926	0.981
Trag_Earlobe	0.918	0.979
Trag_Gonion	0.869	0.971
Trag_Sel	0.987	0.993
Trag_Subman	0.711	0.993
Trag_Subnas	0.995	0.971
TragtoTrag_Contour	0.988	0.997
TragtoTrag_Linear	0.978	0.998

*The red highlighted cells are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900.

Coder B	Time 1	Time 3
Landmarks	ICC	ICC
Alare_Contour	0.817	0.974
BckHD_Glab	0.770	0.974
Bizyg_Width	0.267	0.944

Bizyg_Width_Linear	0.687	0.915
Cheill_Contour	0.916	0.968
DUMMY1	0.181	0.996
Gonion_Subman	0.870	0.950
Nas_Root_Brdth	0.886	0.757
ProNas_AL_Linear	0.843	0.974
ProNas_Alare	0.751	0.960
ProNas_SubNas	0.550	0.941
ProNas_SubNas_Linear	0.607	0.931
Sel_Pronasale	0.932	0.977
Sel_Pronasale_Linear	0.945	0.978
Sell_Dorsal	0.692	0.920
Sellion_Ment	0.934	0.969
SubNas_Ment	0.917	0.921
Subman_Ment	0.858	0.740
Subman_Ment_Linear	0.810	0.692
Subnas_Ment_Linear	0.921	0.849
TopHD_Obt	0.781	0.970
Trag_Earlobe	0.876	0.935
Trag_Gonion	0.762	0.969
Trag_Sel	0.967	0.990
Trag_Subman	0.965	0.997
Trag_Subnas	0.041	0.988
TragtoTrag_Contour	0.988	0.989
TragtoTrag_Linear	0.951	0.992

*The red highlighted cells are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900.

Coder C	Time 1	Time 3
Landmarks	ICC	ICC
Alare_Contour	0.402	0.974
BckHD_Glab	-0.201	0.856
Bizyg_Width	0.734	0.849
Bizyg_Width_Linear	0.866	0.692
Cheill_Contour	0.922	0.924
DUMMY1	0.938	0.963
Gonion_Subman	0.665	0.919
Nas_Root_Brdth	0.865	0.384
ProNas_AL_Linear	0.718	0.961

ProNas_Alare	0.612	0.945
ProNas_SubNas	0.752	0.686
ProNas_SubNas_Linear	0.814	0.955
Sel_Pronasale	0.810	0.962
Sel_Pronasale_Linear	0.820	0.970
Sell_Dorsal	0.542	0.503
Sellion_Ment	0.879	0.977
SubNas_Ment	0.870	0.963
Subman_Ment	0.369	0.885
Subman_Ment_Linear	0.464	0.879
Subnas_Ment_Linear	0.889	0.899
TopHD_Obt	0.928	0.857
Trag_Earlobe	0.939	0.946
Trag_Gonion	0.822	0.871
Trag_Sel	0.937	0.993
Trag_Subman	0.765	0.965
Trag_Subnas	0.966	0.983
TragtoTrag_Contour	0.971	0.993
TragtoTrag_Linear	0.390	0.995

*The red highlighted cells are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900.

Coder D	Time 1	Time 3
Landmarks	ICC	ICC
Alare_Contour	0.816	0.969
BckHD_Glab	0.886	0.947
Bizyg_Width	0.063	0.962
Bizyg_Width_Linear	0.594	0.926
Cheill_Contour	0.687	0.949
DUMMY1	0.933	0.995
Gonion_Subman	0.912	0.944
Nas_Root_Brdth	0.824	0.915
ProNas_AL_Linear	0.872	0.982
ProNas_Alare	0.827	0.964
ProNas_SubNas	0.893	0.967
ProNas_SubNas_Linear	0.861	0.975
Sel_Pronasale	0.890	0.992
Sel_Pronasale_Linear	0.888	0.992
Sell_Dorsal	-0.277	0.943

Sellion_Ment	0.697	0.997
SubNas_Ment	0.500	0.989
Subman_Ment	0.641	0.970
Subman_Ment_Linear	0.691	0.964
Subnas_Ment_Linear	0.626	0.986
TopHD_Obt	0.894	0.964
Trag_Earlobe	0.736	0.823
Trag_Gonion	0.769	0.777
Trag_Sel	0.941	0.732
Trag_Subman	0.911	0.907
Trag_Subnas	0.958	0.740
TragtoTrag_Contour	0.981	0.932
TragtoTrag_Linear	0.744	0.995

*The red highlighted cells are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900.

Appendix VII. Inter-rater reliability ICC values and F-Test for 28 facial measurements for all coders for Times 1 – 3.

Time 1		95% CI		F-Test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df 1*	*df2*	p-value
Alare_Contour	0.874	0.673	0.964	9.22	9	24.9	5.28E-06
BckHD_Glab	0.827	0.476	0.966	6.16	6	20.5	0.0008
Bizyg_Width	-0.201	-0.718	0.575	0.60	9	1.1	0.7690
Bizyg_Width_Linear	0.413	-0.158	0.803	2.60	9	10.0	0.0759
Cheill_Contour	0.886	0.598	0.973	17.50	8	8.6	0.0001
DUMMY1	0.131	-1.430	0.781	1.15	8	25.6	0.3660
Gonion_Subman	0.602	0.005	0.902	3.58	7	12.4	0.0243
Nas_Root_Brdth	0.757	0.253	0.935	9.71	9	6.7	0.0039
ProNas_AL_Linear	0.825	0.543	0.951	7.16	9	20.8	0.0001
ProNas_Alare	0.750	0.370	0.928	4.91	9	21.	0.0012

						4	
ProNas_SubNas	0.715	0.279	0.919	3.62	9	29.8	0.0037
ProNas_SubNas_Linear	0.775	0.435	0.936	5.00	9	26.4	0.0005
Sel_Pronasale	0.860	0.641	0.961	7.23	9	30.0	1.65E-05
Sel_Pronasale_Linear	0.865	0.654	0.962	7.88	9	28.9	9.04E-06
Sell_Dorsal	0.269	-0.244	0.726	1.88	9	11.3	0.1580
Sellion_Ment	0.864	0.434	0.975	21.10	6	5.6	0.0011
SubNas_Ment	0.841	0.390	0.970	16.20	6	6.1	0.0017
Subman_Ment	0.148	-2.588	0.849	1.15	6	18.4	0.3720
Subman_Ment_Linear	0.285	-1.895	0.872	1.35	6	18.6	0.2850
Subnas_Ment_Linear	0.860	0.443	0.974	18.30	6	6.2	0.0011
TopHD_Obt	0.906	0.644	0.989	10.80	4	15.0	0.0002
Trag_Earlobe	0.753	0.338	0.936	5.52	8	15.7	0.0002
Trag_Gonion	0.783	0.317	0.951	8.84	7	8.3	0.0027
Trag_Sel	0.954	0.931	0.989	41.10	8	9.6	1.78E-06
Trag_Subman	0.568	-0.097	0.907	3.06	6	13.5	0.0413
Trag_Subnas	0.973	0.921	0.993	49.00	8	18.5	7.54E-11
TragtoTrag_Contour	0.976	0.900	0.994	86.50	8	8.4	3.25E-07
TragtoTrag_Linear	0.659	0.120	0.909	4.69	8	10.7	0.6590

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Time 2

		95% CI		F-Test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df 1*	*df2*	p-value
Alare_Contour	0.938	0.826	0.983	21.80	9	18.5	4.72E-08
BckHD_Glab	0.923	0.607	0.995	10.70	3	10.3	0.0017

Bizyg_Width	0.641	0.069	0.898	8.21	9	5.3	0.0139
Bizyg_Width_Linear	0.629	0.069	0.895	14.00	9	4.0	0.0110
Cheill_Contour	0.885	0.607	0.970	17.20	9	8.9	0.0001
DUMMY1	0.983	0.953	0.996	56.40	8	26.8	8.71E-15
Gonion_Subman	0.921	0.755	0.980	18.70	8	14.1	3.05E-06
Nas_Root_Brdth	0.365	-0.051	0.759	5.13	9	4.4	0.0553
ProNas_AL_Linear	0.917	0.741	0.978	19.40	9	12.6	5.31E-06
ProNas_Alare	0.879	0.573	0.969	17.50	9	8.1	0.0002
ProNas_SubNas	0.674	0.152	0.915	3.26	8	26.2	0.0104
ProNas_SubNas_Linear	0.811	0.491	0.951	6.50	8	20.2	0.0003
Sel_Pronasale	0.959	0.851	0.990	46.90	9	9.7	7.92E-07
Sel_Pronasale_Linear	0.960	0.861	0.990	44.90	9	10.6	3.15E-07
Sell_Dorsal	0.828	0.493	0.953	9.51	9	11.6	0.0004
Sellion_Ment	0.666	0.099	0.929	29.60	6	3.5	0.0054
SubNas_Ment	0.493	0.007	0.869	10.80	6	3.8	0.0221
Subman_Ment	0.619	0.009	0.917	4.85	6	7.8	0.0236
Subman_Ment_Linear	0.656	0.063	0.927	4.81	6	9.3	0.0170
Subnas_Ment_Linear	0.437	0.022	0.840	18.10	6	3.3	0.0134
TopHD_Obt	0.400	11.069	0.988	1.48	2	6.5	0.2960
Trag_Earlobe	0.908	0.652	0.980	21.80	7	8.3	9.95E-05
Trag_Gonion	0.850	0.531	0.966	10.30	7	12.1	0.0003
Trag_Sel	0.936	0.802	0.986	20.70	7	16.4	5.30E-07
Trag_Subman	0.984	0.936	0.997	104.00	6	10.3	1.21E-08

Trag_Subnas	0.958	0.858	0.992	33.00	6	14.2	1.47E-07
TragtoTrag_Contour	0.982	0.938	0.996	82.30	7	13.3	6.79E-10
TragtoTrag_Linear	0.997	0.991	0.999	537.00	7	14.8	1.27E-16

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Time 3		95% CI		F-test			
Landmarks	ICC	Lower Bound	Upper Bound	F-value	*df1*	*df2*	p-value
Alare_Contour	0.952	0.838	0.987	35.70	9	11.3	5.22E-07
BckHD_Glab	0.043	-0.002	0.224	8.74	7	3.2	0.0438
Bizyg_Width	0.689	0.116	0.915	12.00	9	4.7	0.0089
Bizyg_Width_Linear	0.538	0.016	0.856	9.58	9	4.1	0.0203
Cheill_Contour	0.966	0.871	0.993	53.40	7	9.8	4.89E-07
DUMMY1	0.979	0.923	0.995	83.60	8	10.6	1.65E-08
Gonion_Subman	0.923	0.779	0.983	13.00	7	24.0	8.38E-07
Nas_Root_Brdth	0.245	-0.130	0.678	2.51	9	6.1	0.1350
ProNas_AL_Linear	0.950	0.833	0.987	34.90	9	11.2	6.30E-07
ProNas_Alare	0.931	0.778	0.982	24.50	9	11.8	2.67E-06
ProNas_SubNas	0.850	0.616	0.958	7.52	9	26.4	2.22E-05
ProNas_SubNas_Linear	0.938	0.825	0.983	21.50	9	18.6	4.92E-08
Sel_Pronasale	0.967	0.892	0.991	48.60	9	12.7	2.05E-08
Sel_Pronasale_Linear	0.968	0.892	0.992	54.50	9	11.3	5.07E-08
Sell_Dorsal	0.825	0.497	0.952	8.94	9	12.6	0.0003
Sellion_Ment	0.873	0.373	0.978	51.40	6	3.9	0.0012
SubNas_Ment	0.809	0.239	0.970	28.60	5	4.0	0.0032
Subman_Ment	0.329	-0.383	0.836	1.83	6	14.3	0.1630

Subman_Ment_Linear	0.299	-0.443	0.830	1.69	6	15.7	0.1880
Subnas_Ment_Linear	0.521	0.009	0.881	10.20	6	3.9	0.0220
TopHD_Obt	0.871	0.452	0.991	9.40	3	10.9	0.0023
Trag_Earlobe	0.776	0.346	0.943	7.49	8	10.5	0.0019
Trag_Gonion	0.880	0.631	0.977	8.53	6	20.9	9.16E-05
Trag_Sel	0.866	0.641	0.966	8.07	8	25.7	2.10E-05
Trag_Subman	0.949	0.849	0.989	17.60	7	21.5	1.39E-07
Trag_Subnas	0.991	0.972	0.998	131.00	7	19.6	4.32E-15
TragtoTrag_Contour	0.958	0.861	0.991	35.10	7	13.7	1.06E-07
TragtoTrag_Linear	0.989	0.963	0.997	121.00	7	15.7	1.86E-12

*The red highlighted cells in the ICC column are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900. Green highlighted cells in the p-value column are statistically significant p-values <0.05.

Appendix VIII. Inter-rater reliability ICC value comparisons for all times.

	Time 1	Time 2	Time 3
Landmarks		ICC	
Alare_Contour	0.874	0.938	0.952
BckHD_Glab	0.827	0.923	0.043
Bizyg_Width	-0.201	0.641	0.689
Bizyg_Width_Linear	0.413	0.629	0.538
Cheill_Contour	0.886	0.885	0.966
DUMMY1	0.131	0.983	0.979
Gonion_Subman	0.602	0.921	0.923
Nas_Root_Brdth	0.757	0.365	0.245
ProNas_AL_Linear	0.825	0.917	0.950
ProNas_Alare	0.750	0.879	0.931
ProNas_SubNas	0.715	0.674	0.850
ProNas_SubNas_Linear	0.775	0.811	0.938
Sel_Pronasale	0.860	0.959	0.967
Sel_Pronasale_Linear	0.865	0.960	0.968

Sell_Dorsal	0.269	0.828	0.825
Sellion_Ment	0.864	0.666	0.873
SubNas_Ment	0.841	0.493	0.809
Subman_Ment	0.148	0.619	0.329
Subman_Ment_Linear	0.285	0.656	0.299
Subnas_Ment_Linear	0.860	0.437	0.521
TopHD_Obt	0.906	0.400	0.871
Trag_Earlobe	0.753	0.908	0.776
Trag_Gonion	0.783	0.850	0.880
Trag_Sel	0.954	0.936	0.866
Trag_Subman	0.568	0.984	0.949
Trag_Subnas	0.973	0.958	0.991
TragtoTrag_Contour	0.976	0.982	0.958
TragtoTrag_Linear	0.659	0.997	0.989

*The red highlighted cells are ICC values < 0.500, orange cells are values between 0.500-0.750, yellow cells are 0.750-0.900, and green cells are values > 0.900.