DISSERTATION

ANALYSIS OF 3D FACIAL ANTHROPOMETRIC MEASUREMENTS FOR RESPIRATOR FIT OUTCOMES

Submitted by

Kayna Hobbs-Murphy

Department of Environmental and Radiological Health Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2023

Doctoral Committee:

Advisor: John Rosecrance

Bill Brazile
Brooke Anderson
Kristen Morris
Ann Hess

ABSTRACT

ANALYSIS OF 3D FACIAL ANTHROPOMETRIC MEASUREMENTS FOR RESPIRATOR FIT OUTCOMES

Anthropometry is central to the development of efficacious products and environments (i.e., personal protective equipment, clothing, sunglasses, chairs, interior spaces, etc.) used by humans. Three-dimensional (3D) scanning is increasingly common for collecting anthropometric data, as it is faster and less intrusive than traditional manual methods. Additionally, 3D anthropometric methods used to derive facial dimensions provide greater contextual application in the development of respirators and facemasks. More than 2,000 3D facial scans were analyzed to assess measurement reliability and the dimensions of 27 facial features. This research represents the largest sample of 3D facial anthropometrics assessed to date.

The three specific aims of the research included: 1) to assess the intra- and inter-rater reliability of 3D facial measurement methods, 2) to compare the 3D facial anthropometric summary statistics from the present study to relevant summary statistics from manual facial measurements found in the literature, and 3) to assess the presence of differences in 3D facial anthropometrics related to respirator fit, based on demographic factors of gender, race/ethnicity, and age. Post hoc analyses were completed to quantify 3D facial measurement differences between demographic groups (within the larger demographic categories of gender, race/ethnicity, and age group). The most notable results of this research include a) high reliability in 3D measurement data collection methods, b) differences in measurement data summary statistics between 3D and manual methods, and c) significant differences in facial measurements between

demographic categories of gender (Male and Female/Other), race/ethnicity (White, Black, LatinX, Asian, and Other), and age (18-34, 35-54, and 55-72).

ACKNOWLEDGEMENTS

There are many people who deserve thanks for the support they provided to me, the author, during the process of this dissertation being completed.

First and foremost, thank you to Dr. John Rosecrance for the continual support, guidance, expertise, and advice that proved to be invaluable during the entire process of this research. Thank you for always challenging and believing in me. Most importantly, thank you for your kindness to me during my most difficult times.

I would like to thank my funding sources, the NIOSH Mountain and Plains Education Research Center and the High Plains Intermountain Center for Agricultural Health and Safety. Your financial support was invaluable in the completion of this dissertation research.

I would like to thank my committee members for their expertise and aid: Dr. Bill Brazile, Dr. Brooke Anderson, Dr. Kristen Morris, and Dr. Ann Hess. I would also like to thank Isabel Olmedo-Nockideneh for her invaluable aid and support in data collection and analysis efforts.

Lastly, I would like to thank my friends and family for their support and encouragement. Thank you to my dogs, Moose and Walter, for providing me with company during many hours of work at home. Thank you to my husband, Jedidiah Murphy, for all you do. Finally, thank you to my rainbow baby, Sloane Hobbs-Murphy, for motivating me to work hard toward a wonderful future.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Anthropometry, the study of human body shape and size (also referred to as anthropometrics), stems from the field of anthropology. Anthropometry has long been central to the development of products and environments with which human bodies interact (i.e., personal protective equipment, clothing, sunglasses, chairs, interior spaces, etc.). Three-dimensional (3D) scanning continues to become increasingly common for collecting anthropometric data, as it offers a fast, low-contact way to gather measurement data that provide more context to facial surface dimensions than those data collected using manual anthropometric methods. Manual anthropometric data have traditionally been collected using large, bulky anthropometric tools. The use of these anthropometric tools to measure the body requires minutes per each measurement, in that each measurement much be collected from a single anthropometric study participant, one measurement at a time. Furthermore, these anthropometric tools can only measure the body in very specific locations for which the tools were designed. 3D scanning requires minimal contact between researcher and participant, is often a fast process for both parties (requiring only 1-2 minutes to gather entire 3D scan), and creates a "virtual twin" (also referred to as a 3D scan) (Kuehnapfel et al., 2016, p. 1) of the scanned participant. Researchers and professionals can revisit the 3D scan as needed to collect detailed measurement information, which is not feasible with manual measurement techniques. Moreover, 3D methods allow researchers and professionals to gather measurement data that provide more context to facial surface dimensions than manual methods, such as data about body shape, surface distances, and surface contours that are not easy to gather using manual methods alone (Bailar et al., 2007; ben Azouz et al., 2006).

3D scanning has been achieved through continually evolving applications of technology and is consistently being improved for greater precision and accuracy. Of interest to this dissertation work is the Artec Eva 3D (Model Eva, Senningerberg, Luxembourg) scanner by Artec, which uses structured-light sensing to gather a 3D image of the scan object or subject (Artec3D, n.d.). Structured-light 3D scanners (such as the Artec Eva 3D) collect 3D information by projecting light patterns on the surface of a scan subject and deriving information based on the deformation of the light pattern on the subject (Ebrahim, 2014; Heymsfield et al., 2018). This phenomenon is illustrated in Figure 1.1. In the Artec Eva 3D scanner, the camera/sensor and structured light projector are combined into one device (Artec3D, n.d.).

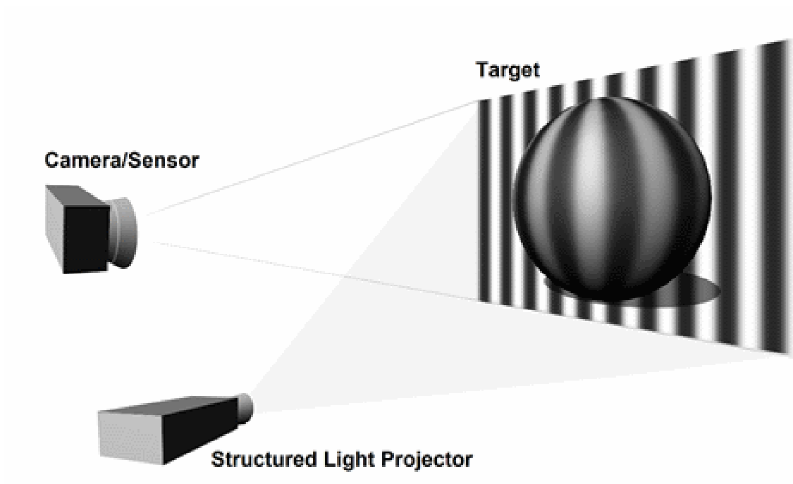

Figure 1.1: Illustration of how structured-light sensor 3D scanning creates 3D scan of an object/subject (Revopoint 3D Technologies Inc., n.d.).

Artec Eva 3D was used to collect 3D facial scans from a sample of 2022 participants by a private company called Human Solutions. These 2022 3D facial scans and relevant measurement data collection software were purchased from Human Solutions for this dissertation research. 3D facial measurement data (also referred to as 3D measurements) collected from the 3D facial scans were analyzed for data collection reliability and facial product design applications. This line of research is valuable to the knowledge base in that it assesses 3D facial measurement data in the context of 3D anthropometric data collection reliability as well as for practical applications such as the design, sizing, and fit of wearable facial products such as respirators. Respirators are wearable facial devices that either remove harmful respirable particulates from the surrounding air or provide clean air to the wearer with a new supply of clean air from a different source (The National Institute for Occupational Health and Safety, 2022). The fit of respiratory protection is imperative to the protection it provides the wearer (Centers for Disease Control and Prevention & The National Institute of Occupational Safety and Health, n.d.; Occupational Safety and Health Administration, 2009). Because facial anthropometric data can provide information about facial shape and size, analyses of these data can be utilized by designers and developers to improve the design, sizing, and fit of respirators and other wearable facial products.

Specific Aims

Specific Aim 1: Intra-Rater and Intra-Rater Reliability of 3D Facial Measurements

Previous studies have aimed to evaluate the reliability and/or accuracy of 3D facial measurements using varying methods of data collection (Aynechi et al., 2011; Düppe et al., 2018; Franco de Sá Gomes et al., 2019; A. J. Kim et al., 2018; Kuehnapfel et al., 2016; Modabber et al., 2016; Wong et al., 2008). Despite previous efforts, ongoing assessments of 3D measurement reliability are needed as the use of 3D scanning and 3D measurements become more accepted to replace manual measurement data collection. Specific aim (*SA1*) sought to assess the intra- and inter-rater reliability of 3D facial measurements gathered by four novice anthropometric raters. In this research, intra-rater reliability (intraRR) is the degree of agreement among collections of a 3D measurement performed on the same subject by a single rater, and inter-rater reliability (interRR) is the degree of agreement among all raters who collect the same 3D measurement on the same subjects.

Four research questions guided SA1:

Specific Aim 1 (SA1): Assess the intra- and inter-rater reliability of 3D facial measurements gathered by four novice anthropometric raters.

> *SA1.RQ1*: What percentage of good to excellent (>0.75 ICC statistic) intra-rater reliability (on average across four raters) can be achieved by the final phase of data collection?
>
> *SA1.RQ2:* What percentage of good to excellent (>0.75 ICC statistic) inter-rater reliability can be achieved by the final phase of data collection?
>
> *SA1.RQ3:* In percentage terms and averaging across four raters, how much does intra-rater reliability improve over two phases of data collection?
>
> *SA1.RQ4:* In percentage terms, how much does inter-rater reliability improve on average over three phases of data collection?

Specific Aim 2: Facial Anthropometrics: A Comparison of Measurements from 3D and Manual Methods

Facial anthropometrics are utilized to design facial products, such as respirators, helmets, hats, glasses, etc. Manual anthropometric methods have long been the standard in anthropometric data collection because they precede the invention of the technology of 3D scanning and therefore the collection of 3D measurements. 3D anthropometric measurements have the potential to provide unique benefits to anthropometric data collection and perhaps be used in place of manual measurements. 3D measurements are typically found to be larger than manual measurements, however, the quantitative discrepancy between manual and 3D measurement methods has not been researched to a large enough degree to fully compare the methods or conclude that all 3D measurement collection processes will result in the same degree of difference from manual measurements. Specific aim (*SA2*) of the research sought to compare the 3D facial anthropometric summary statistics from the present study to relevant summary statistics from manual facial measurements found in the literature (Zhuang et al., 2007; Zhuang & Bradtmiller, 2005).

Two research questions guided SA2:

Specific Aim 2 (SA2): Compare measurements collected using 3D and manual methods.

SA2.RQ1: What are the summary statistics for the 3D measurement data collected?

SA2.RQ2: How do the 3D measurement summary statistics compare to manual measurement summary statistics found in the literature?

Specific Aim 3: Demographic Differences in 3D Facial Anthropometrics Related to Respirator Fit

In response to limitations in seminal facial anthropometric research efforts as well as the increasingly diversifying workforce, researchers have attempted to define the presence or absence of differences in respirator-related facial measurements between different demographic groups. These research efforts have resulted in mixed findings regarding found differences in facial measurements between gender, race/ethnicity, and age groups. No research to date has utilized a large 3D scanned sample to assess differences in 3D facial measurements by three major demographic categories (gender, age, and race/ethnicity). Specific aim (*SA3*) sought to assess the presence of differences in 3D facial anthropometric measurements related to respirator fit, based on demographic factors of gender, race/ethnicity, and age in a sample of 2022 3D scans. This aspect of the research helps determine if people (within the sample population) of different genders, races/ethnicities, or ages can be expected to have different 3D facial measurements relevant to respirator fit.

Three research questions guided SA3:

Specific Aim 3 (SA3): Assess the presence of differences in 3D facial anthropometric measurements related to respirator fit.

SA3.RQ1: Are differences in 3D measurements present between gender groups?

SA3.RQ2: Are differences in 3D measurements present between race/ethnicity groups?

SA3.RQ3: Are differences in 3D measurements present between age groups?

Post Hoc Analyses: Quantification of Significant Differences in 3D Facial Measurements

Post hoc analyses were completed to quantify 3D facial measurement differences between demographic groups (within the larger demographic categories of gender, race/ethnicity, and age group). The results of these analyses are expected to have practical implications for respirator designers and manufacturers.

CHAPTER 2: INTRA-RATER AND INTER-RATER RELIABILITY OF 3D FACIAL MEASUREMENTS

Summary

Three-dimensional (3D) body scanning technology has applications for the collection of anthropometric data in many different fields. The reliability of 3D measurements gathered from 3D scans must be assessed to understand the degree to which this technology can be used in place of manual anthropometric methods. This study assessed the intra- and inter-rater reliabilities of 3D measurement data collected by 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. The results of the present study indicate 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.

Introduction

Three-dimensional (3D) body scanning allows for the rapid, contact-free collection of anthropometric measurement data, which has many applications for medical, scientific, and design-based research and innovation. 3D scanning creates a “virtual twin” (also referred to as a 3D scan) (Kuehnapfel et al., 2016, p. 1) of the scanned person/object, which allows researchers and professionals to revisit the 3D scan as needed to collect detailed measurement information. Measurement data collected from 3D scans (also referred to as 3D measurements) can 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; ben Azouz et al., 2006).

Previous studies have aimed to evaluate the reliability and/or accuracy of 3D facial measurements using varying methods of data collection (Aynechi et al., 2011; Düppe et al., 2018; Franco de Sá Gomes et al., 2019; A. J. Kim et al., 2018; Kuehnapfel et al., 2016; Modabber et al., 2016; Wong et al., 2008). Despite previous efforts, ongoing assessments of 3D measurement reliability are needed as the use of 3D measurements to replace traditional manual measurements becomes more accepted due to benefits regarding time, contact, and measurement context. Authors of a meta-analysis of 3D facial measurement reliability suggest that researchers analyze reliability without manual landmarking the face prior to 3D scanning (Gibelli et al., 2020). Furthermore, as 3D scanning technology advances and diversifies, assessments of reliability should be done more frequently to ensure the appropriateness of 3D scanning in anthropometric research. The specific aim (*SA1*) of this study was to assess the intra- and inter-rater reliability of 3D facial measurements gathered by four novice anthropometric raters using 3D landmarking methods. In the present study, intra-rater reliability (intraRR) is the degree of agreement among collections of a 3D measurement performed on the same subject by a single rater, and inter-rater reliability (interRR) is the degree of agreement among all raters who collect the same 3D measurement on the same subjects. Four research questions guided SA1:

SA1.RQ1: What percentage of good to excellent (>0.75 ICC statistic) intra-rater reliability (on average across four raters) can be achieved by the final phase of data collection?

SA1.RQ2: What percentage of good to excellent (>0.75 ICC statistic) inter-rater reliability can be achieved by the final phase of data collection?

SA1.RQ3: In percentage terms and averaging across four raters, how much does intra-rater reliability improve over two phases of data collection?

SA1.RQ4: In percentage terms, how much does inter-rater reliability improve on average over three phases of data collection?

Methods

3D scan data for this study were purchased from Human Solutions (Human Solutions, n.d.). The company collected 3D facial scans from participants using the handheld Artec Eva 3D (Model Eva, Senningerberg, Luxembourg) structured-light scanner (Artec3D, n.d.). For the purposes of this work, Human Solutions' proprietary 3D scan software, Anthroscan (Version 3.6.1, Kaiserslautern, Germany), was used to collect 3D measurements from each 3D scan.

3D Measurement Collection

The process of gathering 3D measurement data from each 3D scan included digitally clicking through pre-determined facial landmarks (illustrated in Figure 2.1, described in Table 2.1), and subsequently allowing a custom-made measurement software program, referred to as a measurement 'wizard', in Anthroscan (Version 3.6.1, Kaiserslautern, Germany) to generate length and contour distances between relevant landmark locations. In this way, landmarking was done in 3D (3D landmarking) as opposed to landmarking done on the face prior to 3D scanning (manual landmarking). The measurement wizard was developed by Human Solutions specifically for the present study. Figure 2.2 illustrates the 27 3D facial measurements collected by the Anthroscan wizard, which were used to assess intraRR and interRR in this study. Table 2.2 describes the name of each measurement, whether the measurement was collected in a linear (direct from point to point) and/or contour (over the surface of the face) fashion, and the abbreviated measurement name. 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 wizard specified that the measurement is collected by going through a third middle specified landmark (such as the Sellion landmark for TrTr_C) or is collected along a specific axis (X, Y, or Z).

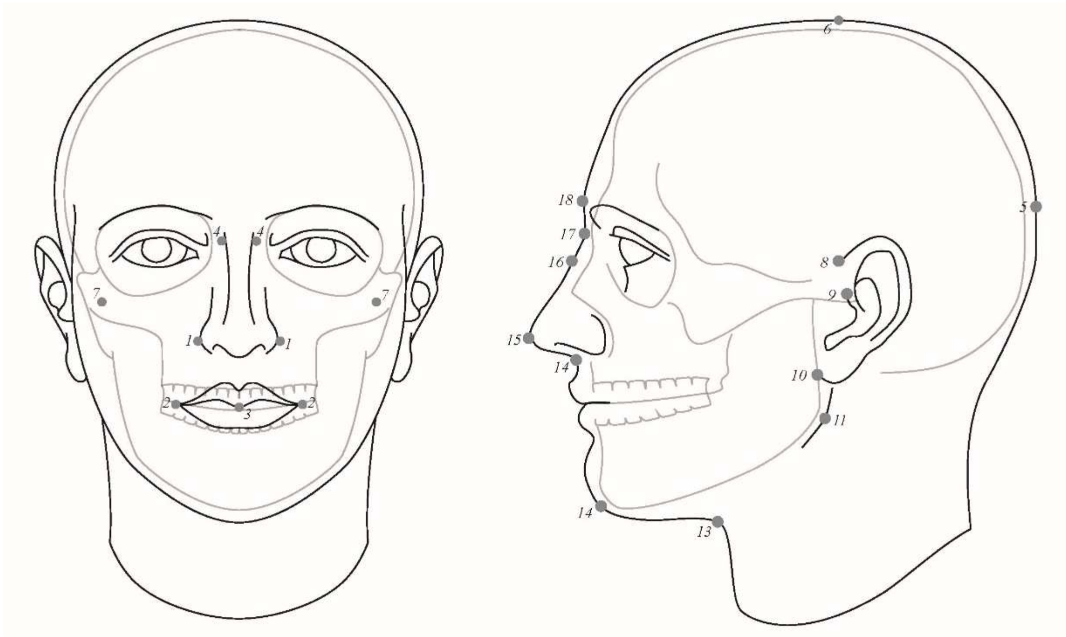

Figure 2.1: Illustration of 3D facial landmarks placed on each 3D scan, used to collect 3D measurements seen in Figure 2.2.

Table 2.1.
3D landmark names (corresponding to Figure 2.1) and an indication of single (center) or left and right (L&R) marking status.

Number in Figure 2.1	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

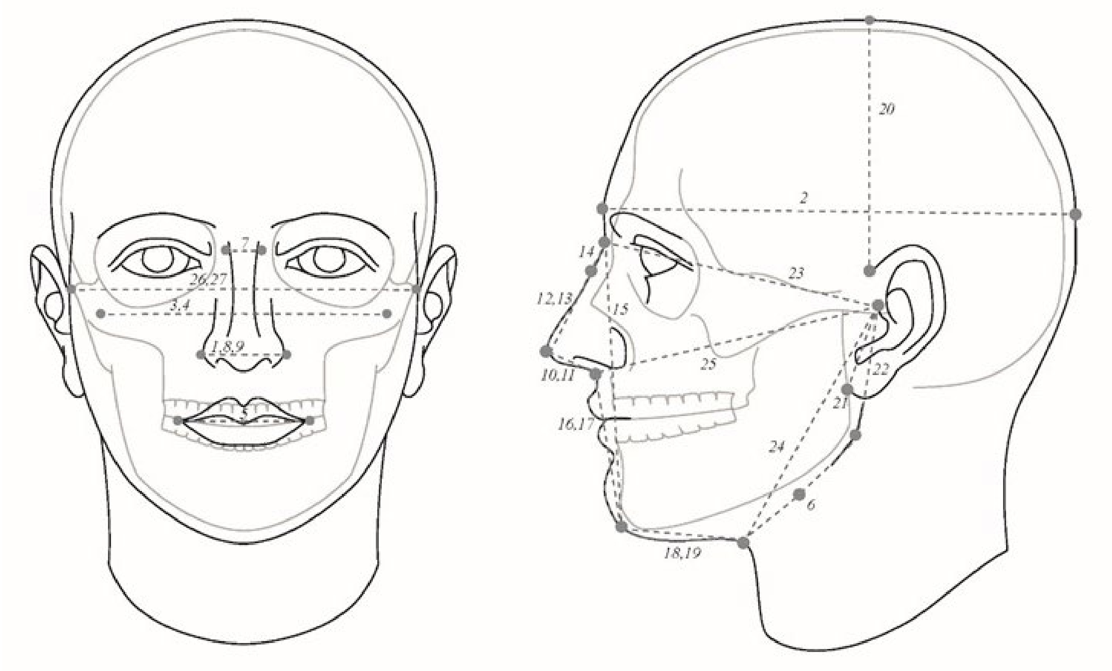

Figure 2.2: Illustration of 3D facial measurements collected from each 3D scan.

Table 2.2.

Measurement names (corresponding to Figure 2.2), measurement type (linear, contour, or both), and abbreviated measurement name.

Number in Figure 2.2	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

Rater Training

Previous research has suggested that providing anthropometric training to raters prior to data collection may improve reliability scores (Düppe et al., 2018). Four raters (Rater A, Rater B, Rater C, and Rater D) were involved in 3D measurement data collection for this research. Before raters collected any 3D measurement data, they were provided with a guide to the 3D facial landmarks and measurements to familiarize themselves with. Raters were asked to watch a video tutorial on how to place 3D landmarks and collect 3D measurements in Anthroscan. In the video tutorial, the primary investigator explained how to use the mouse to zoom in, move the 3D scan to identify the 3D landmark, place a 3D landmark at the appropriate XYZ coordinates, and check visually each 3D measurement for accuracy. 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. If a landmark could not be confidently identified and placed, the corresponding 3D measure/s were not obtained from that 3D scan. Access to the training video and written instructions were provided to the raters, allowing them to revisit instructions throughout data collection. Lastly, raters met with senior research staff prior to starting the data collection process, to address questions.

When raters felt confident in their understanding of the 3D facial landmarks and 3D measurement data collection process, they were asked to take a quiz testing their ability to recognize the 3D landmarks. If raters did not receive full credit on the 3D landmark quiz, they reviewed the training materials and re-attempted the quiz until full credit was received. Once raters received full credit on the 3D landmark quiz, they were able to begin 3D landmarking facial scans for training and subsequently for 3D measurement data collection. Raters were asked to fill out a questionnaire asking which 3D landmarks were most difficult to place on each 3D scan and meet with the senior research staff again for additional training after completing the first phase of data collection (described below).

Three-Phase Data Collection

3D measurement data was collected by the same set of raters throughout the present study. 3D measurement data were collected from a total of 30 3D scans (i.e., 30 scan subjects) in a three-phase data collection process, allowing for analysis of the rate of improvement in intraRR and interRR over time as the result of increased training and experience. Phase 1 of the study assessed intraRR and interRR by asking all four raters to collect 3D measurements from 10 scans (subjects #1-10), three times each. Phase 2 of the study assessed only interRR by asking all four raters to collect 3D measurements from 10 new scans (subjects #11-20), only one time each. Lastly, Phase 3 of the study again assessed intraRR and interRR by asking all four raters to collect 3D measurements from 10 new scans (subjects #21-30), three times each. Table 2.3 provides an outline of the three phases of data collection.

Table 2.3.
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, collection of 3D measurement data from random scans #1-10 was repeated a 2nd time	Yes, collection of 3D measurement data from random scans #1-10 was repeated a 3rd time
Phase 2	interRR	Collect 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	Collect 3D measurement data from random scans #21-30 (1st time)	Yes, collection of 3D measurement data from random scans #21-30 was repeated a 2nd time	Yes, collection of 3D measurement data from random scans #21-30 was repeated a 3rd time

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 & 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 (see Table 2.3 above). In the case of a missing measurement value for one or more raters, the data point was assigned as 0 (millimeters) as opposed to a missing value, to allow for assessment of agreement in landmark placement (or lack thereof) for all measurements. Based on published ICC statistic guidelines (Koo & Li, 2016), a two-way mixed effects model, with "mean of the k raters" type and "absolute agreement" data definition was chosen (pp. 157-159). Analyses were conducted using RStudio (R Core Team, 2022d) with packages tidyverse, irr, lpSolve (Berkelaar & others, 2022; Gamer et al., 2019; Wickham et al., 2019). The irr package was chosen over other available packages that analyze ICC statistics because it allowed the researcher to denote ICC characteristics in coding, such as the "two-way" model and the "agreement" definition.

Results

ICC statistics were calculated using 3D measurement data gathered in metric units. Guidelines for evaluating reliability, as related to ICC statistics, followed what is found in the ICC literature which guided this work (Koo & Li, 2016). ICC statistics below 0.50 (including negative values) were considered to indicate poor reliability (Liljequist et al., 2019; Taylor, 2009). ICC statistics between 0.50 and 0.75 were considered to indicate moderate reliability. ICC statistics between 0.75 and 0.90 were considered to indicate good reliability, and ICC statistics greater than 0.90 were considered to indicate excellent reliability in this study. The 95% confidence interval (CI) for each ICC statistic is presented in the results, allowing for reliability to be assessed on a range of reliability categories, if applicable (Koo & Li, 2016). Table 2.4 quantifies ICC statistic ranges and the corresponding reliability demonstrated in this study.

Table 2.4.
ICC statistic range and corresponding reliability, following guidelines from Koo & Li, 2016.

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

Phase 1

IntraRR. In Phase 1, the range of intraRR ICC statistics was -0.28 to 0.99. Negative ICC statistics indicate very poor reliability and can be more common when the number of data points analyzed is small (Liljequist et al., 2019; Taylor, 2009), which is true in the present study. Figure 2.3 illustrates each rater's Phase 1 intraRR ICC statistic and 95% CI by abbreviated measurement location, with dashed lines indicating ICC statistical range limits for reliability categories presented in Table 2.4. Each rater had at least one ICC statistic indicating poor intraRR. 3D measurements with poor intraRR were mainly 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.

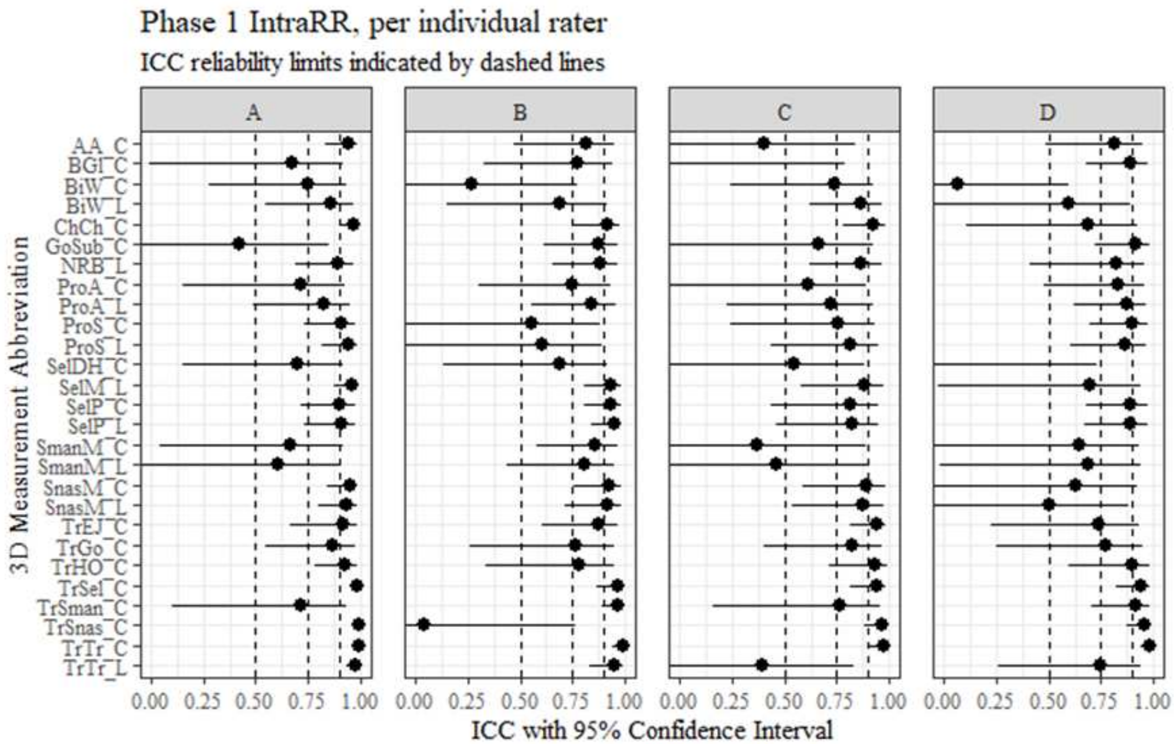

Figure 2.3: Phase 1 intraRR ICC statistics and 95% confidence interval, for each individual rater (Rater A, Rater B, Rater C, and Rater D). Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.50-0.75, good 0.75-0.90, and excellent > 0.90).

InterRR. In Phase 1, the range of interRR ICC statistics was -0.20 to 0.98. Figure 2.4 illustrates Phase 1 interRR ICC statistics and 95% CI by abbreviated measurement location, 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 50.00%, and excellent interRR for 14.81% of the 27 3D measurements.

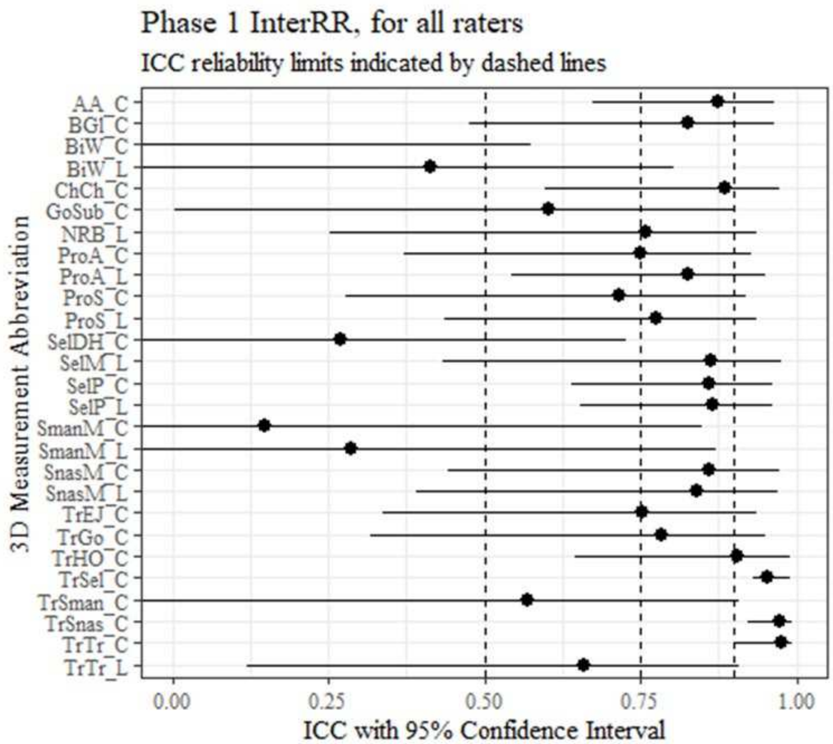

Figure 2.4: Phase 1 interRR ICC statistics and 95% confidence interval, for all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.50-0.75, good 0.75-0.90, and excellent > 0.90).

Phase 2

InterRR. In Phase 2, the range of interRR ICC statistics was 0.37 to 0.99. Figure 2.5 illustrates Phase 2 interRR ICC statistics and 95% CI by abbreviated measurement location, 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.

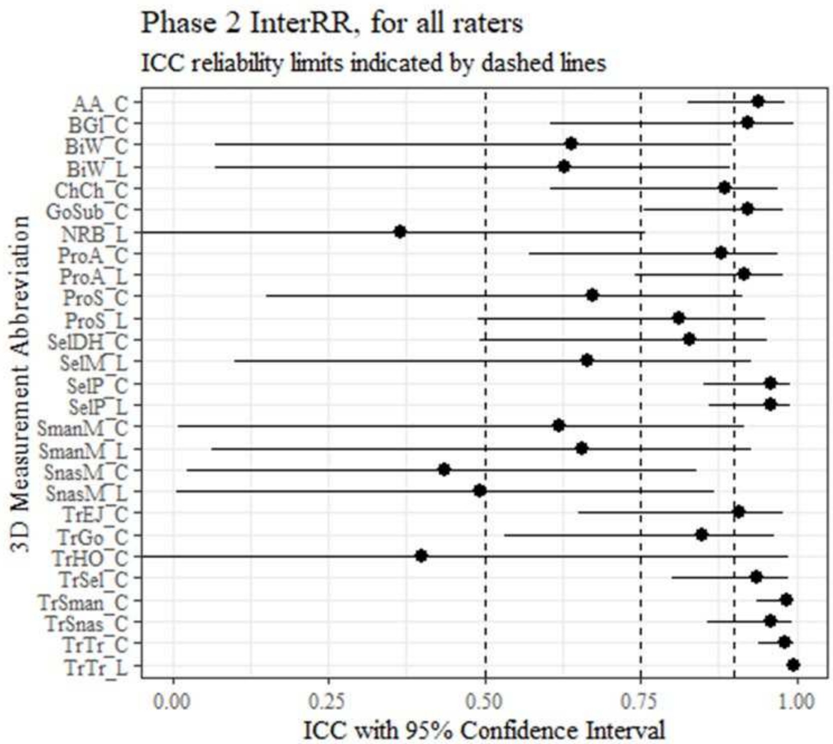

Figure 2.5: Phase 2 interRR ICC statistics and 95% confidence interval, for all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.50-0.75, good 0.75-0.90, and excellent > 0.90).

Phase 3

IntraRR. In Phase 3, the range of intraRR ICC statistics was 0.38 to 0.99. Figure 2.6 illustrates each rater's Phase 1 intraRR ICC statistic and 95% CI by abbreviated measurement location, 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.

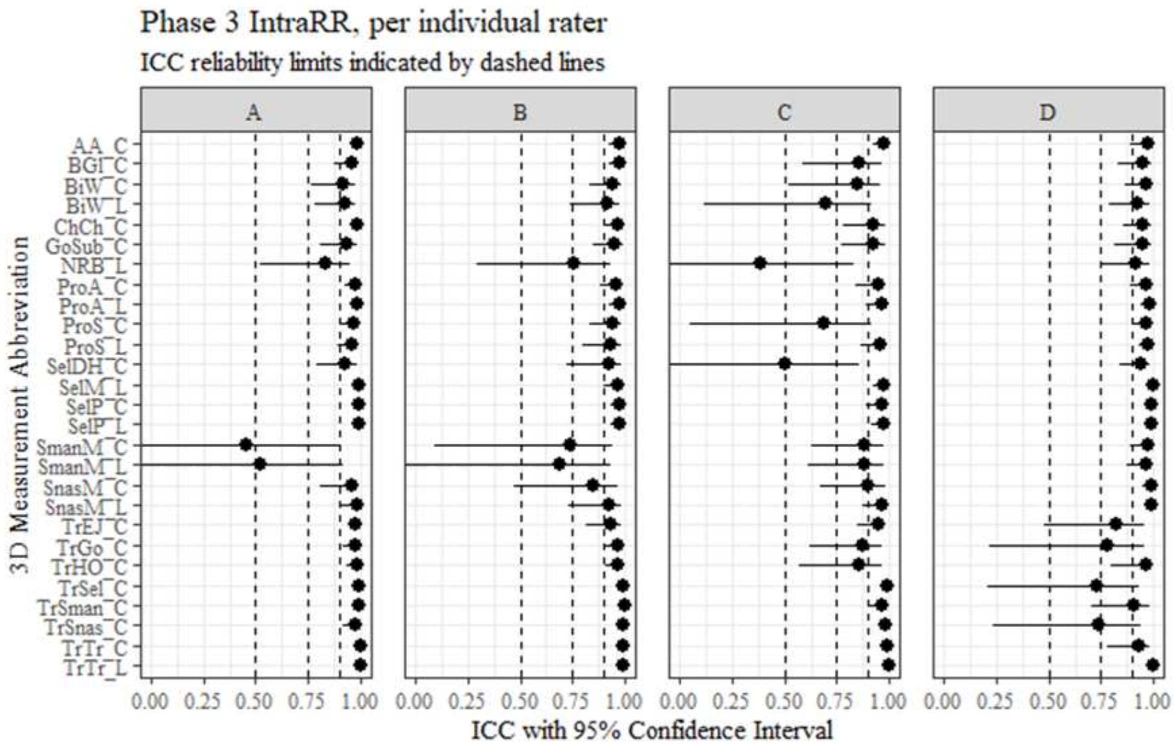

Figure 2.6: Phase 3 intraRR ICC statistics and 95% confidence interval, for each individual rater (Rater A, Rater B, Rater C, and Rater D). Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

InterRR. In Phase 3, the range of interRR ICC statistics was 0.04 to 0.99. Figure 2.7 illustrates Phase 3 interRR ICC statistics and 95% CI by abbreviated measurement location, 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 for 11.11%, good interRR for 29.63%, and excellent interRR for 44.44% of the 27 3D measurements.

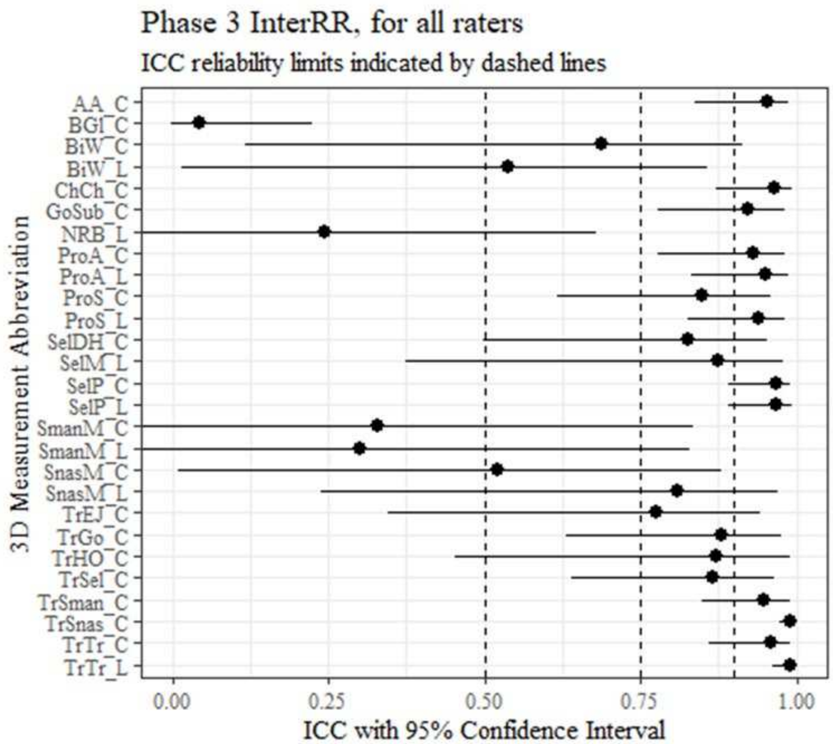

Figure 2.7: Phase 3 interRR ICC statistics and 95% confidence interval, for all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.50-0.75, good 0.75-0.90, and excellent > 0.90).

Discussion

Specific Aim 1 evaluated intra- and inter-rater reliability of 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 (Figure 2.1, Table 2.1) resulted in the collection of 27 3D facial anthropometric measurements per scan, per phase (Figure 2.2, Table 2.2). In response to SA1.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 SA1.RQ2, 74.07% good to excellent (>0.75 ICC) inter-rater reliability was achieved by Phase 3.

In the only example of relevant and comparable research utilizing 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 results of 72.73% excellent intraRR and 54.55% excellent interRR (Franco de Sá Gomes et al., 2019). In the present study, the raters averaged 90.74% intraRR (over the four raters) and achieved 74.07% interRR above 0.75 (excellent cutoff standards by Franco de Sá Gomes et al., 2019) by the third phase of data. Considering that raters the in the present study collected more than double the number of 3D measurements collected by Franco de Sá Gomes et al. (2019) (27 vs. 11), intraRR and interRR in the present study indicate very high averages of intraRR and interRR compared to those found in previous relevant literature.

Improvements in IntraRR

Figures 2.8-2.15 shows each rater's Phase 1 and Phase 3 ICC statistics and intraRR respectively compared (Rater A: Figures 2.8 and 2.9, Rater B: Figures 2.10 and 2.11, Rater C: Figures 2.12 and 2.13, Rater D: Figures 2.14 and 2.15). Each figure caption describes the percent of improved, constant, and/or decreased intraRR values for each rater. In response to SA1.RQ3, ICC scores for intra-rater reliability improved by 58.34% (averaged across four raters) over the two phases in which intraRR was assessed (Phase 1 and Phase 3). Compared to Phase 1, there was a higher proportion of excellent reliability for all 3D measurements in Phase 3. Furthermore, Phase 3 results held fewer ICC statistics in the poor intraRR range for all raters. These findings indicate that increased training and experience with the 3D landmarking and 3D measurement collection process may have improved each rater's intraRR in this study. In different terms, training to improve, and/or experience with, the data collection process may have improved each rater's ability to collect similar measurement data from each 3D scan. Likewise, previous anthropometric research endeavors have attributed improved intraRR to multiple training sessions for raters (Androutsos et al., 2020; de Miguel-Etayo et al., 2014).

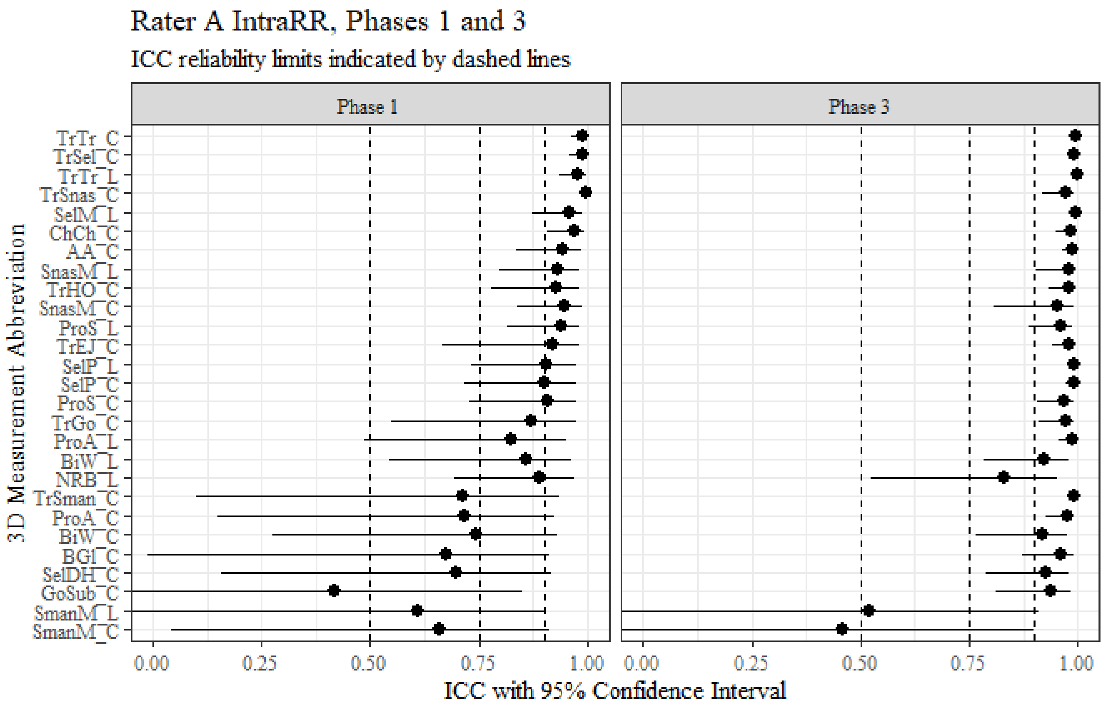

Figure 2.8: Phases 1 and 3 ICC statistics and 95% confidence interval for Rater A. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

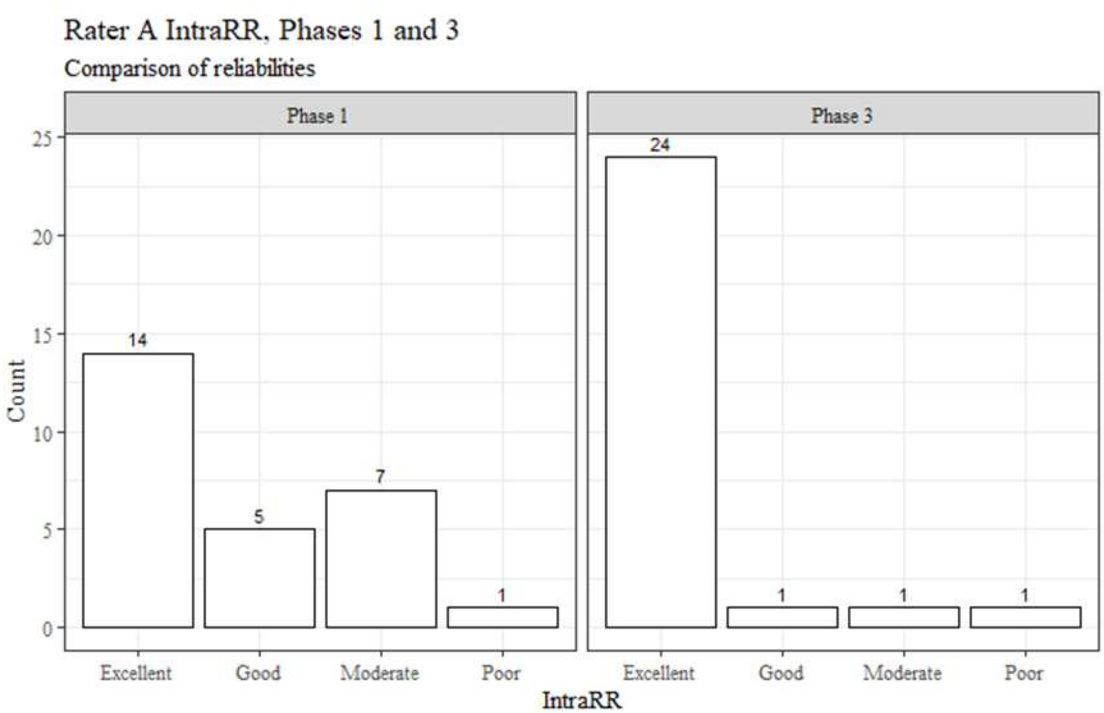

Figure 2.9: Comparison of intraRR at Phase 1 and Phase 3 for Rater A. IntraRR was improved for 10 (37.04%), remained constant for 16 (59.26%), and decreased for 1 (3.70%) of 27 3D measurement locations.

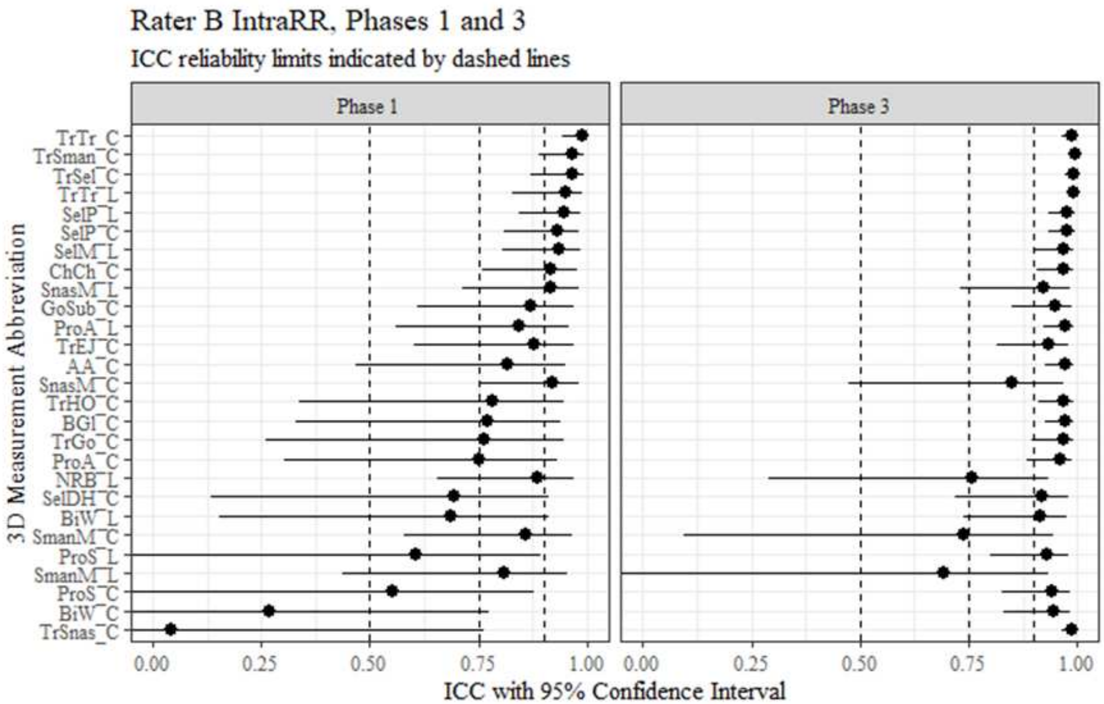

Figure 2.10: Phases 1 and 3 ICC statistics and 95% confidence interval for Rater B. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

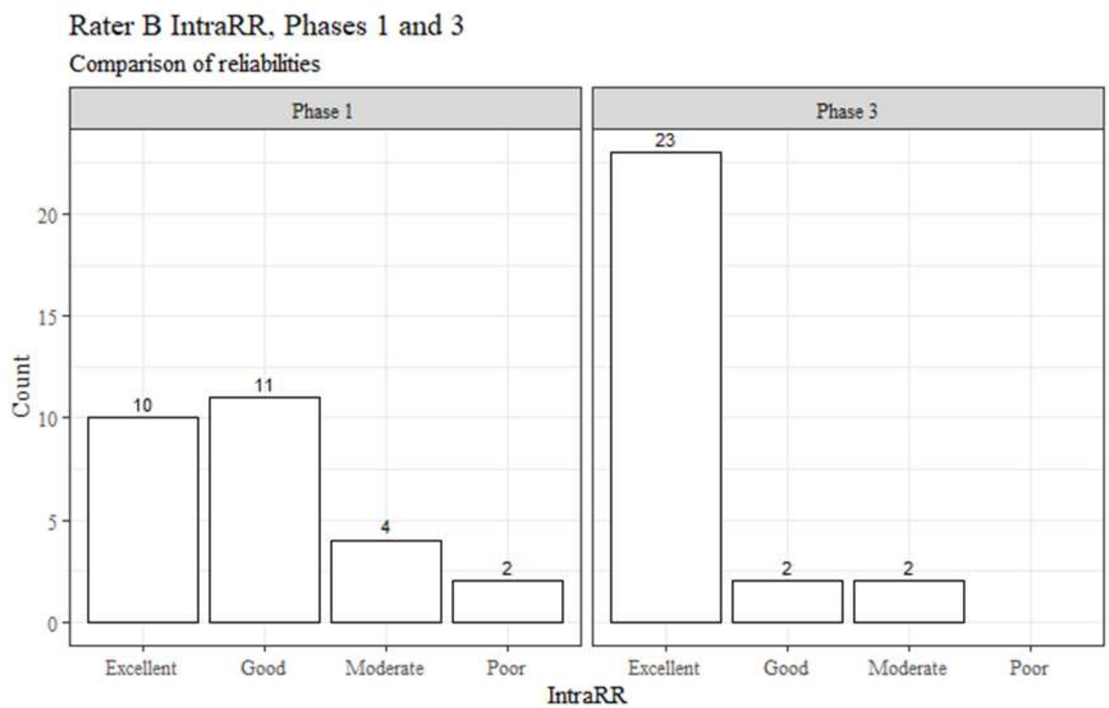

Figure 2.11: Comparison of intraRR at Phase 1 and Phase 3 for Rater B. IntraRR was improved for 16 (59.26%), remained constant for 10 (37.04%), and decreased for 1 (3.70%) of 27 3D measurement locations.

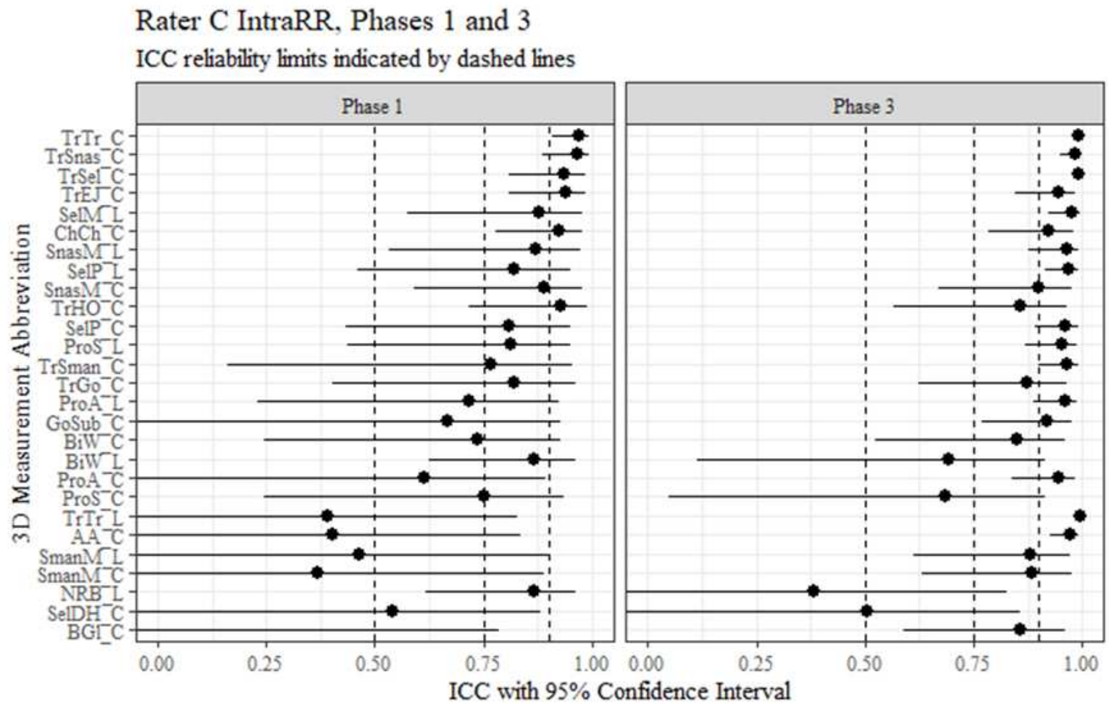

Figure 2.12: Phases 1 and 3 ICC statistics and 95% confidence interval for Rater C. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

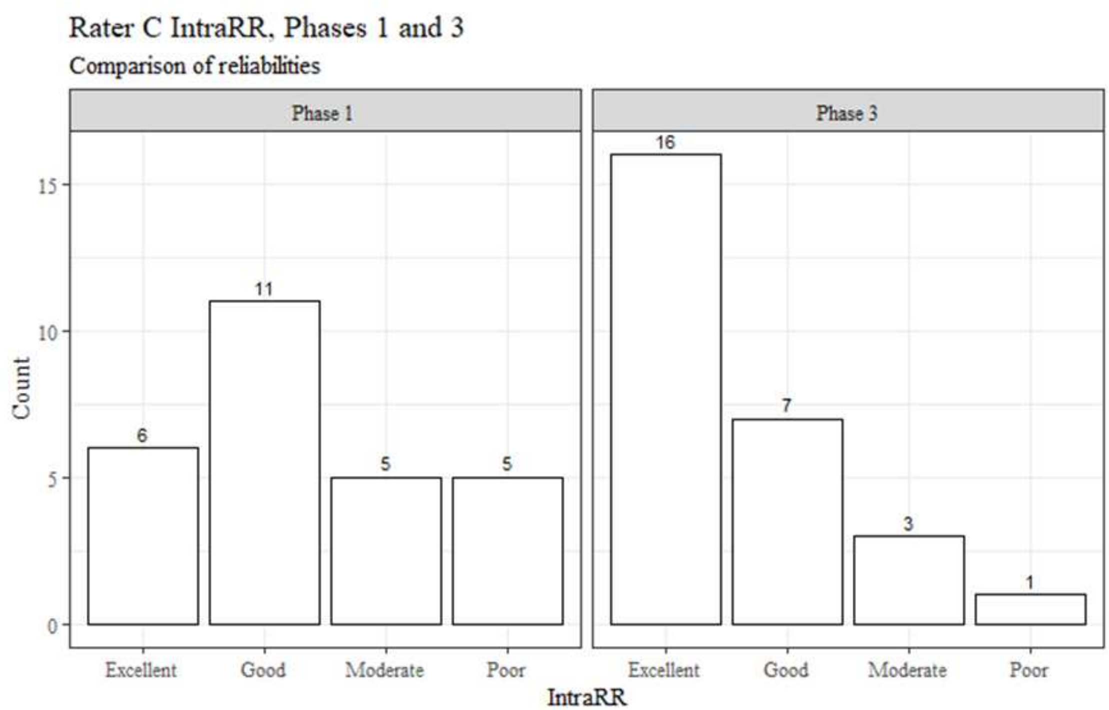

Figure 2.13: Comparison of intraRR at Phase 1 and Phase 3 for Rater C. IntraRR was improved for 16 (59.26%), remained constant for 8 (29.63%), and decreased for 3 (11.11%) of 27 3D measurement locations.

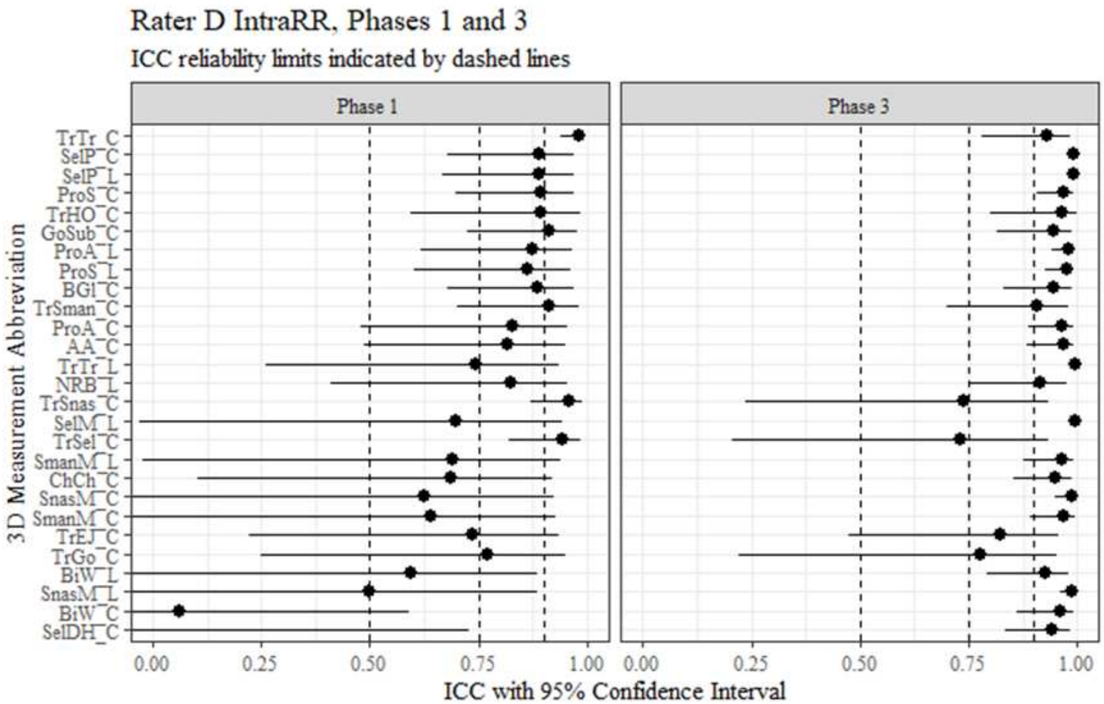

Figure 2.14: Phases 1 and 3 ICC statistics and 95% confidence interval for Rater D. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

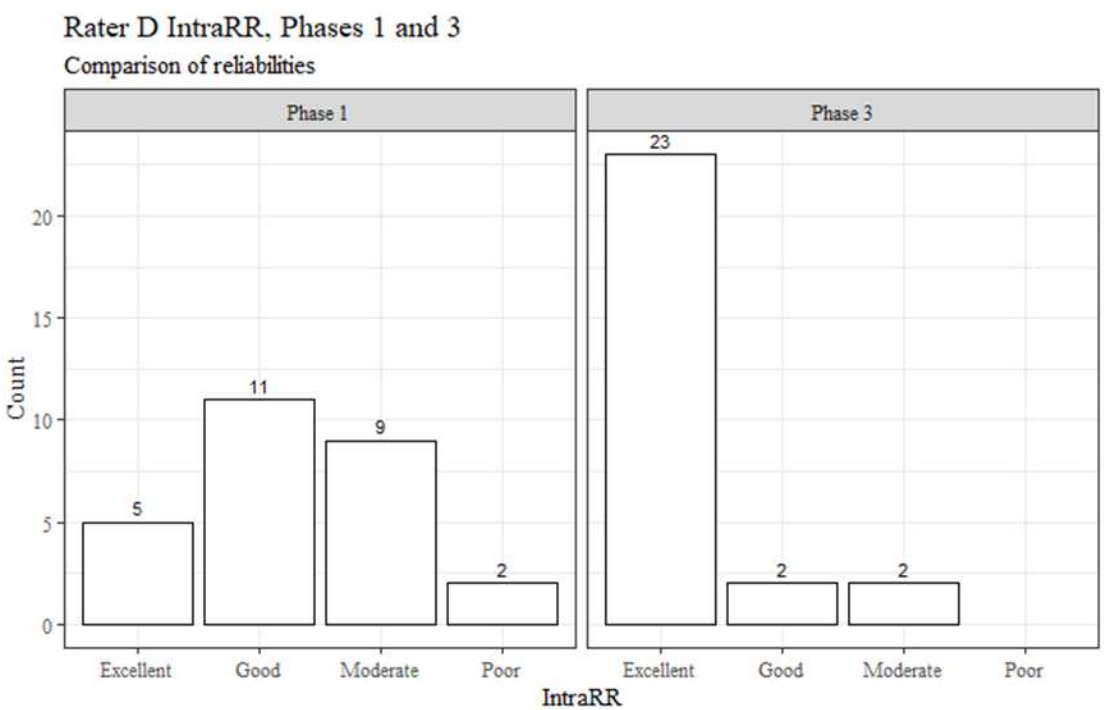

Figure 2.15: Comparison of intraRR at Phase 1 and Phase 3 for Rater D. IntraRR was improved for 21 (77.78%), remained constant for 4 (14.81%), and decreased for 2 (7.41%) of 27 3D measurement locations.

Improvements in InterRR

Figures 2.16 and 2.17 compare the interRR ICC statistics across all three phases at each 3D measurement location. In response to SA1.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 decrease in the overall proportion of ICC statistics indicating poor interRR.

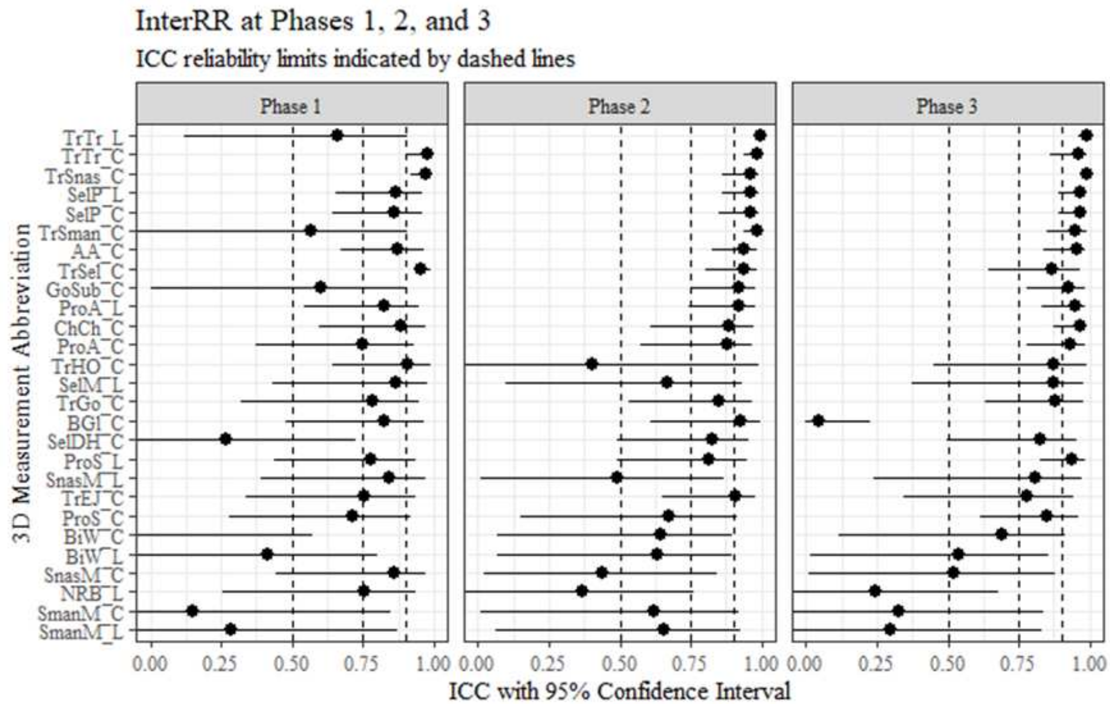

Figure 2.16: Phases 1, 2, and 3 ICC statistics and 95% confidence interval for all raters. Dashed lines represent ICC statistic limits as related to reliability class (poor < 0.50, moderate 0.5-0.75, good 0.75-0.90, and excellent > 0.90).

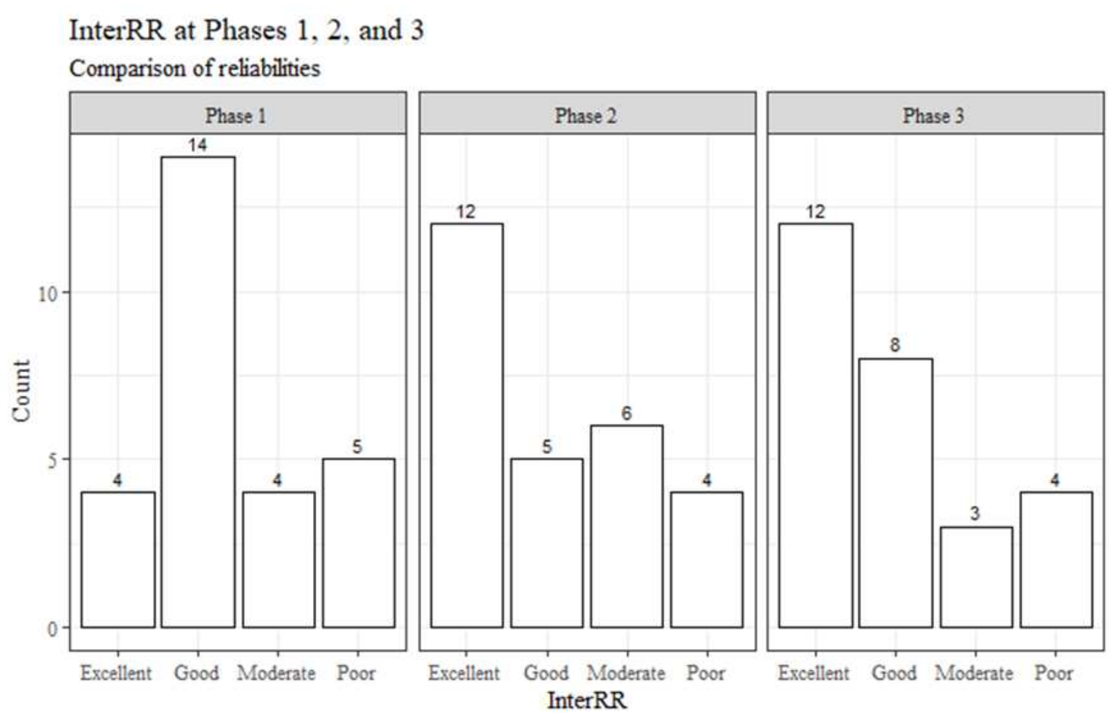

Figure 2.17: Comparison of interRR at Phases 1, 2, and 3 for all raters.

Generally, improved ICC statistics over each phase would suggest that increased rater experience had a positive impact on reliability between all. Because ICC statistics indicated more dramatic improvements in intraRR between phases, it is possible that raters became more skilled 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 more 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 research has attributed improved interRR to multiple group training sessions for raters (Androutsos et al., 2020; de Miguel-Etayo et al., 2014).

3D Landmark Placement

Overall, the results of this study suggest that increased 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. However, as mentioned, group training for 3D landmark placement may have helped to further raise interRR ICC statistics between Phases 2 and 3 in this study. 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. Research regarding the reliability and variation of 3D landmark placement 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 (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, and the inability to palpate the face (as one would when using manual landmarking).

Previous literature has suggested that manually landmarking the participant's face with visual dots (using marker, eyeliner, or small stickers) prior to the 3D scan process can allow for increased reliability in 3D gathered measurements, given that data collectors can palpate bony landmarks more accurately (Aynechi et al., 2011; 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 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. Despite the benefits of manual landmarking prior to 3D scanning, researchers have acknowledged that more research is needed to assess 3D measurement reliabilities in the case of 3D landmarking, as manual landmarking on the face requires more time and closer contact with scan participants (Franco de Sá Gomes et al., 2019; Gibelli et al., 2020).

Limitations

It is possible that an assessment of the reliability of 3D landmark placement (in X, Y, Z axis coordinate measurements) could provide a better understanding of overall 3D measurement reliability. The present study utilized a custom-made measurement wizard in Anthroscan, developed by Human Solutions specifically for this research, which inherently limits the generalizability of this study's findings. However, measurement wizards like the one used in the present may be necessary for future work that utilizes 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.

Because the 3D scan data used in this work were collected by and purchased from Human Solutions, the researchers did not have any hand in assuring the quality of each scan. For example, researchers could not assure 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.

Conclusions

3D scanning is becoming more popular for use in anthropometric research and industry. As researchers and practitioners implement 3D scanning in their works, assessments of reliability allow for a better understanding of the strengths and limitations of the tool prior to implementation. In address of SA1, the present study 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 research studies have assessed the reliabilities of multiple raters of 3D measurement data by using 3D landmarking only. The results of the present study indicate 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. Future research is needed to continually assess the reliability of 3D landmarking, 3D measurement data collection, and 3D scanning as a tool for anthropometric surveying.

CHAPTER 3: 3D FACIAL ANTHROPOMETRICS: A COMPARISON OF MEASUREMENTS FROM 3D AND MANUAL METHODS

Summary

Facial anthropometric measurements are used by designers, researchers, and product developers to properly design and size facial products. Three-dimensional (3D) body scanning technology offers a fast, low-contact, way to gather facial measurement data that provide more context to facial surface dimensions than data collected using manual anthropometric data collection methods. To allow for a better understanding of the applicability of each method, the aim of this study was to compare measurements collected using 3D methods and manual methods found in the literature. In comparison to manual facial measurements, the 3D facial measurements summarized in this research provided more information regarding the surface contour lengths of facial measurement locations. However, precision (ascertained through standard deviation) was found to be generally lower for 3D measurements compared to manual measurements in the present study.

Introduction

Anthropometrics, defined as “the science that defines the physical measures of a person’s size, form, and functional capacities” can help engineers and designers understand how to size and build environments and wearable products (Gordon et al., 2014; The National Institute for Occupational Safety and Health & Centers for Disease Control and Prevention, n.d.). Of particular interest in the field of anthropometrics is the rise in popularity of three-dimensional (3D) scanning, which creates a “virtual twin” of the scanned anatomical area (Kuehnapfel et al., 2016, p. 1). 3D scans provide researchers with technological 3D objects that are representative of the body and able to be referred to indefinitely. 3D scanning can be a faster anthropometric data collection process for the participants/subjects (compared to manual methods) and requires minimal contact between researcher and participant (Gordon et al., 2014). Moreover, 3D anthropometry allows researchers and professionals to gather contextual anthropometric data about body shape, surface distances, and surface contours that are not easy to gather using manual methods alone (Bailar et al., 2007; ben Azouz et al., 2006).

Due to the low participant contact, the speed of 3D scan collection, and the nature of measurement data, 3D anthropometric measurements (referred to as 3D measurements or 3D measurement data) have the potential to provide unique benefits to anthropometric data collection and perhaps be used in place of anthropometric data collected manually. However, manually-collected anthropometric measurements (referred to as manual measurements) have long been the standard in anthropometric data collection because they precede the technology of 3D scanning and therefore the collection of 3D measurements. Furthermore, 3D measurements are typically found to be larger than manual measurements because the human body tends to be compressed during the palpation of soft tissues during manual measurement techniques (Bailar et al., 2007; Gordon et al., 2014; Han et al., 2010; Heymsfield et al., 2018). 3D scan software cannot compress the body; therefore, 3D measurement data are gathered from the surface of the 3D scan without compression. This lack of compression in the collection of 3D measurement data collection can sometimes result in large measurement differences between manual and 3D measurements. Han et al. (2010) found that their 3D measurements differed from their manual measurements to a degree that was larger than allowable errors specified by reputable anthropometric survey references (Clauser et al., 1988; Gordon et al., 1989, 2014).

The literature has sparse information regarding summary statistics of 3D measurement data, specifically regarding 3D measurements of the face. Facial anthropometrics are utilized to design facial products, such as respirators, helmets, hats, glasses, etc. The unique benefits of 3D measurements may allow a better understanding of facial surface anthropometrics and therefore better design of facial products. However, the quantitative discrepancy between manual and 3D measurement methods has not been researched to a large enough degree to fully compare the methods or conclude that the processes of 3D measurement collection will result in the same degree of difference from manual measurements for all measurements. The specific aim (*SA2*) of present study was to compare the 3D facial anthropometric summary statistics from the present study to relevant summary statistics from manual facial measurements found in the literature (Zhuang et al., 2007; Zhuang & Bradtmiller, 2005). Two research questions guided SA2:

SA2.RQ1: What are the summary statistics for the 3D measurement data collected?

SA2.RQ2: How do the 3D measurement summary statistics compare to manual measurement summary statistics found in the literature?

Methods

3D scans for this study were purchased from Human Solutions (Human Solutions, n.d.). The company collected 3D facial scans from 2022 participants using the handheld Artec Eva 3D (Model Eva, Senningerberg, Luxembourg) structured-light scanner (Artec3D, n.d.). Human Solutions' proprietary 3D scan software, Anthroscan (Version 3.6.1, Kaiserslautern, Germany), was used to collect 3D measurements from each 3D scan. Self-reported demographic data was collected at the time of 3D scanning for each of the 2022 participants.

3D Measurement Collection

Collection of 3D facial measurement data was completed by a team of four trained but novice raters. The specifics of 3D facial measurement data collection for the 2022 3D facial scans are described in Chapter 2, which assessed the rater-based reliabilities of gathered 3D measurements before 3D measurement data collection on the entire sample of 2022 participants/scan subjects began. Figure 3.1 illustrates the 27 3D facial measurements (or 3D measurement locations) collected by the four raters using Anthroscan software. Table 3.1 describes the name of each 3D measurement, whether the measurement was collected in a linear (direct from point to point) and/or contour (over the surface of the face) fashion, and the abbreviated measurement name.

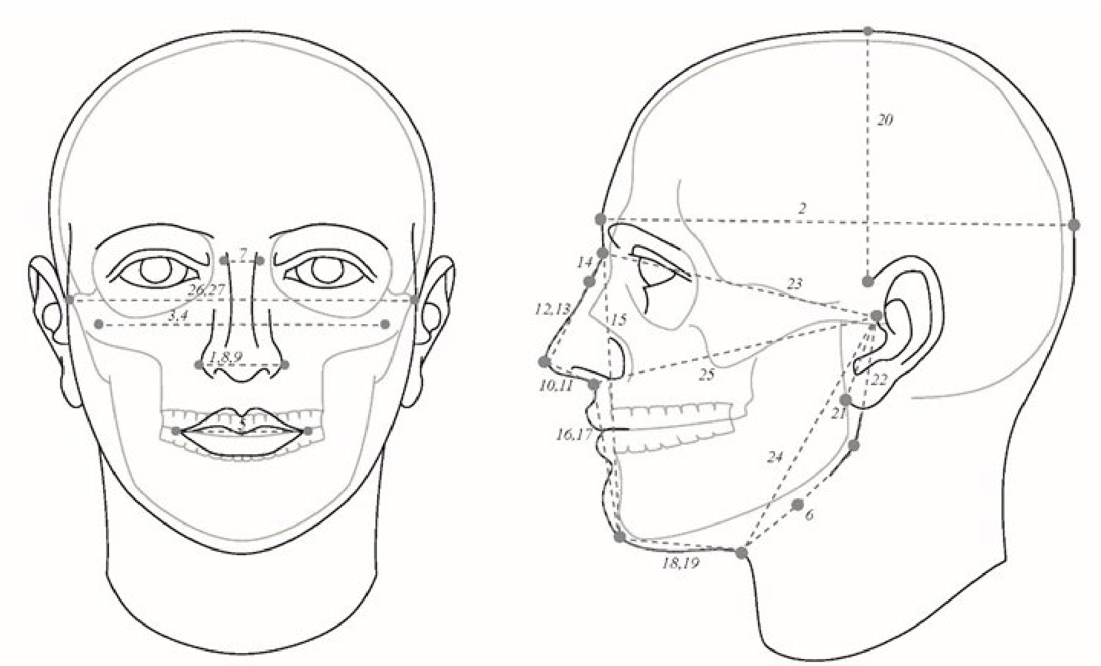

Figure 3.1: Illustration of 3D facial measurement locations. 3D measurement data were collected from these measurement locations for each 3D scan (n=2022).

Table 3.1.
Measurement names (corresponding to Figure 3.1), measurement type (linear, contour, or both), and abbreviated measurement name.

Number(s) in Figure A	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

Statistical Analysis

Summary statistics. 3D measurement summary statistics of facial anthropometrics were calculated for the entire study population as well as separately for Male and Female gender groups. Summary statistics that were calculated included minimum, maximum, mean, standard deviation, standard error, and selected percentile data based on z-scores (5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles). Summary statistics were calculated using R (R Core Team, 2022a) and R packages tidyverse, readxl, extrafont, flextable, and forcats (Chang, 2022; Gohel, 2022; Wickham, 2021, 2022; Wickham & Bryan, 2022). Because some of the 27 3D measurement locations had missing data, the sample size varied for each 3D measurement. The sample size for each 3D measurement location is reported with the summary statistics. Figure 3.2 illustrates the count of missing values for each 3D measurement location.

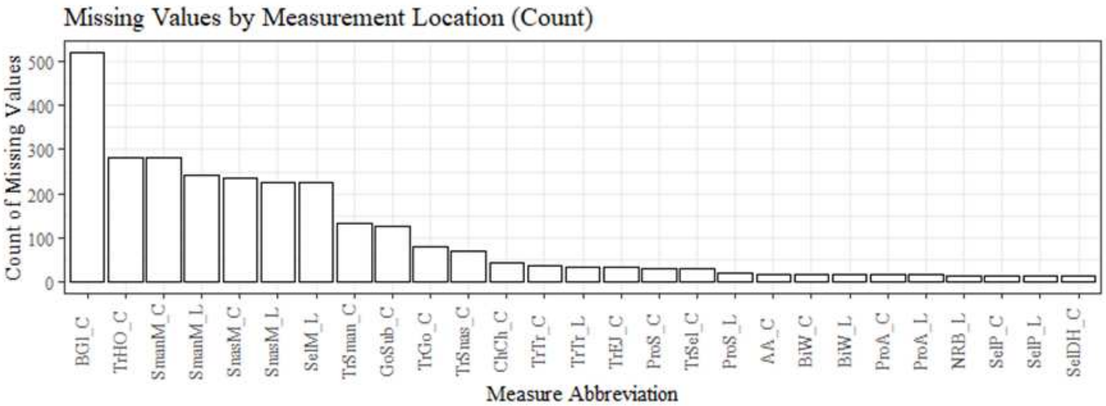

Figure 3.2: Count of missing values per measurement location (n=2022 for each measure).

3D vs manual. This research utilized 3D measurement data, which, unlike manually-collected measurement data, has the capacity to include contour measurements that are measured over the surface of the face (Bailar et al., 2007). However, comparisons to summary statistic data found in other studies are needed to qualify the similarities and differences in 3D and manual measurements. Therefore, summary statistics of 3D measurement locations are compared with similar measurements from previous research (Zhuang & Bradtmiller, 2005). Research by Zhuang and Bradtmiller (2005) was selected for comparison, as it was found to be the only published research that provided summary statistics for manual measurements similar to 3D measurements collected in this study. Table 3.2 describes the manual measurement collected by Zhuang and Bradtmiller (2005), identifies the similar 3D measurement collected in this research, and describes differences in measurement collection methods (to be referred to as measurement method differences or MMDs). The expected effects of the MMDs on the summary statistics of each similar measure are also presented in Table 3.2.

A comparison of the precision of each method (3D vs. manual) was done by assessing differences in variance using two-tailed F-tests for two standard deviations (3D standard deviation vs. manual standard deviation) for each measurement (separately for each gender), with a significance level of $p<0.05$. Because it can be expected that measurements gathered from any given population would be normally distributed, some anthropometric researchers have suggested that standard deviation can be used as a quality assessment tool and standard deviation cut-offs should inform data exclusion in anthropometric surveys (Mei & Grummer-Strawn, 2007). Other researchers argue that standard deviation should not be used as a quality assessment measure in anthropometry, due to uncontrolled variance in any given sample population (Sandler, 2021). Furthermore, standard deviations in anthropometric research may be subject to the measurement quantity (i.e., larger measurements may have higher standard deviations) or measurement collection methods (i.e., measurements that are more difficult to collect may have higher standard deviations). Despite differing opinions, statistical understanding posits that the standard deviation is predictive of the precision of collected data, or that a lower standard deviation is predictive of higher precision (Menditto et al., 2007).

To visually compare 3D and manual measurement data, face length (Sellion to Menton) and face width (Bizygomatic Width Linear) were plotted over a bivariate panel created by the National Institute for Occupational Safety and Health (NIOSH) from manual measurement data collected to assess sizing of respirators (Zhuang et al., 2007). Based on found differences between the NIOSH Bivariate Panel and 3D measurement data, an updated 3D Bivariate Panel was created by adjusting the bivariate panel based on differences in median (50^{th} percentile) between manual and 3D methods. These bivariate panel plots were created using R (R Core Team, 2022a) and R packages tidyverse, readxl, extrafont, flextable, and rstatix (Chang, 2022; Gohel, 2022; Kassambara, 2021; Wickham, 2022; Wickham & Bryan, 2022).

Table 3.2.
Name and description of similar measurements between Zhuang and Bradtmiller (2005) and the present study. Description of Zhuang and Bradtmiller (2005) manual measurement, description of present study 3D measurement, differences in measurement collection methods between manual and 3D (referred to as measurement method differences or MMDs), and effects of MMDs (italicized).

Zhuang and Bradtmiller's (2005) measurement	Description of Zhuang and Bradtmiller (2005) manual measurement*	Present study similar measurement	Description of present study 3D measurement	MMD and *effects of MMD*
Nose Breadth	"Width across nostrils" (Hack et al., 1973) or "The straight-line distance between the right and left alare landmarks on the sides of the nostrils" (Bailar et al., 2007)	AA_C	Contour (over the surface) distance between right and left alare landmarks on the sides of the nostrils.	The manual method collects straight-line distance between alare landmarks. *The 3D measurement is larger than the manual measurement.*
Head Length	"The [straight-line] distance from the glabella landmark between the browridges to the posterior point on the back of the head is measured with a spreading caliper." (Gordon et al., 1989)	BGl_C	Contour (over the surface) distance between the glabella and the back of the head.	The manual method collects the straight-line distance between the relevant landmarks. *The 3D measurement is larger than the manual measurement.*
Bizygomatic Breadth	"The maximum horizontal breadth of the face (between the zygomatic arches) is measured with a spreading caliper." (Gordon et al., 1989)	BiW_L	Straight-line distance between the most prominent zygion points.	The manual spreading caliper requires zygion landmarks to be more distal from the center of the body. *The 3D measurement is smaller than the manual measurement.*
Lip Length	"Width of lips, mouth closed" (Hack et al., 1973) or "The straight-line distance between the left and right cheilion landmarks at the corners of the mouth" (Bailar et al., 2007)	ChCh_C	Contour (over the surface) distance between the left and right cheilion landmarks.	The manual method collects the straight-line distance between the relevant landmarks. *The 3D measurement is larger than the manual measurement.*
Nasal Root Breadth	"The horizontal breadth of the nose at the level of the deepest depression of the root (Sellion landmark) and at a depth equal to one-half the distance from the bridge of the nose to the eyes." (Bailar et al., 2007)	NRB_L	Straight-line distance between nasal root landmarks (left and right), with the landmark being placed at the deepest protrusion despite the distance from the eye and nose bridge.	The manual method landmarked the nasal root at one-half distance between the inner eye and nose bridge. *The 3D measurement is larger than the manual measurement.*
Nose Protrusion	"The straight-line distance between the pronasale landmark at the tip of the nose and the subnasale landmark under the nose." (Bailar et al., 2007)	ProS_L	Straight-line distance between the pronasale and subnasale landmarks.	No measurement method differences between 3D and manual methods (no MMD). *No difference in measurement summary statistics was expected.*
Mention-Sellion Length	"The distance between the menton landmark at the bottom of the chin and the Sellion landmark at the deepest point of the nasal root depression is measured with a sliding caliper. The teeth are lightly occluded." (Gordon et al., 1989)	SelM_L	Straight-line distance between menton and sellion landmarks.	No measurement method differences between 3D and manual methods (no MMD). *No difference in measurement summary statistics was expected.*
Bitragion Coronal Arc	"The surface distance between the right and left tragion landmarks across the top of the head is measured with a tape. The head is in the Frankfort plane." (Gordon et al., 1989)	TrHO_C	Contour (over the surface) distance between the right tragion to the top of the head.	The 3D measurement was half the manual measurement. As a result, *the 3D measurement is smaller than the manual measurement.*
Bitragion Subnasale Arc	"The surface distance between the right and left tragion landmarks across the Subnasale landmark just under the nose is measured with a tape." (Gordon et al., 1989)	TrSnas_C	Contour (over the surface) distance between the right tragion to the subnasale.	The 3D measurement was half the manual measurement. As a result, *the 3D measurement is smaller than the manual measurement.*

*Zhuang and Bradtmiller (2005) cite Gordon et al. (1989) as the describing literature for their manually-collected measurements. However, not all measurements collected in the study are found in the cited literature. Additional literature defining manually-gathered measurement was sought in the case of the absence of a description of the measurement in the cited literature.

Results and Discussion

3D Summary Statistics

3D measurement data were collected from 1063 Females, 939 Males, 5 Non-Binary or Other participants, and 9 participants who did not report their gender (8 prefer to not say, 1 missing gender data). Figures 3.3, 3.4, and 3.5 illustrate the self-reported demographic composition of the sample population, including gender, race/ethnicity, and age respectively.

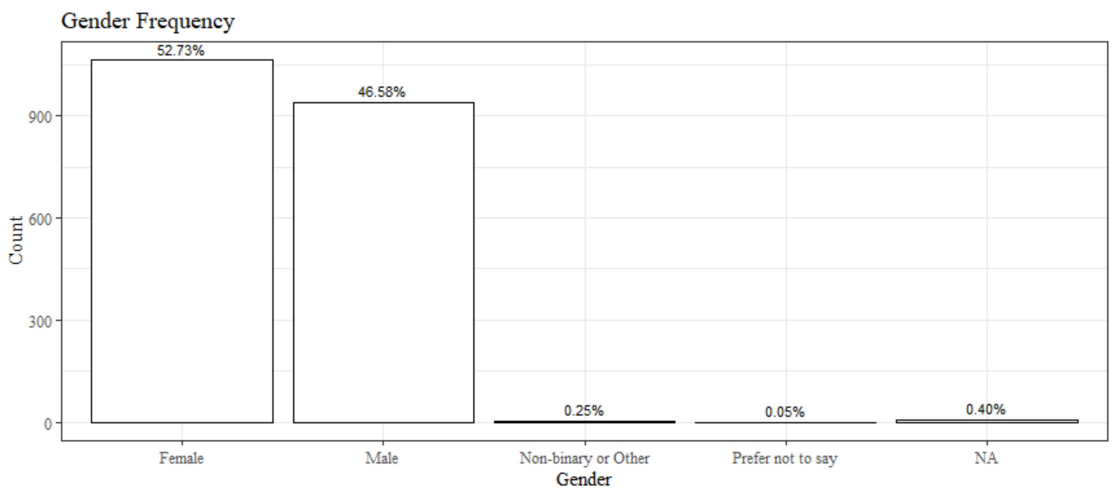

Figure 3.3: Self-reported gender of sample participants (n=2022).

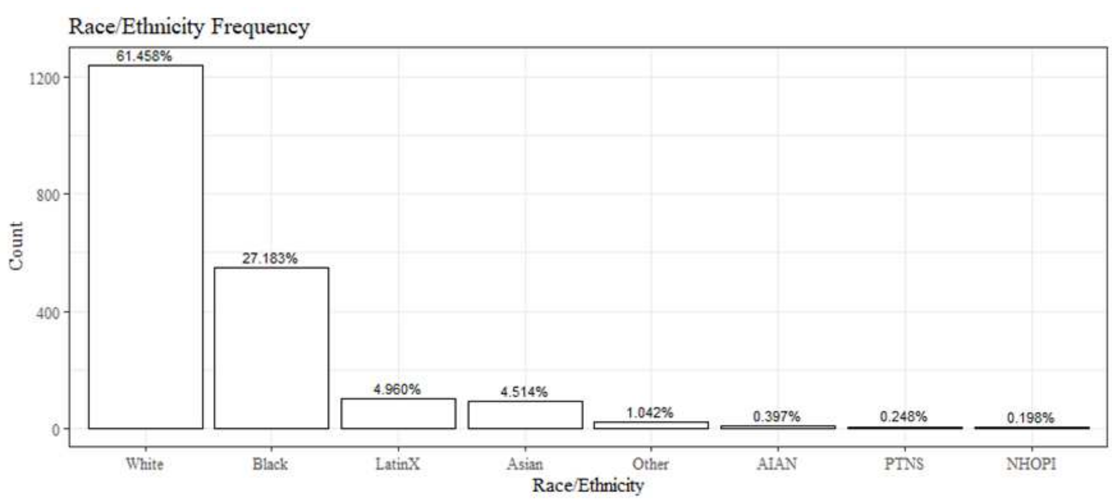

Figure 3.4: Self-reported race/ethnicity of sample participants (n=2022).

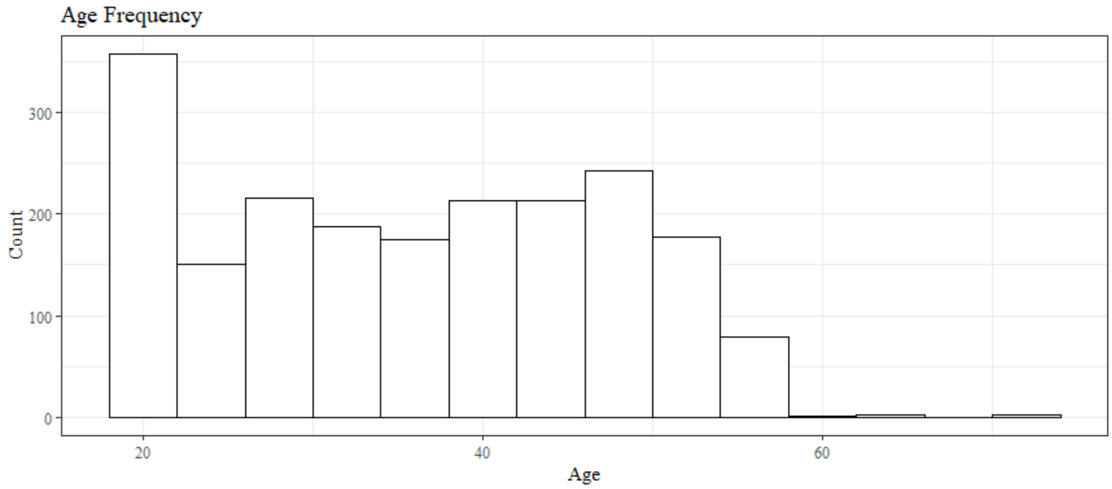

Figure 3.5: Self-reported age of sample participants (n=2022).

Table 3.3 provides summary statistics for all 2022 3D scan participants. Table 3.4 provides summary statistics for 1063 female participants, and Table 3.5 provides summary statistics for 939 male participants. Other gender categories were omitted from the analysis due to a) a small sample size and b) a lack of representation in comparable research from Zhuang and Bradtmiller (2005). The summary statistic data provided in Tables 3.3, 3.4, and 3.5 address and answer SA2.RQ1.

Table 3.3.

Summary statistics for all genders. Measure correlates to the 3D measurement abbreviation seen in Table 3.1.

Measure	N	Min.	Max.	Mean	SD	SE	5th	25th	50th	75th	95th
AA_C	1,999	44	87	61.25	6.24	0.14	52	57	61	65	72
BGl_C	1,496	215	350	292.83	15.13	0.34	269	283	293	303	318
BiW_L	1,999	82	152	111.20	9.78	0.22	96	104	111	118	128
BiW_C	1,999	101	188	133.44	12.92	0.29	114	124	133	141	156
ChCh_C	1,973	47	97	67.05	7.43	0.17	55	62	67	72	80
GoSub_C	1,888	45	217	98.88	15.45	0.34	75	88	99	108	125
NRB_L	2,000	3	40	17.95	4.74	0.11	11	15	18	21	27
ProA_L	2,000	19	39	27.94	3.03	0.07	23	26	28	30	33
ProA_C	2,000	20	44	30.12	3.50	0.08	25	28	30	32	36
ProS_L	1,997	12	42	19.16	2.70	0.06	15	17	19	21	23
ProS_C	1,985	12	43	21.10	3.35	0.07	16	19	21	23	26
SelP_L	2,001	16	65	44.53	4.41	0.10	38	42	44	47	52
SelP_C	2,001	18	66	45.01	4.49	0.10	38	42	45	48	53
SelDH_C	2,000	1	31	13.05	2.95	0.07	9	11	13	15	18
SelM_L	1,792	69	145	116.24	9.40	0.21	101	110	116	123	131
SnasM_L	1,791	40	128	67.93	8.66	0.19	52	63	68	74	81
SnasM_C	1,779	44	125	75.05	10.61	0.24	57	68	75	82	92
SmanM_L	1,772	7	92	44.48	12.01	0.27	26	37	44	52	65
SmanM_C	1,731	7	97	45.66	12.92	0.29	27	37	45	53	69
TrHO_C	1,734	135	213	166.98	10.00	0.22	152	160	167	173	184
TrEJ_C	1,983	20	60	38.03	4.69	0.10	31	35	38	41	46
TrGo_C	1,935	35	114	60.03	8.35	0.19	48	54	60	65	74
TrSel_C	1,985	120	168	141.96	7.48	0.17	130	137	142	147	155
TrSman_C	1,883	64	246	153.32	14.43	0.32	132	143	152	162	178
TrSnas_C	1,926	122	184	150.36	9.34	0.21	136	144	150	157	167
TrTr_C	1,978	241	332	282.71	14.34	0.32	261	272	282	293	308
TrTr_L	1,982	127	173	146.50	7.73	0.17	135	141	146	152	160

Table 3.4. *Summary statistics for females. Measure correlates to the 3D measurement abbreviation seen in Table 3.1.*

Measure	N	Min.	Max.	Mean	SD	SE	5th	25th	50th	75th	95th
AA_C	1,051	44	77	58.32	5.00	0.15	51	55	58	62	67
BGl_C	609	242	332	286.40	13.72	0.42	264	278	286	295	310
BiW_L	1,051	82	142	109.08	9.47	0.29	94	102	108	116	125
BiW_C	1,051	101	165	129.00	11.86	0.36	111	121	127	137	150
ChCh_C	1,044	47	97	65.22	7.24	0.22	54	60	65	70	78
GoSub_C	1,037	45	208	93.67	13.92	0.43	72	84	93	103	117
NRB_L	1,054	4	32	18.17	4.76	0.15	10	15	18	21	27
ProA_L	1,052	19	36	26.48	2.49	0.08	23	25	26	28	31
ProA_C	1,052	20	41	28.51	2.94	0.09	24	27	28	30	34
ProS_L	1,052	12	42	18.74	2.74	0.08	15	17	19	20	23
ProS_C	1,051	13	42	20.46	3.16	0.10	16	18	20	22	25
SelP_L	1,053	16	57	42.85	3.87	0.12	37	40	43	45	49
SelP_C	1,053	18	58	43.26	3.93	0.12	37	41	43	46	49
SelDH_C	1,052	1	26	12.52	2.75	0.08	8	11	13	14	17
SelM_L	1,046	69	140	113.10	8.28	0.25	99	108	114	118	126
SnasM_L	1,046	40	111	66.24	8.09	0.25	51	62	67	72	78
SnasM_C	1,040	46	125	73.02	10.05	0.31	55	67	73	80	88
SmanM_L	1,043	9	81	42.51	11.43	0.35	25	35	42	50	62
SmanM_C	1,039	9	82	43.69	12.13	0.37	25	36	43	52	64
TrHO_C	849	135	202	164.41	9.94	0.30	151	157	164	170	182
TrEJ_C	1,040	22	53	37.35	4.39	0.13	30	34	37	40	44
TrGo_C	1,039	35	98	58.24	7.69	0.24	47	53	58	63	70
TrSel_C	1,041	120	168	138.63	6.66	0.20	129	134	138	143	150
TrSman_C	1,031	87	195	146.82	11.88	0.36	130	139	145	154	167.5
TrSnas_C	1,016	122	181	146.25	8.32	0.26	135	140	145	151	161
TrTr_C	1,037	241	329	275.77	12.34	0.38	257	267	275	283	297
TrTr_L	1,040	127	165	141.87	5.83	0.18	133	13	142	145	152

Table 3.5.
Summary statistics for males. Measure correlates to the 3D measurement abbreviation seen in Table 3.1.

Measure	N	Min.	Max.	Mean	SD	SE	5th	25th	50th	75th	95th
AA_C	934	47	87	64.58	5.85	0.19	56	60	64	68	74
BGl_C	878	215	350	297.45	14.36	0.47	275	288	298	307	320
BiW_L	934	86	152	113.66	9.54	0.31	99	107	113	119	130
BiW_C	934	105	188	138.55	12.15	0.40	121	131	137	145	162
ChCh_C	915	48	95	69.13	7.09	0.23	58	64	69	74	81
GoSub_C	837	52	217	105.32	14.82	0.48	83	96	105	114	130
NRB_L	932	3	40	17.74	4.72	0.15	11	15	17	20	26
ProA_L	934	23	39	29.62	2.70	0.09	26	28	30	31	34
ProA_C	934	23	44	31.95	3.17	0.10	27	30	32	34	38
ProS_L	931	12	30	19.65	2.58	0.08	16	18	20	21	24
ProS_C	920	12	43	21.86	3.41	0.11	17	20	22	24	27
SelP_L	934	31	65	46.48	4.15	0.14	40	44	46	49	53
SelP_C	934	31	66	47.03	4.22	0.14	40	44	47	50	54
SelDH_C	934	4	31	13.69	3.02	0.10	9	12	14	15	18
SelM_L	732	93	145	120.81	8.98	0.29	106	115	121	127	135
SnasM_L	731	42	128	70.38	8.84	0.29	55	65	71	76	83
SnasM_C	725	46	105	77.99	10.69	0.35	60	71	79	85	94
SmanM_L	715	7	92	47.25	12.34	0.40	29	39	46	55	71
SmanM_C	679	7	97	48.59	13.58	0.44	30	40	47	56	76
TrHO_C	875	141	213	169.45	9.43	0.31	154	163	169	175	185
TrEJ_C	929	20	60	38.80	4.91	0.16	31	35	39	42	47
TrGo_C	882	35	114	62.19	8.61	0.28	49	56	62	67	76
TrSel_C	930	125	168	145.72	6.55	0.21	135	141	146	150	157
TrSman_C	838	64	246	161.36	13.19	0.43	143	153	160	169	184
TrSnas_C	896	127	184	155.07	8.20	0.27	142	149	155	160	169
TrTr_C	927	254	332	290.53	12.33	0.40	271	282	290	299	312
TrTr_L	928	132	173	151.72	6.18	0.20	142	147	152	156	162

Overall, 3D facial measurement means were larger for males than for females. Figure 3.6 illustrates the differences between the measurement means for male and female genders. Previous research has similarly reported female faces to be smaller than male faces (Brazile et al., 1998; Gross & Horstman, 1990; Oestenstad & Perkins, 1992; Zhuang et al., 2010). This finding continues to have implications for the design and manufacture of head and face products, specifically those products that are meant to protect the wearer (such as personal protective equipment). Proper sizing of these products is of high importance, as improper fit could lead to safety hazards. Head and face products may need to be made available in multiple sizes to accommodate differences in face sizes between genders.

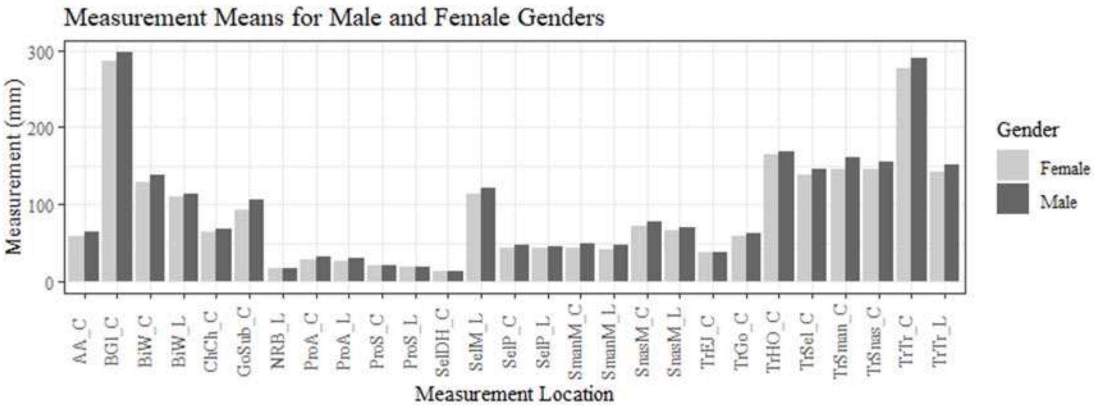

Figure 3.6: 3D measurement means for male and female genders. Overall, 3D facial measurements for males were larger than 3D facial measurements for females.

3D vs. Manual Methods

The following section of discussion addresses and answers SA2.RQ2. Table 3.6 compares 3D measurement summary statistics from this study and manual measurement summary statistics from previously conducted research for female participants (Zhuang & Bradtmiller, 2005). Table 3.7 compares 3D measurement summary statistics from this study and manual measurement summary statistics from previously conducted research for male participants (Zhuang & Bradtmiller, 2005).

Table 3.6.

3D vs. manual (Zhuang and Bradtmiller, 2005) measurements, female participants.

Measurement	Method	N	Min.	Max.	Mean	SD	5th	25th	50th	75th	95th
AA_C	3D	1,051	44	77	58.32	5.00	51	55	58	62	67
Nose breadth	Manual	1,454	22	54	33.20	3.90	28	31	33	35	41
BGl_C	3D	609	242	332	286.40	13.72	264	278	286	295	310
Head length	Manual	1,454	152	215	187.50	7.20	175	183	187	192	199
BiW_L	3D	1,051	82	142	109.08	9.47	94	102	108	116	125
Bizygomatic breadth	Manual	1,454	115	157	135.10	6.50	124	131	135	140	146
ChCh_C	3D	1,044	47	97	65.22	7.24	54	60	65	70	78
Lip length	Manual	1,454	35	63	48.00	4.00	42	45	48	51	55
NRB_L	3D	1,054	4	32	18.17	4.76	10	15	18	21	27
Nasal root breadth	Manual	1,454	10	25	16.30	2.00	13	15	16	18	20
ProS_L	3D	1,052	12	42	18.74	2.74	15	17	19	20	23
Nose protrusion	Manual	1,454	11	29	19.80	2.70	16	18	20	21	25
SelM_L	3D	1,046	69	140	113.10	8.28	99	108	114	118	126
Menton-Sellion length	Manual	1,454	91	135	113.40	6.10	104	109	113	118	124
TrHO_C	3D	849	135	202	164.41	9.94	151	157	164	170	182
Bitragion coronal arc	Manual	1,454	290	425	339.30	15.00	315	330	340	350	365
TrSnas_C	3D	1,016	122	181	146.25	8.320	135	140	145	151	161
Bitragion subnasale arc	Manual	1,454	238	335	277.50	13.10	258	269	277	285	300

Table 3.7.

3D vs. manual (Zhuang and Bradtmiller, 2005) measurements, male participants.

Measurement	Method	N	Min.	Max.	Mean	SD	5th	25th	50th	75th	95th
AA_C	3D	934	47	87	64.58	5.85	56	60	64	68	74
Nose breadth	Manual	2,543	26	58	36.60	4.10	31	34	36	39	45
BGl_C	3D	878	215	350	297.45	14.36	275	288	298	307	320
Head length	Manual	2,543	174	225	197.30	7.40	185	192	197	202	210
BiW_L	3D	934	86	152	113.66	9.54	99	107	113	119	130
Bizygomatic breadth	Manual	2,542	120	170	143.50	6.90	132	139	143	148	155
ChCh_C	3D	915	48	95	69.13	7.09	58	64	69	74	81
Lip length	Manual	2,543	40	70	51.10	4.20	44	48	51	54	58
NRB_L	3D	931	12	30	19.65	2.58	16	18	20	21	24
Nasal root breadth	Manual	2,543	10	29	16.60	2.30	13	15	16	18	20
ProS_L	3D	931	12	30	19.65	2.58	16	18	20	21	24
Nose protrusion	Manual	2,543	13	32	21.10	2.70	17	19	21	23	26
SelM_L	3D	732	93	145	120.81	8.98	106	115	121	127	135
Menton-Sellion length	Manual	2,543	100	156	122.70	7.00	111	118	123	127	135
TrHO_C	3D	896	127	184	155.07	8.20	142	149	155	160	169
Bitragion coronal arc	Manual	2,543	310	405	350.70	13.90	330	340	350	360	375
TrSnas_C	3D	896	127	184	155.07	8.20	142	149	155	160	169
Bitragion subnasale arc	Manual	2,543	253	345	294.80	13.20	275	285	295	305	315

To assess the equality of variance between 3D and manual summary statistics (Zhuang & Bradtmiller, 2005), F-test statistics, standard deviations, and p-values for two-tailed F-tests (significance level: $p<0.05$) are reported in Table 3.8 for each similar measurement and for each gender. The F-test results suggested equal variance between 3D and manual measurement standard deviations for both genders for the ProS_L and Nose Protrusion measurements. All other F-tests results suggested unequal variance between methods, for both genders, and for all other tested measurement locations (with and without MMDs). 3D measurements tended to have larger standard deviations, except for in the case of Bitragion Coronal Arc and Bitragion Subnasale Arc (manual measures had larger standard deviations for these two measures).

Table 3.8.

F-test statistics for similar measurements, separated by gender. Standard deviations for 3D vs. manual methods, p-values (for two-tailed F-tests), and conclusions regarding equal or unequal variance status (significance level p>0.05) between 3D and manual methods are presented.

Measurement (3D and manual)	Gender	F-test statistic	3D Standard Deviation	Manual Standard Deviation	p-value	Variances?*
AA_C and Nose breadth	Female	1.64	5.00	3.90	<0.01	Unequal
	Male	2.04	5.85	4.10	<0.01	Unequal
BGl_C and Head length	Female	3.63	13.72	7.20	<0.01	Unequal
	Male	3.77	14.36	7.40	<0.01	Unequal
BiW_L and Bizygomatic breadth	Female	2.12	9.47	6.50	<0.01	Unequal
	Male	1.91	9.54	6.90	<0.01	Unequal
ChCh_C and Lip length	Female	3.28	7.24	4.00	<0.01	Unequal
	Male	2.85	7.09	4.20	<0.01	Unequal
NRB_L and Nasal root breadth	Female	5.66	4.76	2.00	<0.01	Unequal
	Male	1.26	2.58	2.30	<0.01	Unequal
ProS_L and Nose protrusion	Female	1.03	2.74	2.70	0.61	Equal
	Male	0.91	2.58	2.70	1.90	Equal
SelM_L and Menton-Sellion length	Female	1.84	8.28	6.10	<0.01	Unequal
	Male	1.65	8.98	7.00	<0.01	Unequal
TrHO_C and Bitragion coronal arc	Female	2.28	9.94	15.00	<0.01	Unequal
	Male	2.87	8.20	13.90	<0.01	Unequal
TrSnas_C and Bitragion subnasale arc	Female	2.48	8.32	13.10	<0.01	Unequal
	Male	2.59	8.20	13.20	<0.01	Unequal

*Based on F-statistic compared to F-critical values, and significance level p=0.05.

Measurements with No Measurement Method Differences (MMDs). The comparison of summary statistics for similar 3D and manual measurement locations quantified MMDs for similar measurements (described in Table 3.2). No MMD was noted for Nose Protrusion and ProS_L, and therefore no difference in measurement summary statistics was expected. For both genders, most summary statistics were within one to two millimeters when comparing 3D and manual, with the manual measurement tending to be slightly larger. These findings contradict previous research findings that 3D measurements tend to be larger than manual measurements (Bailar et al., 2007; Gordon et al., 2014; Han et al., 2010; Heymsfield et al., 2018). According to the two-tailed F-test results presented in Table 3.8, the variance for Nose Protrusion and ProS_L was equal for the two methods. Statistical understanding posits that the standard deviation (a measure of variance) is predictive of the precision of collected data, or that a lower standard deviation is predictive of higher precision (Menditto et al., 2007). Thus, these findings may indicate comparable precision between 3D and manual measurement methods for the collection of nose protrusion measurement data.

No MMD was noted for Menton-Sellion Length and SelM_L (i.e., face length), and therefore no difference in measurement summary statistics was expected. For both genders, summary statistics were comparable. This finding mirrors the expected outcome listed in Table 3.2 yet contradicts previous research outcomes of 3D measurements being larger in general (Bailar et al., 2007; Gordon et al., 2014; Han et al., 2010; Heymsfield et al., 2018). However, the variance for face length measurement was unequal between the two methods (Table 3.8), with a higher standard deviation for 3D measurement collection. This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of face length measurements. Furthermore, it should be noted that facial hair volume is more easily adjusted when gathering manual measurements, due to the ability to compress facial hair and identify the menton (chin) landmark as needed to gather the measurement. For male participants of the 3D scan anthropometric data collection, there were many missing values due to the inability to compress facial hair and gather face length measurements. Researchers and professionals must weigh the strengths and weaknesses of 3D measurements when utilizing this method for face length measurement collection.

Measurements with Measurement Method Differences (MMDs). The comparison of summary statistics for similar 3D and manual measurement locations quantified MMDs for similar measurements (described in Table 3.2). Alare to Alare Contour, a 3D surface contour measurement, was expected to be larger than Nose Breadth, a manual straight-line measurement. This expectation was found to be true in the measurement summary statistics for both genders. Variances were unequal for the two methods, with a higher standard deviation for 3D measurement collection (Table 3.8). This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of nose breadth measurements.

Back of Head to Glabella, a 3D surface contour measurement, was expected to be larger than Head Length, a manual straight-line measurement. This expectation was found to be true in the measurement summary statistics for both genders. Variances were unequal for the two methods, with a much higher standard deviation for 3D measurement collection (Table 3.8). This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of head size measurements.

Bizygomatic Linear Width (3D linear measurement) and Bizygomatic Breadth (manual measurement) were both collected as straight-line measurements to assess face width. However, the 3D measurement was expected to be smaller due to differences in 3D and manual landmark placement (see Table 3.2 for more details). Based on summary statistics, the expected effect of the MMD was found to be accurate for both genders. Variances were unequal for the two methods, with a higher standard deviation for 3D measurement collection (Table 3.8). This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of face width measurements.

Cheilion to Cheilion, a 3D surface contour measurement, was expected to be larger than Lip Length, a manual straight-line measurement. This expectation was found to be true in the measurement summary statistics for both genders. Variances were unequal for the two methods, with a much higher standard deviation for 3D measurement collection (Table 3.8). This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of lip length measurements.

Nasal Root Breadth (3D linear measurement) and Nasal Root Breadth (manual measurement) were both collected as straight-line measurements to assess the width of the nasal root. However, the 3D measurement was expected to be smaller due to differences in 3D and manual landmark placement (see Table 3.2 for more details). Based on summary statistics, the expected effect of the MMD was found to be accurate for both genders. Variances were unequal for the two methods, with a higher standard deviation for 3D measurement collection (Table 3.8). This may indicate that the 3D measurement method provides less precision than the manual measurement method for the collection of nasal root breadth measurements.

Top of Head to Otobasion (3D contour measurement) and Bitragion Coronal Arc (manual measurement) were both collected as surface contour measurements to assess the distance from the tragion (ear landmark) to the top of the head. However, the 3D measurement was expected to be smaller as it collected half the measurement length of the manual measurement (see Table 3.2 for more details). More specifically, the 3D measurement was expected to be about half of the manual measurement. Based on summary statistics, this expectation was found to be true for females, however, the 3D measurement mean for males was slightly less than half the length of the manual measurements based on summary statistics. This may indicate that the male participants of the 3D scan anthropometric data collection had overall shorter tragion to top of head lengths when compared to those male participants of the manual anthropometric data collection (Zhuang & Bradtmiller, 2005), or may indicate an unknown flaw in one of the measurement collection methods. Variances were unequal for the two methods, with a higher standard deviation for manual measurement collection (Table 3.8). This may indicate that the 3D measurement method provides more precision than the manual measurement method for the collection of tragion to top of head measurements. However, it should be noted that head hair volume is more easily adjusted for when gathering manual measurements, due to the ability to compress the hair as needed to gather the measurement. For female participants of the 3D scan anthropometric data collection, there were many missing values due to the placement of hair on top of the head. Thus, each measurement method presents strengths and limitations for the collection of tragion to top-of-head measurements.

Tragion to Subnasale (3D contour measurement) and Bitragion Subnasale Arc (manual measurement) were both collected as surface contour measurements to assess the distance from the tragion (ear landmark) to the subnasale (under the nose). However, the 3D measurement was expected to be smaller, as it collected half the measurement length of the manual measurement (see Table 3.2 for more details). More specifically, the 3D measurement was expected to be about half of the manual measurement. Based on summary statistics, this expectation was found to be true for males, however, 3D measurements for females were slightly less than half the length of the manual measurements based on summary statistics. This may indicate that the female participants of the 3D scan anthropometric data collection had overall shorter distances from tragion to subnasale when compared to those female participants of the manual anthropometric data collection (Zhuang & Bradtmiller, 2005), or may indicate an unknown flaw in one of the measurement collection methods.. Variances were unequal for the two methods, with a higher standard deviation for manual measurement collection (Table 3.8). This may indicate that the 3D measurement method provides more precision than the manual measurement method for the collection of tragion to subnasale measurements. However, it should be noted that facial hair volume (particularly for Male gender) is more easily managed when gathering manual measurements, due to the ability to compress the hair as needed to gather the measurement.

Thus, each measurement method presents strengths and limitations for the collection of tragion to subnasale measurements.

Bivariate Panel. Figure 3.7 illustrates the face width (Bizygomatic Width Linear) and face length (Sellion to Menton Length) 3D measurement data from this work overlayed on the NIOSH bivariate panel created from similar manual measurements in previous research (Zhuang et al., 2007). The NIOSH bivariate panel (Zhuang & Bradtmiller, 2005) comparison to the 3D measurement data for face length and face width mirrors expected and observed findings regarding comparisons (i.e., MMDs or no MMDs) of relevant measurement locations per either method (see Table 3.2). It was expected that the 3D measurement for face width (Bizygomatic Linear Width) would be smaller than the manual face width measure, due to differences in landmark placement as the result of each method's tools. For the 3D measurement data, this was shown to be true in summary statistics and is similarly visually observed in the bivariate panel comparisons (i.e., data points are shifted to the left of the bivariate panel boxes on the x-axis). No effects of MMD were expected or observed for the 3D or manual measurement of face length (Sellion to Menton Linear and Menton-Sellion Length, respectively), and this is also visually observed in the bivariate panel comparisons (i.e., data points are vertically centered near bivariate panel boxes on the y-axis).

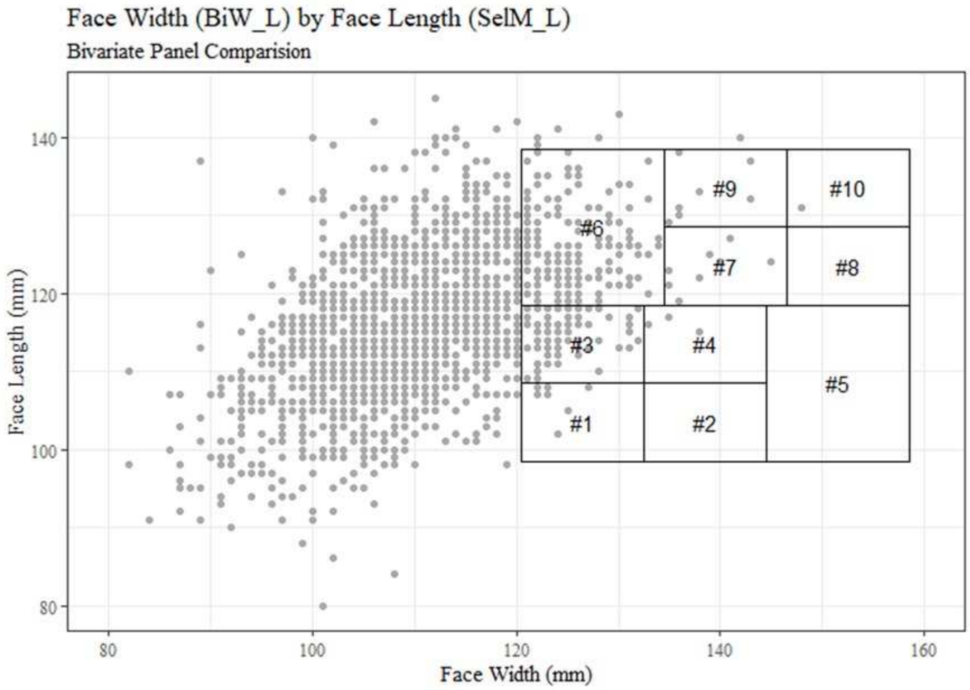

Figure 3.7: Comparison of 3D bivariate face length and face width 3D measurement data (plotted points) and NIOSH bivariate panel classifications (numbered boxes) developed from manual face length and face width measurement data (Zhuang et al., 2007).

To illustrate what an updated bivariate panel would look like based on 3D measurement data collected in the present study, the bivariate panel was adjusted based on the differences in median (or 50th percentile) between 3D and manual methods for face width. Based on the summary statistics found in Tables 3.6 and 3.7, face width panel boundaries were adjusted 30mm smaller for the new 3D bivariate panel. Figure 3.8 illustrates the face width (Bizygomatic Width Linear) and face length (Sellion to Menton Length) 3D measurement data from this work overlayed on an updated 3D bivariate panel based on differences between 3D and manual measurement medians (50th percentile).

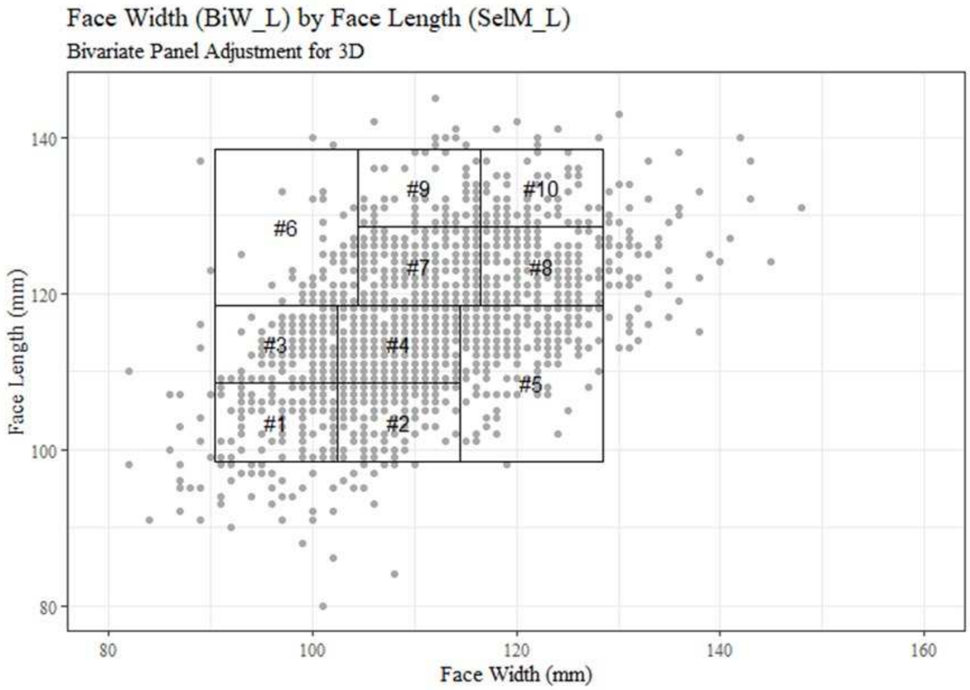

Figure 3.8: 3D bivariate face length and face width 3D measurement data (plotted points) and adjusted 3D bivariate panel classifications (numbered boxes) based on differences in median (50th percentile) between face width manual measurement and BiW_L 3D measurement.

Limitations

All anthropometric data collection efforts are limited by the sample population they collect measurement data from, as well as by both the people and equipment they use to collect measurement data. The present study is no exception to these generalities. Furthermore, summary statistics of measurement data can only be used to provide a summary-level understanding of the measurement data collected. Summary statistics cannot be used to ascertain significant differences between data. Thus, the present study highlights only summary-based differences between 3D and manual measurements. Regarding measurements MMDs, few direct comparisons could be made due to the similar (yet not identical) nature of the method of measurement collection between 3D and manual methods. Because the MMDs in this research had expected effects on the comparability of summary statistics between methods (3D vs manual), information regarding the detailed differences between summary statistics of the two methods could not be elicited. In the case of no effects from MMDs, differences in summary statistics could still not be fully assessed due to comparison between two different sample populations. Future research should seek to collect 3D and manual measurement data from the same sample population, to assess measurement method differences (MMDs) more accurately.

Conclusions

In address of SA2, the present study examined the similarities and differences, as well as the strengths and weaknesses, of 3D measurement data gathered from 2022 3D scans, as compared to manual measurement data gathered by previous researchers (Zhuang et al., 2007; Zhuang & Bradtmiller, 2005). The summary statistics presented in this study provide measurement information that may be relevant to designers of respirators, face masks, glasses frames, hats, and other facial products. In comparison to manually collected (manual) facial measurements, the 3D measurements summarized in this research provide more contextual information regarding the surface contour lengths of facial measurement locations. However, 3D measurements may differ from manual measurements in terms of precision and actual found measurement. For example, the present study found that standard deviations of manual methods were generally higher, and therefore manual measurement collection methods may be more precise. Furthermore, even in the case of no measurement method differences (no MMDs) between the manual and 3D methods, summary statistics differed in a way that contradicted previous literature (i.e., 3D measurements were either larger than manual or the same as manual). Future research is needed to continue to assess important differences between 3D and manual measurement collection methods for facial anthropometric data.

CHAPTER 4: DEMOGRAPHIC DIFFERENCES IN 3D FACIAL ANTHROPOMETRICS RELATED TO RESPIRATOR FIT

Summary

In response to limitations in seminal anthropometric research efforts as well as the increasingly diversifying workforce, researchers have attempted to define the presence or absence of differences in respirator-related facial measurements between different demographic groups. Three-dimensional (3D) body scanning technology offers a faster way to gather facial measurement data that provide more context to facial surface dimensions than those data collected using manual anthropometric data collection methods. The purpose of the present study was to assess the presence of differences in 3D facial measurements related to respirator fit, based on demographic factors of gender, race/ethnicity, and age in a sample of 2022 3D scans. Principal components analysis (PCA) and Multivariate Analysis of Variance (MANOVA) were used to determine the presence or absence of these differences. Results indicated that 3D measurements related to respirator fit were significantly different for all groups within each demographic category (gender, race/ethnicity, and age).

Introduction

Many previous research efforts, completed by government and academic institutions, have attempted to quantify the facial anthropometrics of working, respirator-wearing populations. In 1973, researchers at the Los Alamos National Laboratory (LANL) conducted research for NIOSH which evaluated the "fit of half-mask, quarter-mask, and full-facepiece respirators" (Hack et al., 1973, p. 1). To evaluate the fit of these respirators, facial anthropometric data were collected from 200 civilian males, 40% of whom were "Spanish-American" (Hack et al., 1973, p. 5). The LANL researchers did not find differences higher than

2mm in the means of all measured face dimensions (Hack et al., 1973). However, in 1975, Leigh (1975) found that of 1467 Dow Chemical employees (127 of whom were female), 12.6% were not represented by the LANL face panel for full respirators. This finding sparked questions regarding the gender-based generalizability of the LANL face panel. Gross and Horstman (1990) utilized three sizes of respirators from three brands (nine respirators total) to conduct quantitative fit tests of respirators on 120 civilians (60 females and 61 males) and found that 95% of the participants were able to fit using the respirators provided, though a) no brand name was given for these respirators and b) the sample lacked diversity regarding age, race, and ethnicity (Gross & Horstman, 1990). Oestenstad & Perkins (1992) conducted research to investigate the facial measurements of 30 females and 38 males and found that the measurements collected did not differ greatly from previous research. The diversity (regarding race, ethnicity, and age) of the sample populations in these gender-focused efforts was not found to be inclusive or representative of working populations at the time (Brazile et al., 1998).

To address limitations from previous research regarding diversity and respirator fit, subsequent research efforts sought to quantify differences in facial dimensions relevant to respirator fit between racial, ethnic, age, and gender-based groups. Brazile et al. (1998) found that face measurements related to respirator fit were significantly different between different groups in gender and race/ethnicity categories. Kim et al. (2003) conducted research to assess the association between Korean facial measurements and respirator fit factors; they found that a) male and female Koreans had significant differences in almost all measurements, and b) respirator fit depended on different measurements for Koreans than for Americans (H. Kim et al., 2003). Zhuang et al. (2010) found that face measurements related to respirator fit were significantly different "between males and females, all racial/ethnic groups, and the subjects who were at least 45 years old when compared to workers between 18 and 29 years of age" (p. 391). Luximon et al. (2010) conducted research to assess the facial anthropometric variation of Chinese women, finding that Chinese females have wider and shorter faces compared to other "cultures" (p. 1). Ball et al. (2010) found that the head shape of Chinese people differed in appearance from White people, "with a flatter back and forehead" (p. 832). Using 22 facial dimensions relevant to pilot oxygen mask design, Lee et al. (2012) found that Korean male pilots' faces differed from both Korean male civilians' and US male pilots' faces, and that Korean female pilots' faces were "significantly smaller" than Korean male pilots (p. 1927). Other important research efforts have sought to confirm that current respirators fit diverse demographic populations, such as South African people (Spies et al., 2011), Chinese people (Zhang et al., 2020), and Chilean people (Rodríguez et al., 2020).

Of particular interest in the field of anthropometrics is the rise in the popularity of 3D scanning, which offers a faster, lower-contact way to analyze facial measurements that provide more context to facial surface dimensions than manual measurements. Measurements gathered from 3D scans have been utilized in previous research regarding demographics and respirator fit (Ball et al., 2010; Lee et al., 2012). The specific aim (*SA3*) of the present study was to assess the presence of differences in 3D facial anthropometric measurements (referred to as 3D facial measurements or 3D measurements) related to respirator fit, based on demographic factors of gender, race/ethnicity, and age in a sample of 2022 3D scans. This study assessed the largest sample of 3D facial anthropometrics seen in the literature to date. This study helps determine if people (within the sample population) of different gender, race/ethnicity, or age groups can be expected to have different 3D facial measurements relevant to respirator fit, which has implications for respirator design, sizing, and fit for diverse workers. Three research questions guided SA3:

SA3.RQ1: Are differences in 3D measurements present between gender groups?

SA3.RQ2: Are differences in 3D measurements present between race/ethnicity groups?

SA3.RQ3: Are differences in 3D measurements present between age groups?

Methods

3D scan data for this study were purchased from Human Solutions (Human Solutions, n.d.). The company collected 3D facial scans from 2022 participants using the handheld Artec 3D structured-light scanner (Model Eva, Senningerberg, Luxembourg). Human Solutions' proprietary 3D scan software, Anthroscan (Version 3.6.1, Kaiserslautern, Germany), was used to collect 3D measurements from each 3D scan. Self-reported gender, racial/ethnic, and age information was collected at the time of 3D scanning for each of the 2022 participants.

3D Measurement Collection

3D facial measurement data collection was completed by a team of four novice raters. The specifics of 3D facial measurement data collection for the 2022 3D facial scans (including training, process, and missing data) are described in Chapter 2, which assessed the rater-based reliabilities of gathered 3D measurements before larger 3D measurement data collection began. Figure 4.1 illustrates 12 3D facial measurements collected using Anthroscan. Table 4.1 describes the name of each measurement, whether the measurement was collected in a linear (direct from point to point) and/or contour (over the surface of the face) fashion, and the abbreviated measurement name. These 12 3D measurements were selected from a larger sample of 27 measurements collected from the sample population in effort to reduce the complexity of the statistical analysis (described in Methods section below). The 12 measurements selected were chosen as most representative of respirator fit based on 1) cited relevance of measurements to respirator fit (Zhuang et al., 2007), 2) correlation with other measurements collected (i.e., in cases of high correlation, only one measurement from the correlated set was included), 3) rater reliability (inter- and intra-, described in Chapter 2), and 4) novelty of measurement to the field of literature (i.e., does this measurement provide something new to the field of literature surrounding respirator facial anthropometrics?). Furthermore, 3D measurement locations with a high percentage of missing values, caused by occlusions present in the 3D scan (described in Chapter 2), were avoided for inclusion in the selected measurements for statistical analysis in the present study.

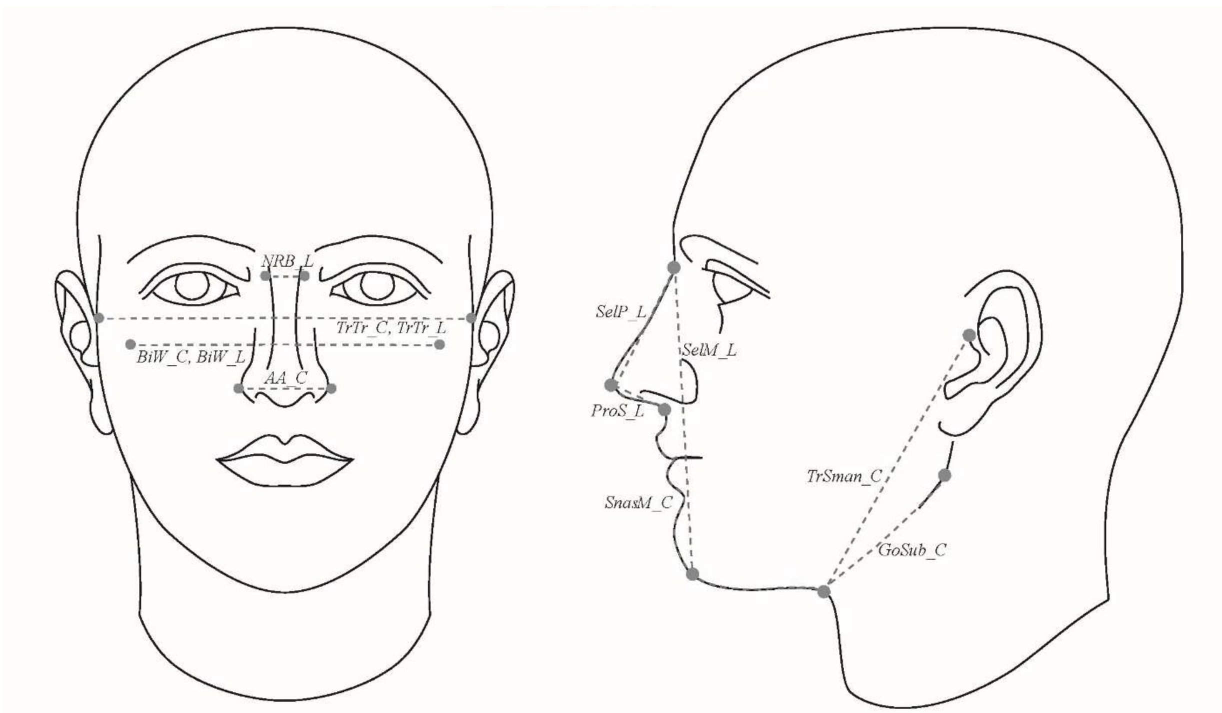

Figure 4.1: Illustration of 3D facial measurements collected from each 3D scan.

Table 4.1.
Measurement names (corresponding to Figure X), measurement type (linear, contour, or both), and abbreviated measurement name.

Measurement Name	Measurement Type	Abbreviated Name
Alare to Alare	Contour	AA_C
Bizygomatic Width	Contour	BiW_C
Bizygomatic Width	Linear	BiW_L
Gonion to Submandibular	Contour	GoSub_C
Nasal Root Breadth	Linear	NRB_L
Pronasale to Subnasale	Linear	ProS_L
Sellion to Pronasale	Linear	SelP_L
Sellion to Menton	Linear	SelM_L
Subnasale to Menton	Contour	SnasM_C
Tragion to Gonion	Contour	TrGo_C
Tragion to Sellion	Contour	TrSel_C
Tragion to Submandibular	Contour	TrSman_C
Tragion to Tragion	Contour	TrTr_C
Tragion to Tragion	Linear	TrTr_L

Statistical Analysis

Principal Components Analysis (PCA). Principal components analysis (PCA) is a statistical dimension-reducing technique that can quantify a dataset's variability through the calculation of principal components (PCs) (Holmes & Huber, 2022; James et al., 2021). As described by Zhuang and Bradtmiller (2007), "PCA defines a new coordinate system using linear combinations of the original variables to describe trends in the data." (p. 649). Subsequent score plotting of each participant's data based on how they relate to each principal component or the new coordinate system (Zhuang et al., 2007) can provide visual context to dataset variability (Holmes & Huber, 2022; James et al., 2021). Because PCA reduces data complex data to a visualizable state, it is commonly used in anthropometric research where the collection of measurement data from multiple locations for each participant typically results in a large dataset with many variables. Furthermore, PCA score plotting can help researchers visually identify important categorical trends in dataset variability. In this study, differences in variability between groups within each demographic category were analyzed using PCA score plotting. PCA and

PCA score plotting were completed using R (R Core Team, 2022b) and packages tidyverse, readxl, extrafont, flextable, writexl, ggfortify, and scales (Chang, 2022; Gohel, 2022; Horikoshi & Tang, 2018; Ooms, 2021; Wickham et al., 2019; Wickham & Bryan, 2022; Wickham & Seidel, 2022).

Multivariate Analysis of Variance (MANOVA). Multivariate Analysis of Variance (MANOVA) is a statistical technique that can assess multiple continuous dependent variables (i.e., 3D measurement variables) to determine the presence of significant differences between multiple categorical independent variables (i.e., demographic variables) (*R in Action*, 2021). Compared to other methods of analysis of variance (ANOVA), MANOVA allows for a higher correlation between continuous variables, which is inherently present in anthropometric research (e.g., participants with wider faces tend to have wider face measurements overall). Previous research from Brazile et al. (1998) used MANOVA to assess demographic differences in respirator-relevant face measurements. In the presence of significant findings in the present study, further examination of significant differences for the 12 3D measurement variables was done using post hoc ANOVA testing. To complete the MANOVA and post hoc ANOVA analyses, R Studio (R Core Team, 2022b) and R packages tidyverse, readxl, extrafont, flextable, car, broom, and emmeans were used (Chang, 2022; Fox et al., 2022; Gohel, 2022; Lenth, 2022; Robinson et al., 2022; Wickham et al., 2019; Wickham & Bryan, 2022).

Results

Figure 4.2 illustrates the number of missing data values (out of 2022) for each of the 12 3D measurements analyzed in this research. Because missing values cannot be used in PCA or MANOVA, participants with missing 3D measurement values were removed from the analyses in this study. The resulting sample size was reduced from 2022 to 1677 participants/scan subjects. PCA and MANOVA analyses were run using the reduced dataset of 1677 subjects as well as an imputed dataset of 2022 participants. Statistical results were found to be consistent between the reduced and imputed datasets. Thus, only the results of the reduced dataset (1677 subjects) are discussed in the subsequent sections of this chapter. Table 4.2 provides the racial/ethnic composition of the reduced sample, Table 4.3 provides the gender composition of the reduced sample, and Table 4.4 provides the age composition of the reduced sample.

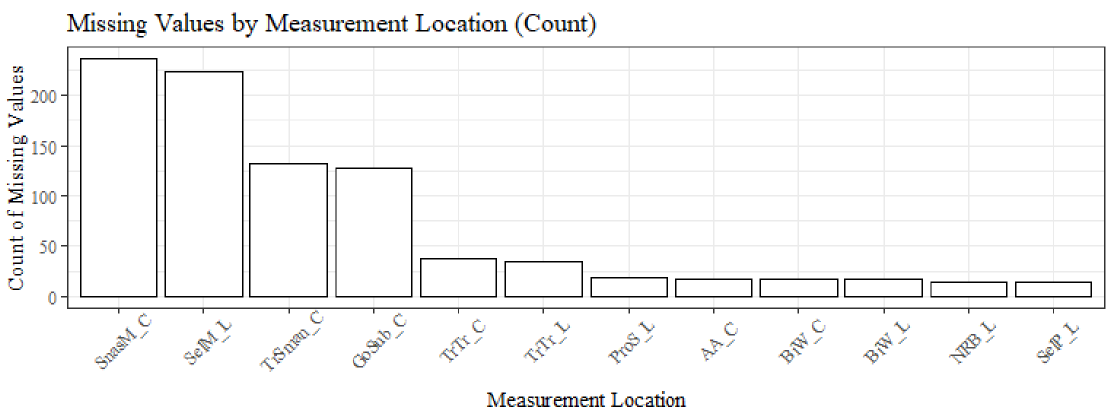

Figure 4.2: Count of missing values per measurement location (out of 2022 for each measure).

Table 4.2.
Racial/ethnic makeup of 3D face scan dataset.

Race/ethnicity	Abbreviated term	n (total n=1677)
White/Caucasian	White	1040
Black, African, or African American	Black	446
Latin/Hispanic	LatinX	84
Asian/Asian American	Asian	81
Other*	Other	26

*Due to small sample size, "Other" represents participants who self-identified as Other (n=13), American Indian or Alaska Native (n=6), Native Hawaiian or Other Pacific Islander (n=3), or Prefer not to say (n=4).

Table 4.3.
Gender makeup of 3D face scan dataset.

Gender	n (total n=1677)
Female/Other*	996
Male	681

* Female/Other represents participants who self-identified as Female (n=994), non-binary or other (n=1), and prefer not to say (n=1).

Table 4.4.
Age makeup of 3D face scan dataset.

Age	Term	n (total n=1677)
18-34	Youngest	826
35-54	Mid-age	777
55-72	Oldest	74

*Age was given by participants as exact numeric and subsequently divided into three groups for data analysis. Group limits were developed based on the oldest participant's age and approximately equal age spacing in each group.

Principal Components Analysis (PCA)

The scree plot in Figure 4.3 illustrates the percent of variability described by each principal component, with total variability described by the 12 PCs equaling 100%. Table 4.5 provides the factor loadings for each measurement location for PC1, PC2, and PC3, together describing 69.91% of the variability in the dataset. PC1 and PC2 described 58.87% of the variability in the dataset and were used as the new coordinate system (Zhuang et al., 2007) to plot each observation (i.e., each participant) on the subsequent PCA score plots (Figures 4.4-4.7). Figure 4.4 illustrates a PCA score plot of the entire dataset, with factor loadings overlaid. Long line lengths (either in the positive or negative direction) indicate large factor loading, or strong variable effect on the principal components (PC1 and PC2 only). Small angles between lines on the factor loading plot indicate a positive correlation between variables. Right (90-degree) angles between lines indicate a lack of correlation. Large (180-degree) angles indicate a negative correlation; however, no negative correlations are seen in the factor loadings plot in Figure 4.4.

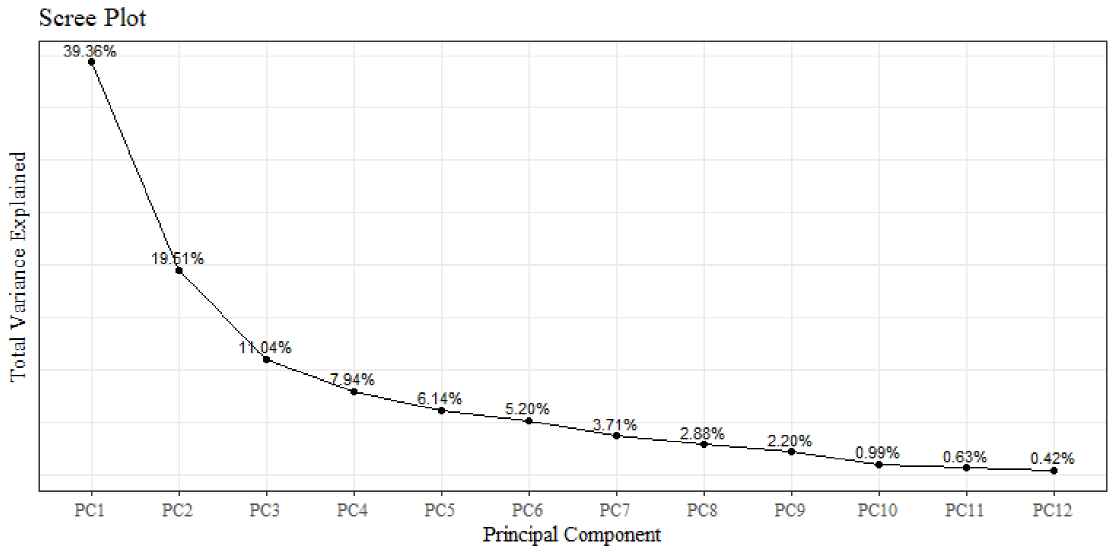

Figure 4.3: Scree plot illustrating the proportion of variance explained by each principal component.

Table 4.5.

PCA Factor Loadings for Principal Component 1 (PC1), Principal Component 2 (PC2), and Principal Component 3 (PC3). Variables with the largest loadings, and therefore with the highest influence on each PC, are bolded.

Measurement	PC1	PC2	PC3
AA_C	0.22610676	**-0.3779251**	-0.27191984
BiW_C	0.31862749	0.3092589	-0.14925876
BiW_L	0.31985590	**0.3603274**	0.06247600
GoSub_C	0.26391016	-0.3194544	**0.39647678**
NRB_L	0.13011314	0.2687254	0.20783339
ProS_L	0.09639314	**-0.4223793**	-0.26287412
SelP_L	0.19424050	-0.1133301	**-0.63600471**
SelM_L	**0.35269639**	0.1874002	-0.28461427
SnasM_C	0.29601013	**0.3659376**	-0.06456278
TrSman_C	**0.36861199**	-0.2192867	0.30097637
TrTr_C	**0.37142151**	-0.1152908	0.15450138
TrTr_L	**0.35440086**	-0.1891442	0.16203007

PCA Plot with Factor Loadings

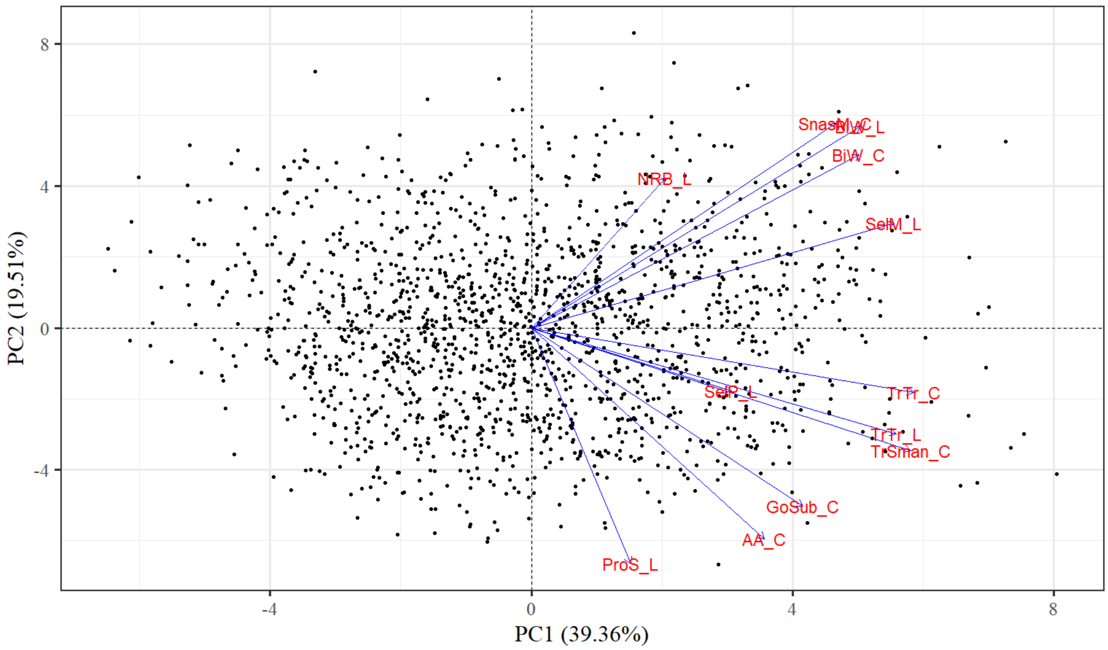

Figure 4.4: PCA score plot with factor loadings

Figure 4.5 illustrates a PCA score plot of the dataset with gender groups identified by colored ellipses. Figure 4.6 illustrates a PCA score plot of the dataset with race/ethnicty groups identified by colored ellipses. Lastly, Figure 4.7 illustrates a PCA score plot of the dataset with age groups identified by colored ellipses. These figures visually illustrate differences in measurement variability for groups within each demographic category.

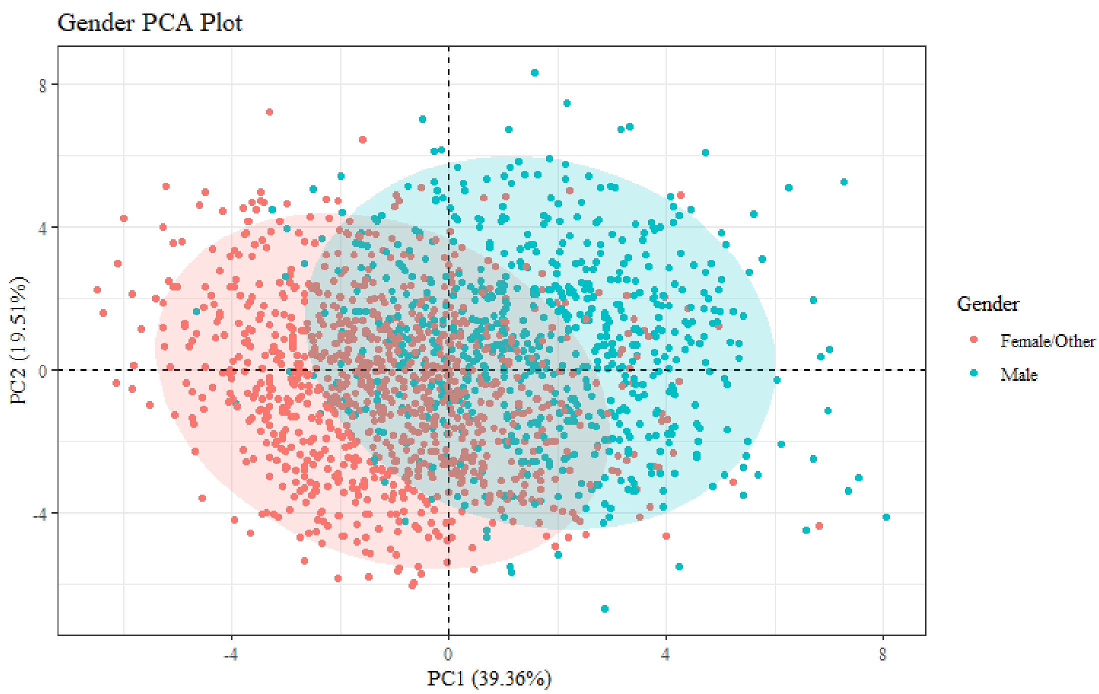

Figure 4.5: PCA score plot with gender category ellipses.

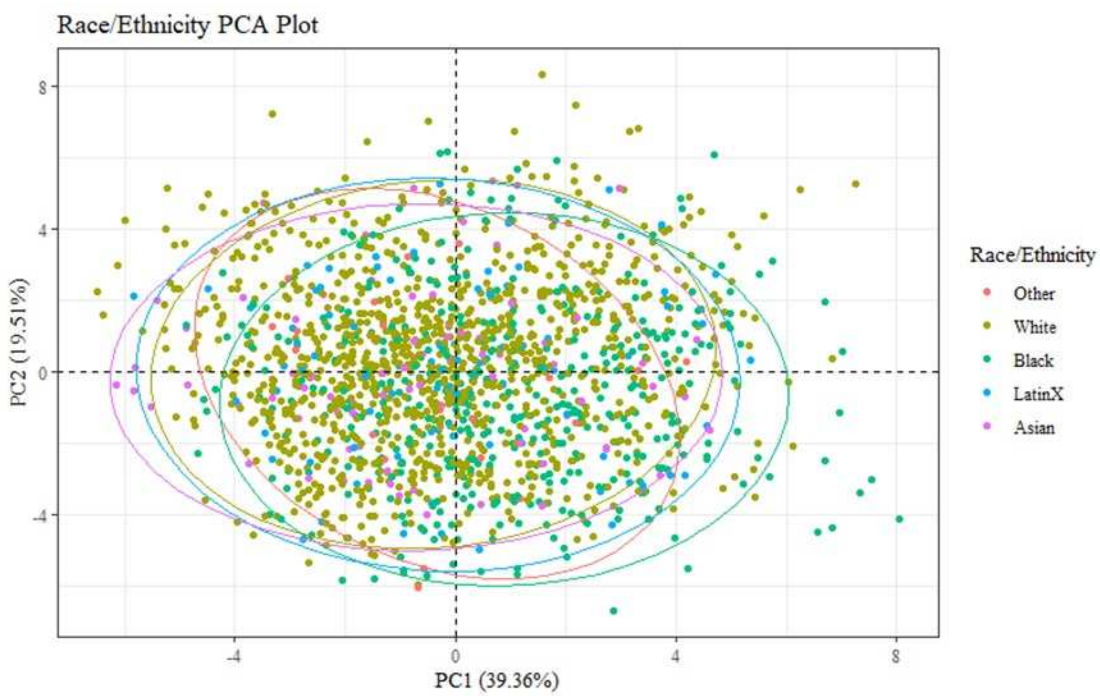

Figure 4.6: PCA score plot with race/ethnicity category ellipses.

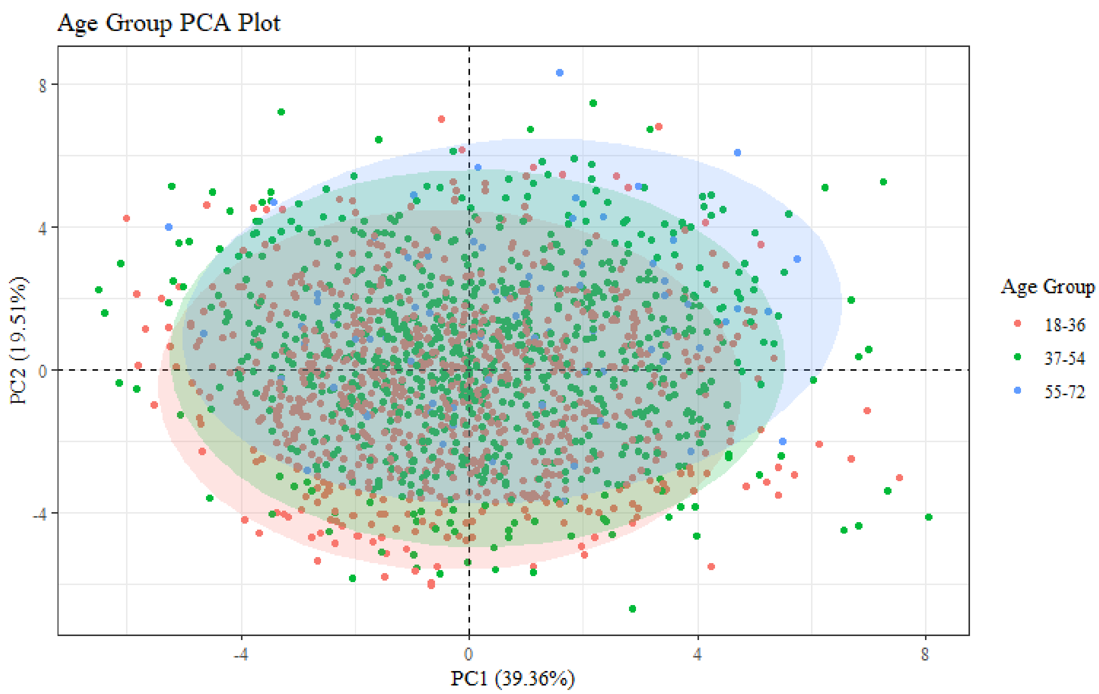

Figure 4.7: PCA score plot with age group category ellipses.

Multivariate Analysis of Variance (MANOVA)

The assumptions for MANOVA testing are independent observations, normality, homogeneity of covariances, and linear response. These assumptions were approximately satisfied within the present study's dataset. Due to low or no representation of some demographic interactions (e.g., there were no participants with race/ethnicity as Other, gender as Male, and age as 55-74), a Type I additive MANOVA model was used in this work. Table 4.6 provides the degrees of freedom (df), Pillai statistic, F-statistic, df1, df2, p-value, and significance status of each demographic factor from the MANOVA output (significance level: $p<0.05$). Based on the MANOVA analysis findings, there were significant measurement differences between groups within each demographic category. In other words, people in different groups within the demographic categories of gender, race/ethnicity, and age can be expected to have one or more different 3D facial measurements (of the 12 3D facial measurements assessed, Figure 4.1). Post hoc ANOVA testing was done to assess which of the

12 measurements were significantly different for each demographic group. ANOVA findings (F-Statistic, p-value, significance based on level p<0.05) presented in Table 4.7 revealed that the majority of the 12 measures assessed in this research were different for people of different gender, different race/ethnicity, or different age groups. Post hoc analyses presented in Chapter 5 explore the predicted measurement differences between measurements for each demographic category in metric units and percentages.

Table 4.6.
Findings from Type I Additive MANOVA Model (significance level = p<0.05).

Demographic Factor	df	Pillai	F-Statistic	df1	df2	p-value	Significant?
Gender	1	0.60	204.32	12	1,658	<0.01	TRUE
Race/Ethnicity	4	0.46	18.14	48	6,644	<0.01	TRUE
Age Group	2	0.17	13.03	24	3,318	<0.01	TRUE
Residuals	1,669	`				NA	

Table 4.7.
Findings from post hoc ANOVA tests (significance level = p<0.05). Degrees of freedom (df) for all 3D measurements: gender=1, race_eth=2, age_group=4.

3D Measurement	Demographic Factor	F-Statistic	p-value	Significant?
AA_C	Gender	581.86	<0.01	TRUE
	Race/Ethnicity	8.47	<0.01	TRUE
	Age Group	41.20	<0.01	TRUE
BiW_L	Gender	65.37	<0.01	TRUE
	Race/Ethnicity	20.60	<0.01	TRUE
	Age Group	3.45	<0.01	TRUE
BiW_C	Gender	210.08	<0.01	TRUE
	Race/Ethnicity	5.40	<0.01	TRUE
	Age Group	1.65	0.19	FALSE
GoSub_C	Gender	293.94	<0.01	TRUE
	Race/Ethnicity	15.74	<0.01	TRUE
	Age Group	80.96	<0.01	TRUE
NRB_L	Gender	8.73	<0.01	TRUE
	Race/Ethnicity	22.79	<0.01	TRUE
	Age Group	2.12	0.12	FALSE*
ProS_L	Gender	76.77	<0.01	TRUE
	Race/Ethnicity	45.15	<0.01	TRUE
	Age Group	33.71	<0.01	TRUE
SelP_L	Gender	357.65	<0.01	TRUE
	Race/Ethnicity	21.86	<0.01	TRUE
	Age Group	5.20	<0.01	TRUE
SelM_L	Gender	343.72	<0.01	TRUE
	Race/Ethnicity	18.68	<0.01	TRUE
	Age Group	5.41	<0.01	TRUE
SnasM_C	Gender	117.56	<0.01	TRUE
	Race/Ethnicity	61.24	<0.01	TRUE
	Age Group	2.65	0.07	FALSE*
TrSman_C	Gender	610.52	<0.01	TRUE
	Race/Ethnicity	21.67	<0.01	TRUE
	Age Group	74.92	<0.01	TRUE
TrTr_C	Gender	561.35	<0.01	TRUE
	Race/Ethnicity	33.29	<0.01	TRUE
	Age Group	9.65	<0.01	TRUE
TrTr_L	Gender	1,124.03	<0.01	TRUE
	Race/Ethnicity	17.44	<0.01	TRUE
	Age Group	6.69	<0.01	TRUE

*Tested to be significant with imputed data

Discussion

The results of the statistical analyses indicated important differences in face measurements between different groups within genders, races/ethnicities, or age categories. Table 4.8 compares the findings of previous studies (discussed in the Introduction) to the findings of this study. Overall, this study found differences in 3D measurements related to respirator fit based on demographic factors beyond what was found in previous literature.

The factor loadings provided in Table 4.5 and illustrated in Figure 4.4 indicate that PC1, accounting for 39.36% of the variability in the dataset, was most influenced by 3D measurement variables Sellion to Menton Linear, Tragion to Submandibular Contour, Tragion to Tragion Linear, and Tragion to Tragion Contour. These 3D measurements are large, across-face 3D measurements that the indicate overall shape and size of the face. Of note is the importance of the Tragion to Submandibular Contour measurement to PC1; this is a 3D measurement contextualizing face length that has not been assessed by previous literature, and thus offers a novel finding to this field of research.

PC2, accounting for 19.51% of the variability in the dataset, was most influenced by 3D measurement variables Alare to Alare Contour, Bizygomatic Width Linear, Pronasale to Subnasale Linear, and Subnasale to Menton Contour. These 3D measurements are generally smaller in metric length than those that influenced PC1. Compared to PC1, which captured large, positive variance, PC2 captured a more nuanced story of positive and negative variance in the dataset. The two largest PC2 factor loadings (provided in Table 4.5 and illustrated in Figure 4.4) indicate that Alare to Alare Contour, a nose width measure, and Pronasale to Subnasale Linear, a nose protrusion measure, affect the overall variance in the 3D facial measurement dataset more than the ten other 3D measurement location variables. Notably, the factor loadings for these 3D measurements, both of which are related to nose shape, were found to be negative. When PC2 scores (for the PCA score plots) were calculated for each observation or participant, these 3D measurements related to nose shape are minimized by the large negative factor loadings. Furthermore, PC3, accounting for 11.04% of the variability in the dataset, was very largely negatively affected by Sellion to Pronasale, a measure of nose bridge length. The influence of nose measurements on PCA are new a finding compared to other similar research, which have tended to find PC1 related to face length and PC2 related to face width (Zhuang et al., 2007, 2010).

Based on the MANOVA testing, different groups within demographic categories of gender, race/ethnicity, and age group can be expected to have significant differences in the 12 tested 3D facial measurements related to respirator fit. These findings show similarities and differences to previous literature findings, which are summarized in Table 4.8. Differences within each demographic category, including results of PCA, MANOVA, and ANOVA analyses, are discussed further below.

Differences between Gender Groups. In address of SA3.RQ1, differences in 3D measurements between gender groups are discussed. Based on the gender-grouped PCA score plot (Figure 4.5), gender appears to have the highest difference in variability between groups (Male vs. Female/Other) out of the three demographic categories (race/ethnicity, gender, age group). Furthermore, the gender-grouped PCA score plot indicates that a) Male faces may be quite larger overall than Female/Other faces (PC1) and b) Males may have slightly larger noses than those identifying as Female/Other (PC2). Based on the MANOVA findings presented in Table 4.6, people of different gender groups within the analyzed sample had significant differences in at least one of the 12 analyzed 3D facial measurements that relate to respirator fit. These results mirror those from Zhuang et al. (2010), who found gender to be the most impactful demographic factor in predicting differences in face size (compared to race/ethnicity and/or age). Post hoc ANOVA analysis (Table 4.7) found that all 12 3D measurements tested were significantly different between people of different genders. Previous relevant research efforts have not found all tested measurements to be significantly different between demographic groups of gender (Brazile et al., 1998), however, this may be attributed to smaller sample size and different measurements collected compared to the present study.

Differences between Race/Ethnicity Groups. In address of SA3.RQ2, differences in 3D measurements between race/ethnicity groups are discussed. The race/ethnicity-grouped PCA score plot (Figure 4.6) indicated differences between race/ethnicity groups in overall face size (PC1) in order from smallest to largest (actual metric size): Asian, LatinX, White, Other, Black. However, the race/ethnicity-grouped PCA score plot illustrated minimal differences in nose size (PC2) between the five race/ethnicity groups. Based on the MANOVA findings presented in Table 4.6, people of different race/ethnicity groups within the analyzed sample had significant differences in at least one of the 12 analyzed 3D facial measurements that relate to respirator fit. These results mirror those from Zhuang et al. (2010), who found significant differences in face size between different race/ethnicity groups. Post hoc ANOVA analysis (Table 4.7) found that all 12 3D measurements tested were significantly different between people of different race/ethnicity groups. Previous relevant research efforts have found not all tested measurements to be significantly different between demographic groups of race/ethnicity (Brazile et al., 1998), however, this may be attributed to smaller sample size and different measurements collected compared to the present study.

Differences between Age Groups. In address of SA3.RQ3, differences in 3D measurements between age groups are discussed. The age-grouped PCA score plot (Figure 4.7) indicates differences in variability between youngest (18-29), mid-age (37-54), and oldest (55-74) age groups. The age-grouped PCA score plot indicated some age-related differences in overall face size (PC1) and nose size (PC2): the youngest age group had the smallest faces and noses, and the oldest age group had the largest faces and noses (with the mid-age group between the two). Based on the MANOVA findings presented in Table 4.6, people of different age groups within the analyzed sample had significant differences in at least one of the 12 analyzed 3D facial measurements that relate to respirator fit. Similar to the findings of the present study, Zhuang et al. (2010) found measurement differences between face size for people in three age groups, although age brackets were assigned somewhat differently with the oldest group being >45. In the present study, ANOVA testing found 9 of the 12 3D measurement locations to be significantly different for people of different age groups. Of relevance to the literature is the finding that Bizygomatic Width Contour (a 3D measurement indicating face width) was not significantly different (testing with both non-imputed and imputed data) for people of different age groups. Zhuang et al. (2010) found that their oldest participants (>45) had longer and narrower faces than their youngest participants (18-29). Therefore, results from this study contradict the results of Zhuang et al. (2010) in that older participants did not have significantly narrower faces in this study.

Table 4.8.
Findings in previous literature compared to findings from this study.

Source	Findings from previous studies regarding differences in facial measurements related to respirator fit	Finding/s from this study regarding differences in facial measurements related to respirator fit
Hack et al., 1973 (LANL)	No measurement differences (<2mm) between facial measurements of participants (n=200, males only, 40% Spanish-American).	Significant measurement differences between demographic groups.
Leigh, 1975	12.6% of participants (n=1467, 120 females) were not represented by LANL.	Significant measurement differences between gender groups but did not compare measurements to LANL.
Gross & Horstman, 1990	5% of participants (n=120) were not able to fit nine selected respirators.	This study did not use respirators to assess fit.
Oestenstad & Perkins, 1992	Measurements did not differ from previous research (n=68).	Significant measurement differences between demographic groups but did not compare measurements to previous studies.
Brazile et al., 1998	Significant measurement differences between gender and race/ethnicity groups (n=186).	Significant measurement differences between gender and race/ethnicity groups, as well as age groups.
Kim et al., 2003	Significant measurement differences between gender (male vs female) for Korean people, as well as Korean people and people of other origins (n=110).	Significant measurement differences between gender and race/ethnicity groups but did not collect information about nationality.
Ball et al., 2010	Measurement differences in head shape between Chinese people and White people (n=1200, males only).	Significant measurement differences in race/ethnicity groups but did not collect information about nationality.
Luximon et al., 2010	From summary statistics, measurement differences between specifically Chinese females and people of other origins (n=772, females only).	Significant measurement differences between gender and race/ethnicity groups but did not collect information about nationality.
Zhuang et al., 2010	Significant measurement differences between gender (male vs. female), all racial/ethnic groups, all sampled occupations, and those aged >45 compared to those aged 18-29 (n=3997).	Significant measurement differences between gender, race/ethnicity, and age groups but did not collect information about occupation.
Spies et al., 2011	86% of South African participants (n=29) were not able to fit a size medium disposable respirator.	Significant measurement differences between race/ethnicity groups but did not collect information about nationality.
Lee et al., 2012	Measurement differences between Korean male pilots and civilians, US pilots, and Korean female pilots (n=336).	Significant measurement differences between gender and race/ethnicity groups but did not collect information about nationality or occupation.
Zhang et al., 2020	Significant measurement differences between Chinese males and females (n=85).	Significant measurement differences between race/ethnicity groups but did not collect information about nationality.
Rodríguez et al., 2020	Measurements of Chilean people are comparable to measurements found in previous research (n=474).	Significant measurement differences between demographic groups but did not a) collect information about nationality or b) compare measurements to previous studies.

Limitations

PCA cannot be used to assess statistical significance due to the lack of formal testing, thus results are open to researcher interpretation. Despite best efforts to assess PCA results in the most logical way, the interpretation of PCA results from the present study should be viewed somewhat as opinion. All anthropometric data collection and research efforts have limitations regarding the diversity of their sample population, with the present study being no exception. Due to model complexity and low representation of certain groups, interactions could not be included in the MANOVA model. By instead using an additive MANOVA model, this research could only determine the presence of significant differences between each demographic group within a single demographic category (as opposed to differences across groups, i.e., if White mid-age Females have different facial measurements than youngest LatinX males). Furthermore, previous research efforts have considered the nationality and occupation of their sample population, which this study did not. Despite these limitations, this research contributes to the knowledge base surrounding respirator-specific 3D facial anthropometrics and demographically-related differences in these measurements.

Conclusions

In address of SA3, the present study utilized a large sample of 2022 3D facial scans to assess demographic differences in 3D measurements related to respirator fit. This work has practical implications for the designers who develop and size respirators, professionals who fit respirators, workers who utilize respirators in their daily work, and researchers who study facial anthropometrics (specifically in relation to respirator fit). Furthermore, this work utilized 3D measurement data, which may have novel practical implications for designers and researchers interested in 3D scanning and anthropometrics. For example, this work found that a novel 3D measurement related to face width was able to predict a large amount of variability in the entire dataset.

In agreement with previously published research, people of different gender, race/ethnicity, or age groups had significantly different face measurements related to respirator fit. Unlike previous studies, this study found that 1) nose shape was negatively predictive of variation in the facial anthropometric dataset, 2) all measurements tested were significantly different for different groups within gender and race/ethnicity categories, and 3) face width was not significantly different between age groups. Future research is needed to continue to assess if diverse demographic factors have significant effects on facial measurements and 3D facial measurements specifically.

CHAPTER 5: POST HOC ANALYSES: QUANTIFICATION OF SIGNIFICANT DIFFERENCES IN 3D FACIAL MEASUREMENTS

Summary

The results of post hoc analyses, which quantified 3D facial measurement differences between demographic groups (within the larger demographic categories of gender, race/ethnicity, and age group), are presented below. This chapter presents the methods by which post hoc analyses were completed, and the results of these analyses. Similar to the practical implications of providing population summary statistics (Chapter 3), is expected that the results of these analyses will have practical implications for respirator designers and manufacturers.

Methods

The 12 3D measurement locations examined for significant differences in Chapter 4 are presented again in Figure 5.1 and Table 5.1. For post hoc analyses presented in this chapter, quantification of differences in the 12 3D measurements was calculated using estimated marginal means for each combination of demographic groups (within a single demographic category) if found significant in Chapter 4 ANOVA testing. Tukey's testing was used to calculate the estimated difference in estimated marginal means as well as the significance level of the estimate ($p<0.05$). Estimated marginal means are calculated using modeled data, allowing them to have less bias than summary statistic group means. When difference estimates in estimated marginal means are calculated for each group in a demographic category using Tukey's testing, all other demographic categories are averaged over (i.e., when calculating estimated differences between Male and Female/Other groups, race/ethnicity groups and age groups are averaged over). Only significant results (Tukey testing $p<0.05$) are reported in the subsequent sections of this chapter. For analysis, the group with the largest sample size in each demographic category served as the reference or baseline group (i.e., for gender, Female/Other was had the largest sample size and served as reference for testing). For reporting, the demographic group with the larger face measurements (larger in terms of metric measurements) is listed first in each section (i.e., for Male - Female/Other, Males generally have larger face measurements overall, and are therefore listed first). Post hoc statistical analyses were completed using R software (R Core Team, 2022c) and packages tidyverse, readxl, extrafont, flextable, and scales (Chang, 2022; Gohel, 2022; Wickham, 2022; Wickham & Bryan, 2022; Wickham & Seidel, 2022).

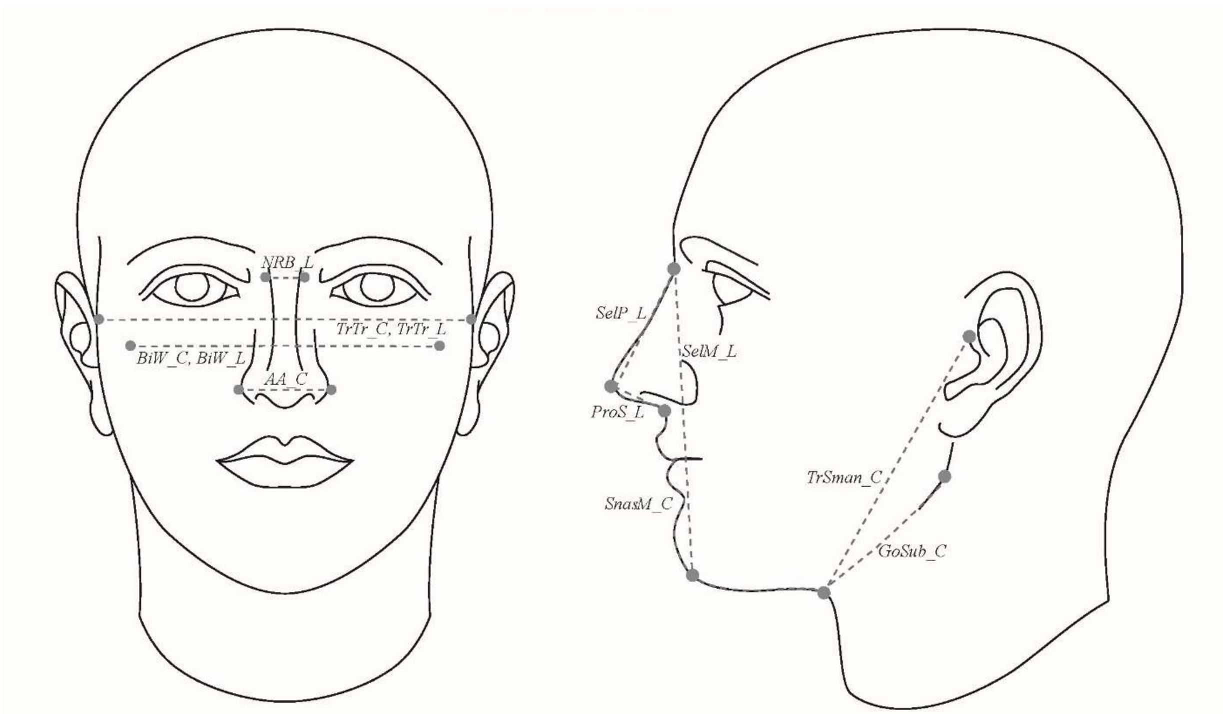

Figure 5.1: Illustration of 12 respirator-related 3D facial measurements collected from each 3D scan.

Table 5.1.
Measurement names (corresponding to Figure X), measurement type (linear, contour, or both), and abbreviated measurement name.

Measurement Name	Measurement Type	Abbreviated Name
Alare to Alare	Contour	AA_C
Bizygomatic Width	Contour	BiW_C
Bizygomatic Width	Linear	BiW_L
Gonion to Submandibular	Contour	GoSub_C
Nasal Root Breadth	Linear	NRB_L
Pronasale to Subnasale	Linear	ProS_L
Sellion to Pronasale	Linear	SelP_L
Sellion to Menton	Linear	SelM_L
Subnasale to Menton	Contour	SnasM_C
Tragion to Gonion	Contour	TrGo_C
Tragion to Sellion	Contour	TrSel_C
Tragion to Submandibular	Contour	TrSman_C
Tragion to Tragion	Contour	TrTr_C
Tragion to Tragion	Linear	TrTr_L

Results

Gender

Male – Female/Other. Based on ANOVA testing in Chapter 4, all 12 3D measurements were found to be significantly different for gender. The gender demographic category was comprised of two gender groups: Male, and Female/Other. Sample sizes for these gender groups are presented in Table 5.2. Tukey's testing indicated that all 12 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement differences between Male and Female/Other gender groups are presented in the context of millimeters in Figure 5.2. Significant differences in estimated marginal means between Male and Female/Other gender groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.3 and 5.4.

Table 5.2.

Gender makeup of 3D face scan dataset.

Gender	n (total n=1677)
Female/Other*	996
Male	681

* Female/Other represents participants who self-identified as Female (n=994), non-binary or other (n=1), and prefer not to say (n=1).

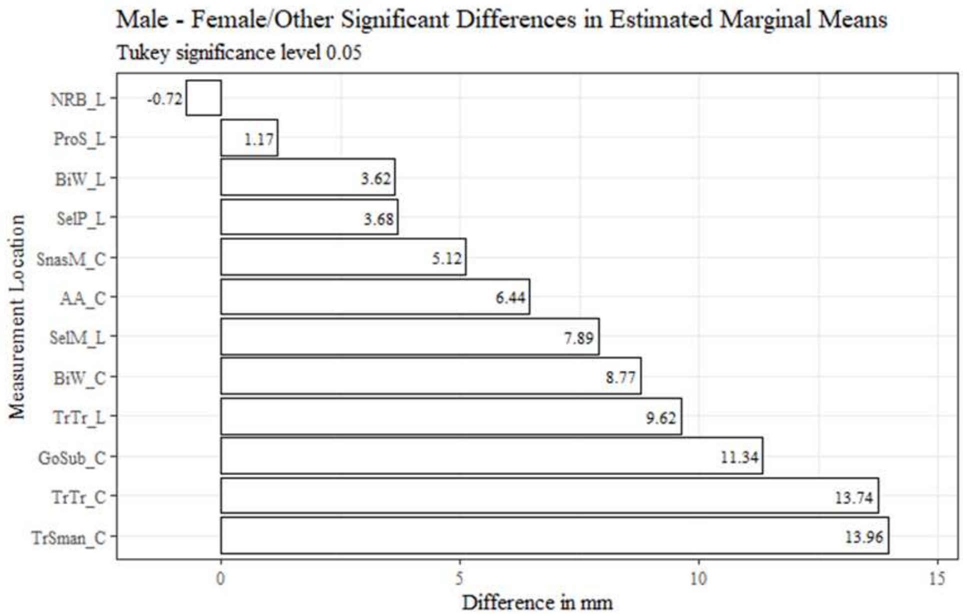

Figure 5.2: Significant differences in estimated marginal means between Male and Female/Other gender groups. Difference estimates are presented in millimeters.

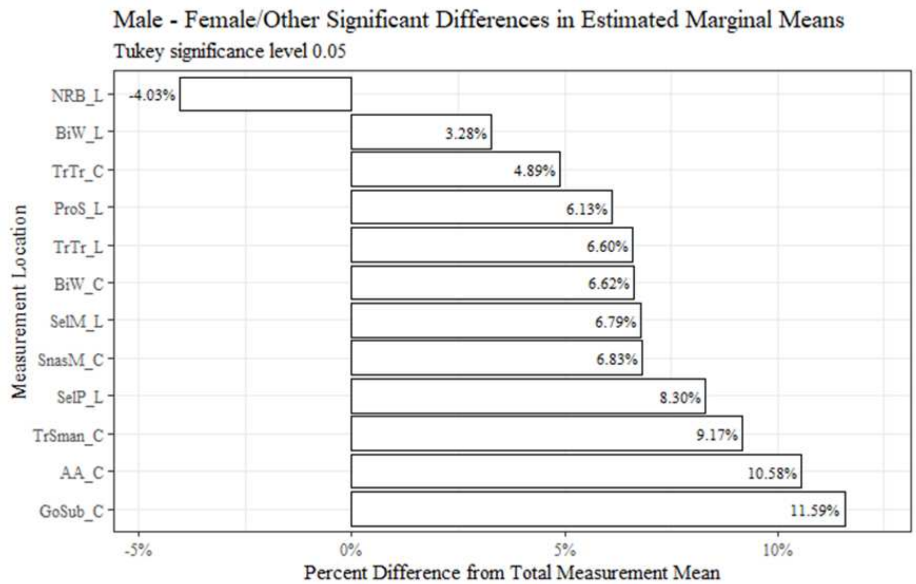

Figure 5.3: Significant differences in estimated marginal means between Male and Female/Other gender groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

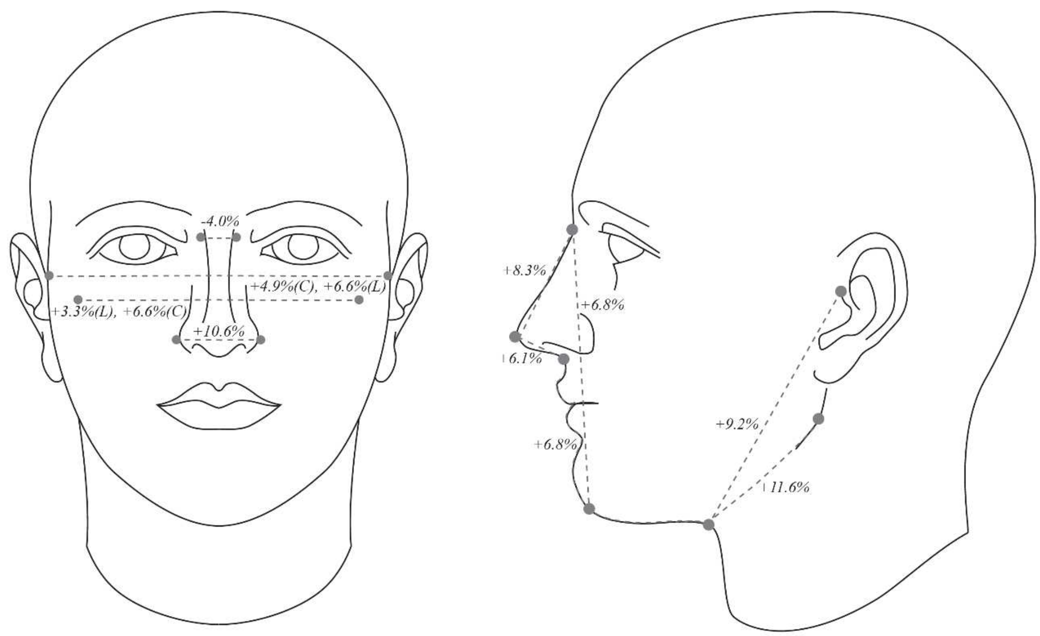

Figure 5.4: Significant differences in estimated marginal means between Male and Female/Other gender groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Race/Ethnicity

Based on ANOVA testing in Chapter, all 12 3D measurements were found to be significantly different for race/ethnicity. Quantified differences between each race/ethnicity group, as found to be significant ($p<0.05$) by Tukey's testing, are described below in order of highest to lowest number of found differences. The race/ethnicity demographic category was comprised of five groups (listed in order of group sample size): White, Black, LatinX, Asian, and Other. Sample sizes for these groups are presented in Table 5.3.

Table 5.3.
Racial/ethnic makeup of 3D face scan dataset.

Race/ethnicity	Abbreviated term	n (total n=1677)
White/Caucasian	White	1040
Black, African, or African American	Black	446
Latin/Hispanic	LatinX	84
Asian/Asian American	Asian	81
Other*	Other	26

*Due to small sample size, "Other" represents participants who self-identified as Other (n=13), American Indian or Alaska Native (n=6), Native Hawaiian or Other Pacific Islander (n=3), or Prefer not to say (n=4).

Black - White. Between Black and White race/ethnicity groups, Tukey's testing indicated that 10 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). Based on the principal components analysis (PCA) score plotting in Chapter 4, differences in variability between Black and White race/ethnicity groups were indicated to be less significant than those between a) Black and Asian and b) Black and LatinX groups. However, Black and White race/ethnicity groups had the highest number of significantly different measurements between any of the race/ethnicity groups according to Tukey's testing. The significant measurement differences between Black and White race/ethnicity groups are presented in the context of millimeters in Figure 5.5. Significant differences in estimated marginal means between Black and White race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.6 and 5.7.

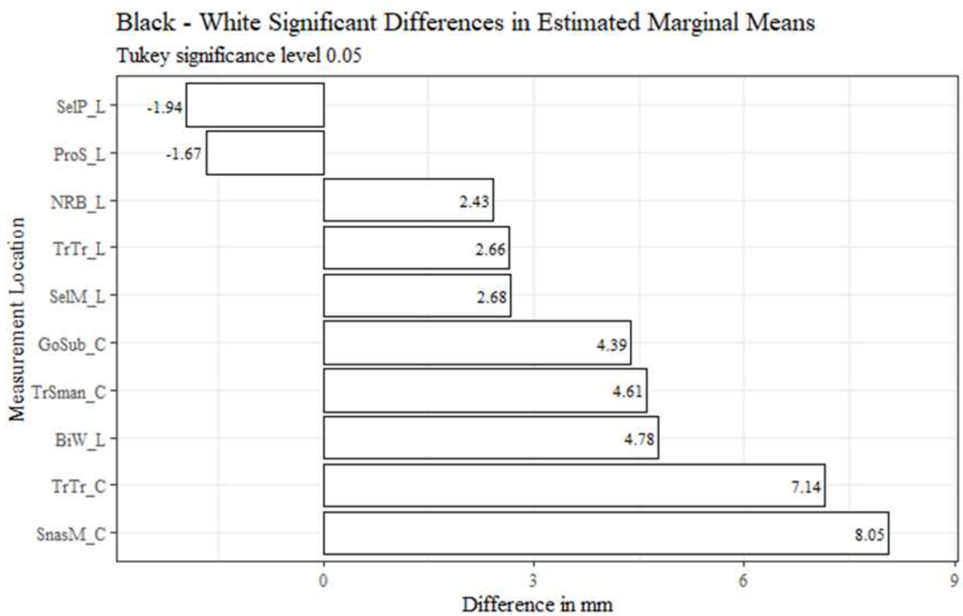

Figure 5.5: Significant differences in estimated marginal means between Black and White race/ethnicity groups. Difference estimates are presented in millimeters.

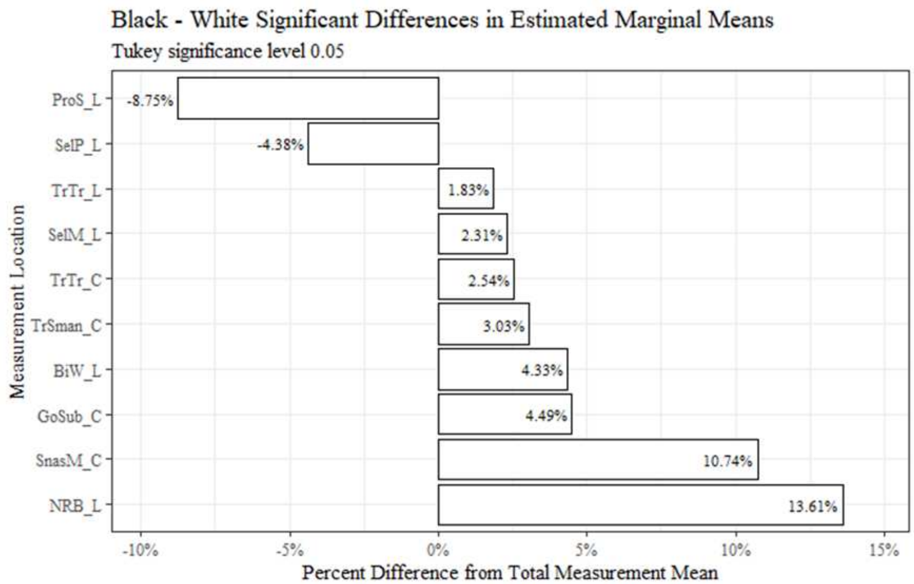

Figure 5.6: Significant differences in estimated marginal means between Black and White race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

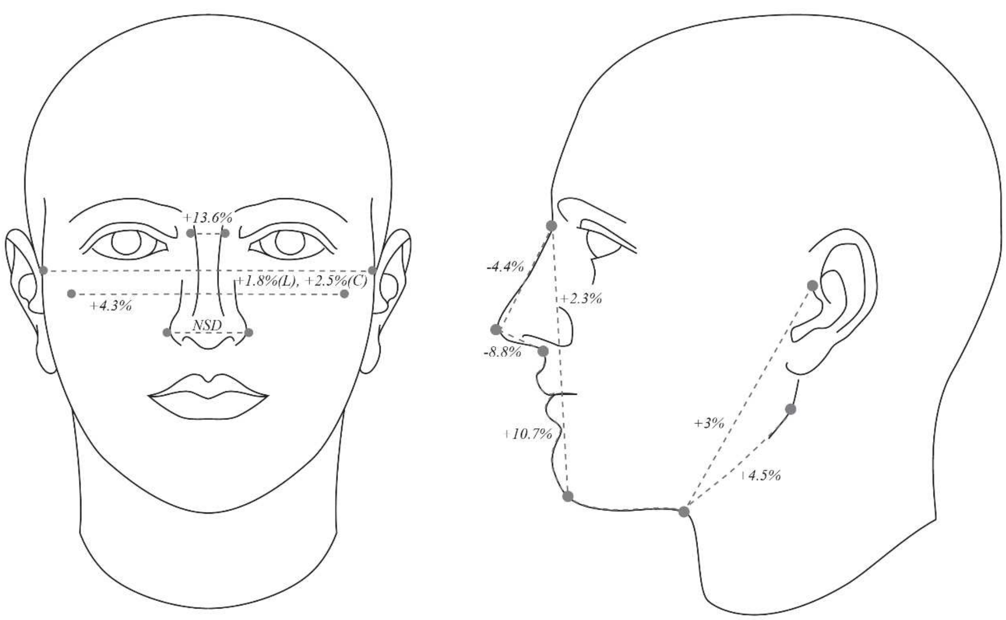

Figure 5.7: Significant differences in estimated marginal means between Black and White race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Black - Asian. Between Black and Asian race/ethnicity groups, Tukey's testing indicated that 9 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). In Chapter 4, PCA score plotting indicated that differences in variability were highest between Black and Asian race/ethnicity groups. The significant measurement differences between Black and Asian race/ethnicity groups are presented in the context of millimeters in Figure 5.8. Significant differences in estimated marginal means between Black and Asian race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.9 and 5.10.

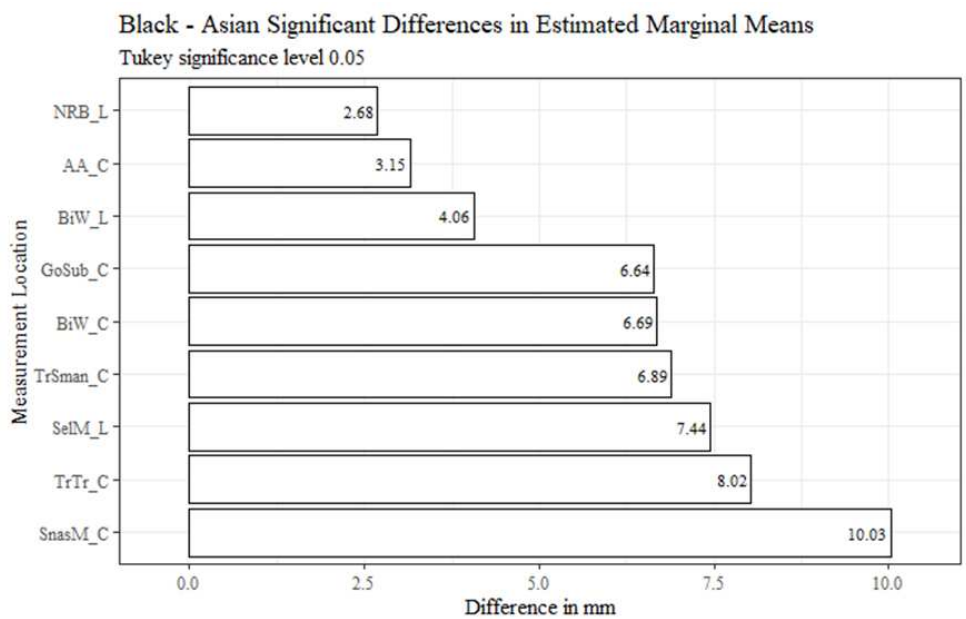

Figure 5.8: Significant differences in estimated marginal means between Black and Asian race/ethnicity groups. Difference estimates are presented in millimeters.

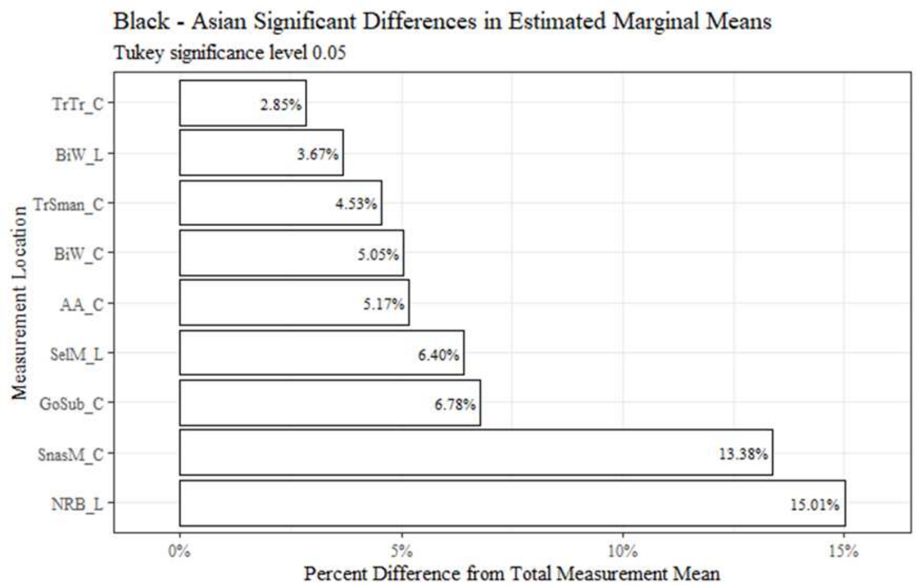

Figure 5.9: Significant differences in estimated marginal means between Black and Asian race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

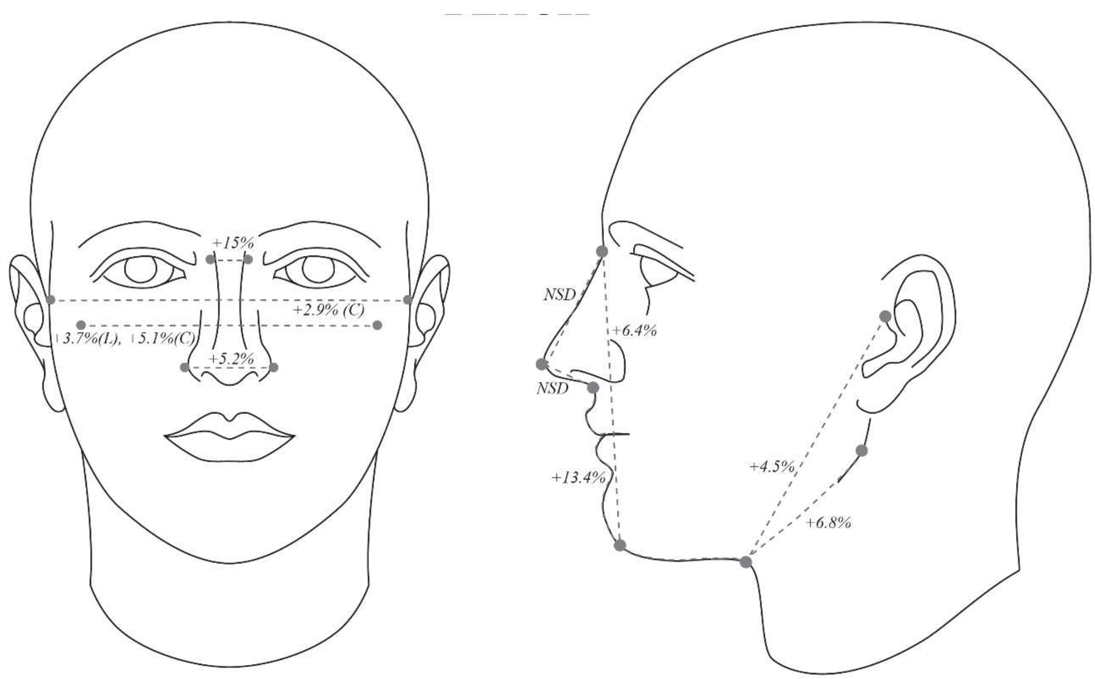

Figure 5.10: Significant differences in estimated marginal means between Black and Asian race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Black - LatinX. Between Black and LatinX race/ethnicity groups, Tukey's testing indicated that 9 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). In Chapter 4, PCA score plotting indicated that differences in variability were second highest between Black and LatinX race/ethnicity groups. The significant measurement differences between Black and LatinX race/ethnicity groups are presented in the context of millimeters in Figure 5.11. Significant differences in estimated marginal means between Black and LatinX race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.12 and 5.13.

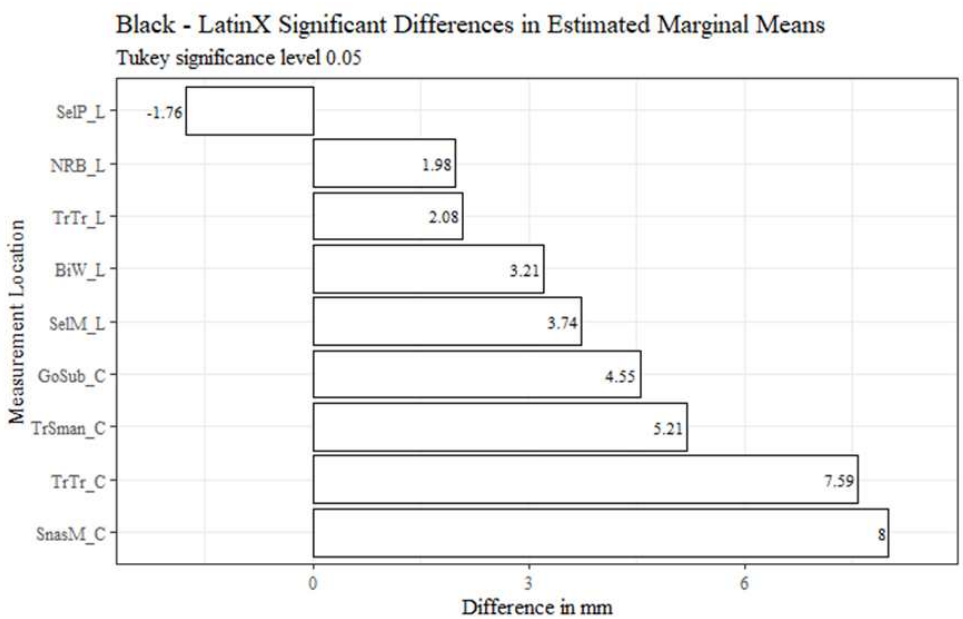

Figure 5.11: Significant differences in estimated marginal means between Black and LatinX race/ethnicity groups. Difference estimates are presented in millimeters.

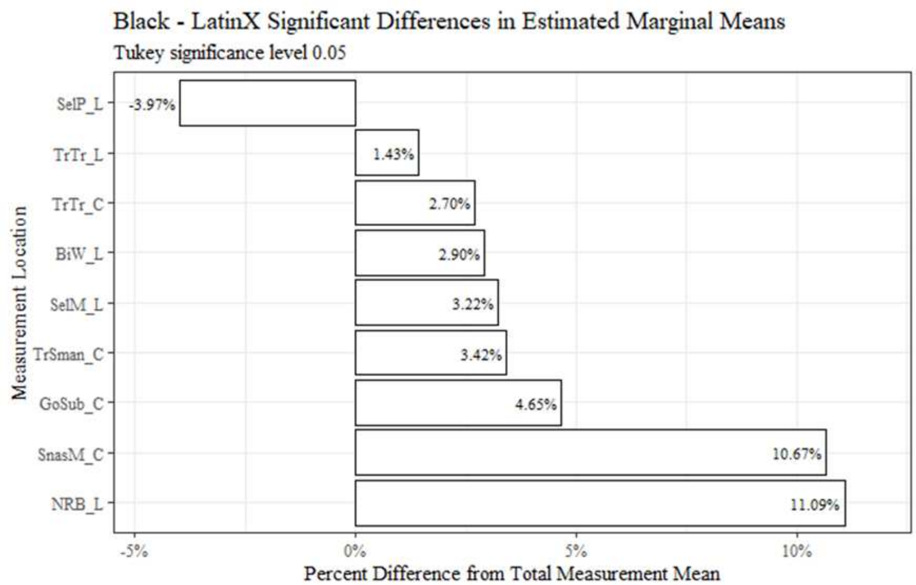

Figure 5.12: Significant differences in estimated marginal means between Black and LatinX race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

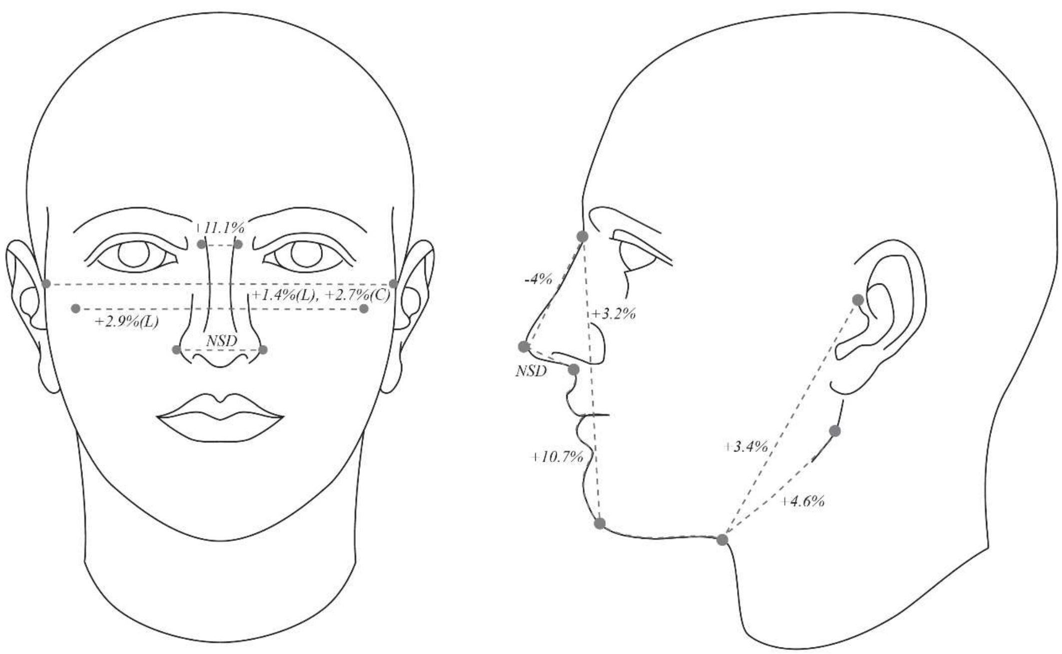

Figure 5.13: Significant differences in estimated marginal means between Black and LatinX race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, $p<0.05$).

White - Asian. Between White and Asian race/ethnicity groups, Tukey's testing indicated that 5 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement differences between White and Asian race/ethnicity groups are presented in the context of millimeters in Figure 5.14. Significant differences in estimated marginal means between White and Asian race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.15 and 5.16.

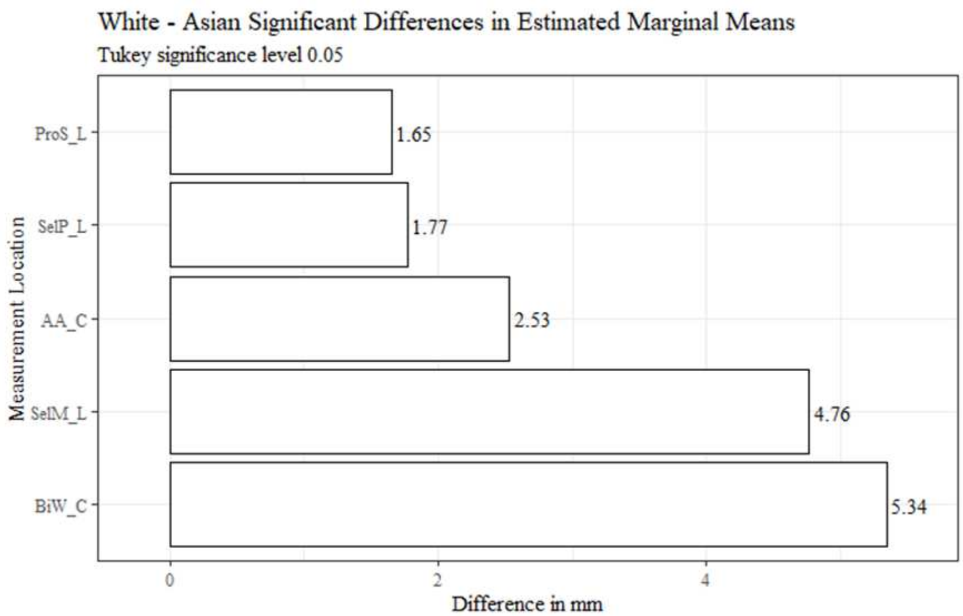

Figure 5.14: Significant differences in estimated marginal means between White and Asian race/ethnicity groups. Difference estimates are presented in millimeters.

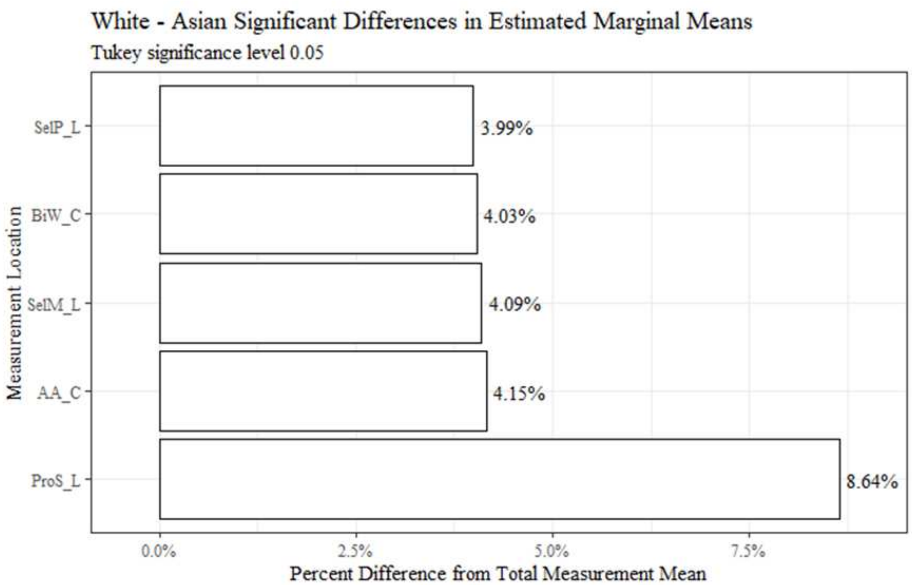

Figure 5.15: Significant differences in estimated marginal means between White and Asian race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

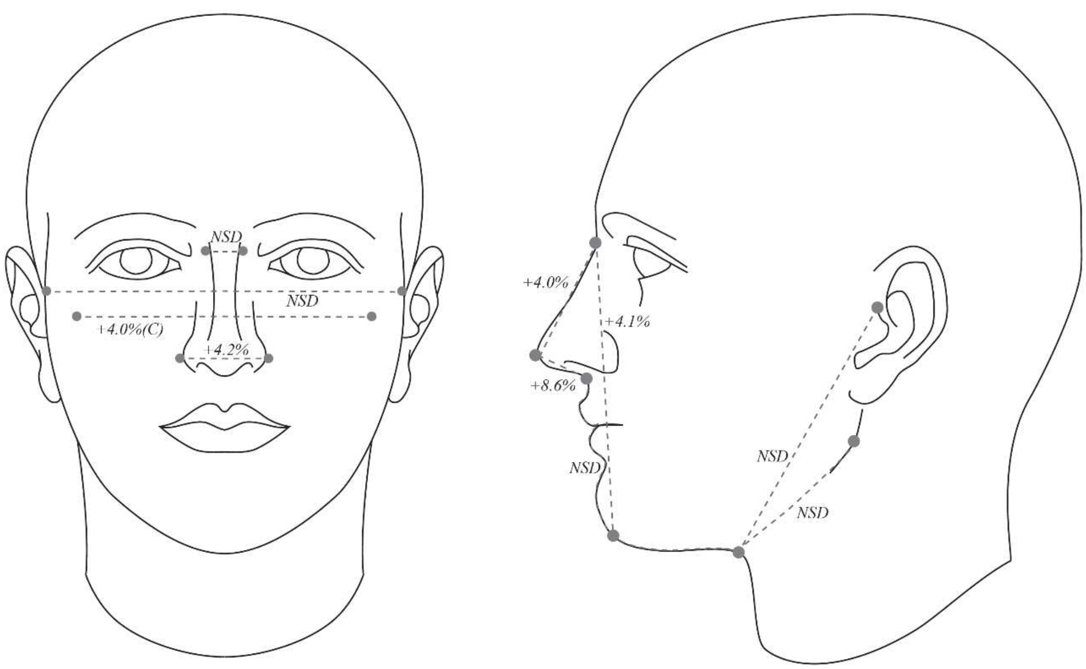

Figure 5.16: Significant differences in estimated marginal means between White and Asian race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Black - Other. The Other race/ethnicity category is comprised of people who self-identified their race/ethnicity as Other (n=13), American Indian or Alaska Native (n=6), Native Hawaiian or Other Pacific Islander (n=3), or Prefer not to say (n=4). Between Black and Other race/ethnicity groups, Tukey's testing indicated that 4 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement differences between Black and Other are presented in the context of millimeters in Figure 5.17. Significant differences in estimated marginal means between Black and Other race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.18 and 5.19.

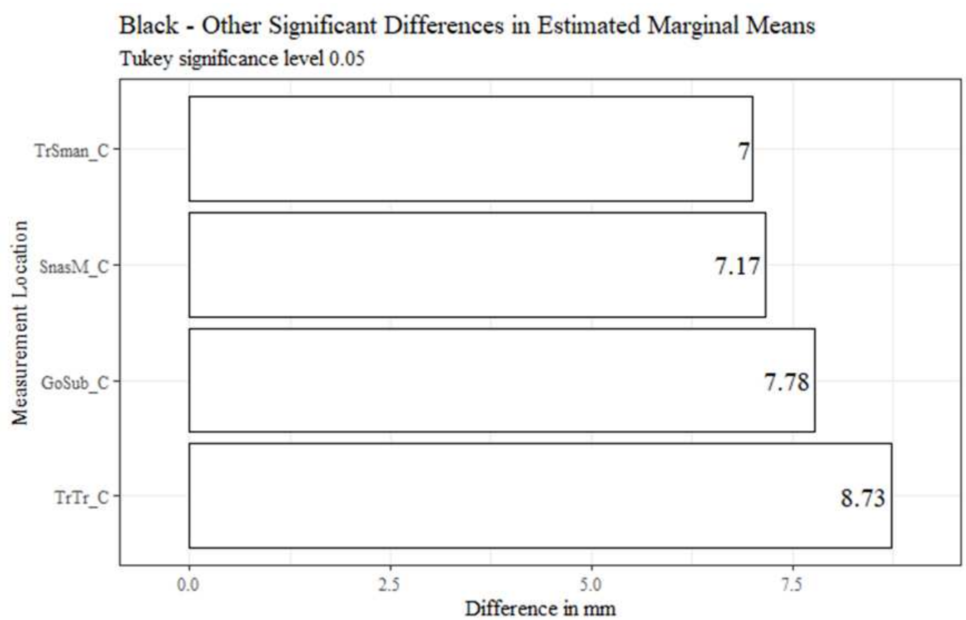

Figure 5.17: Significant differences in estimated marginal means between Black and Other race/ethnicity groups. Difference estimates are presented in millimeters.

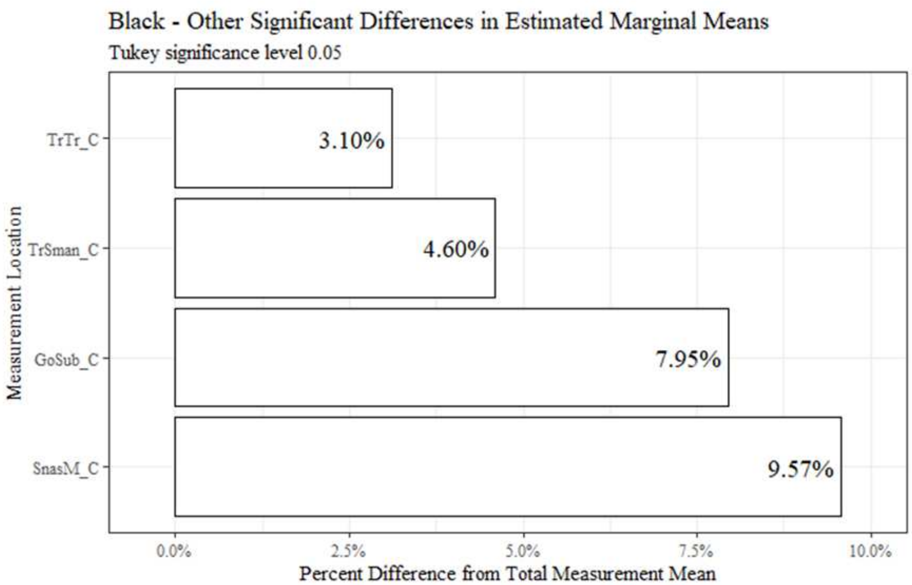

Figure 5.18: Significant differences in estimated marginal means between Black and Other race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

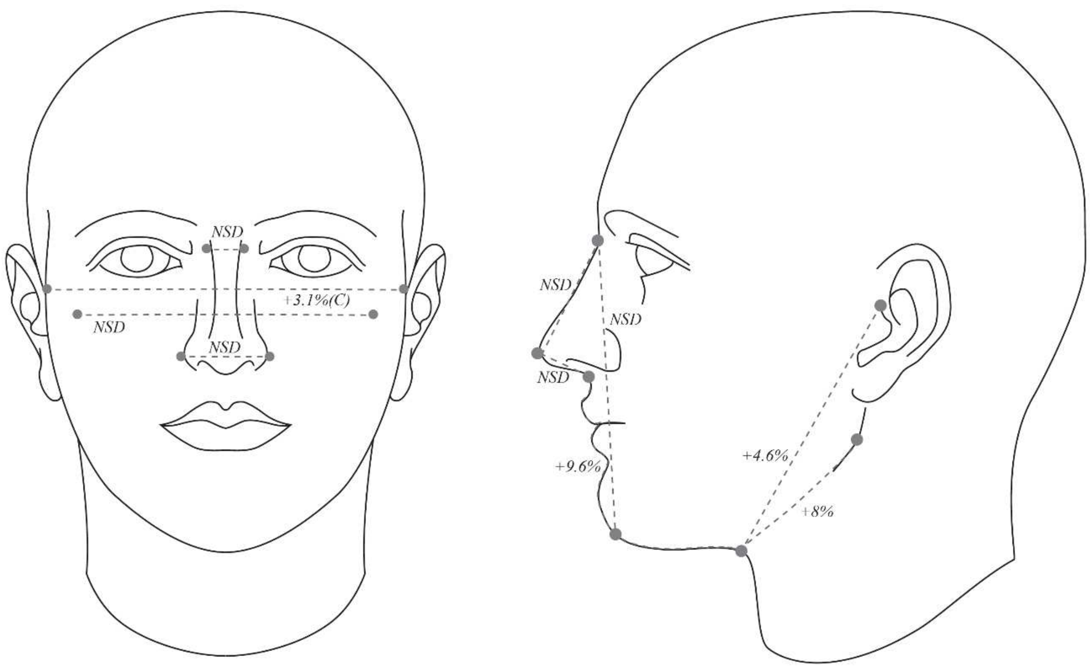

Figure 5.19: Significant differences in estimated marginal means between Black and Other race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

LatinX - Asian. Between LatinX and Asian race/ethnicity groups, Tukey's testing indicated that 3 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement differences between LatinX and Asian race/ethnicity groups are presented in the context of millimeters in Figure 5.20. Significant differences in estimated marginal means between LatinX and Asian race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.21 and 5.22.

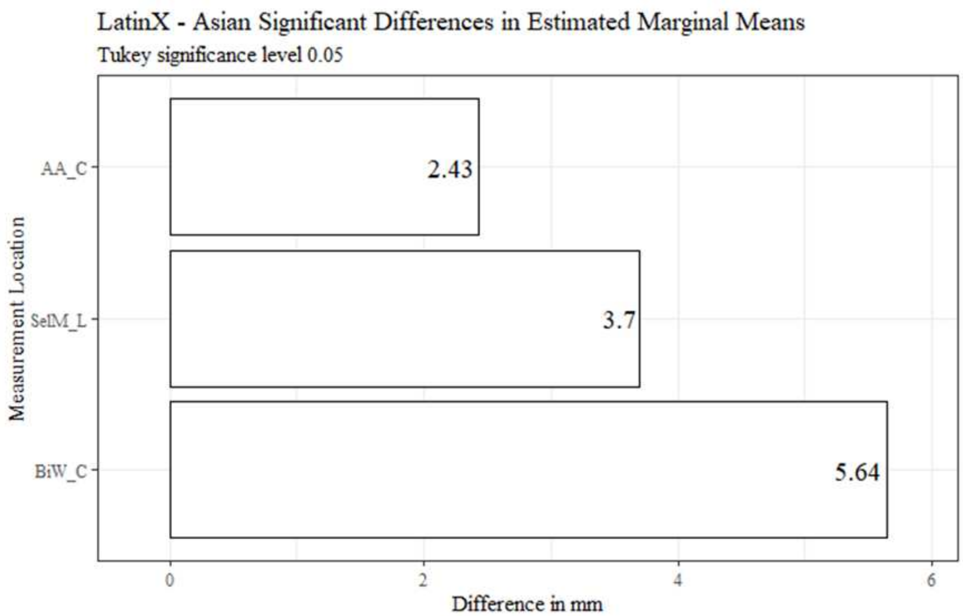

Figure 5.20: Significant differences in estimated marginal means between LatinX and Asian race/ethnicity groups. Difference estimates are presented in millimeters.

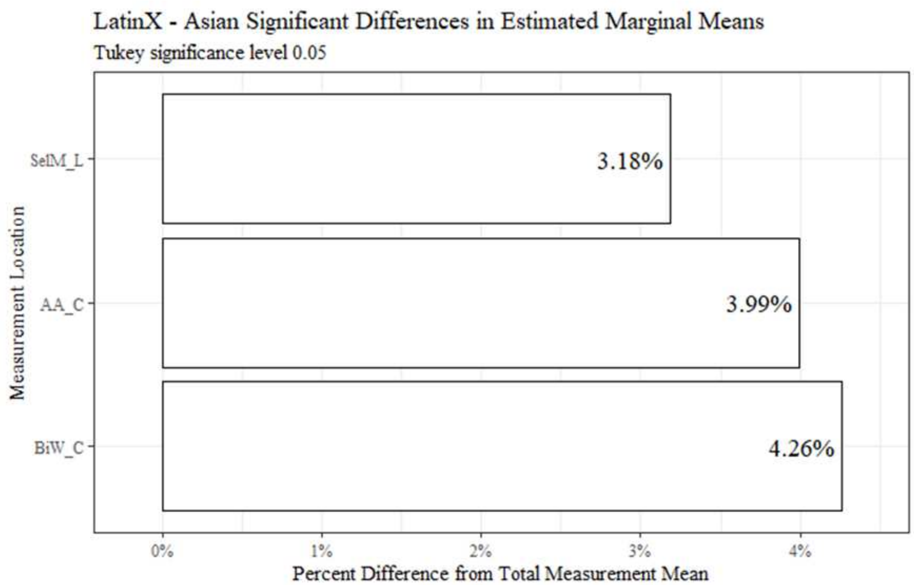

Figure 5.21: Significant differences in estimated marginal means between LatinX and Asian race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

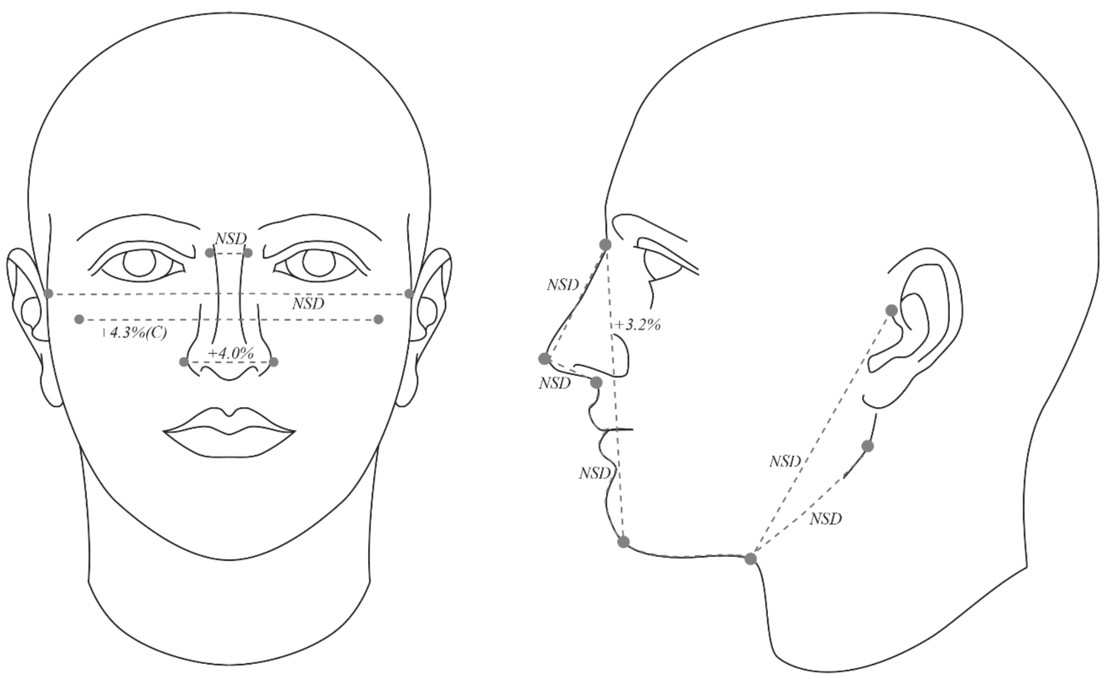

Figure 5.22: Significant differences in estimated marginal means between LatinX and Asian race/ethnicity groups (Asian as reference group). Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

White – Other. Between White and Other race/ethnicity groups, Tukey's testing indicated that 1 out of 12 measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement difference between groups was for the Pronasale to Subnasale Linear measurement, which was different by 2.11mm. The difference in Pronasale to Subnasale Linear estimated marginal means between White and Other race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, is presented in Figure 5.23.

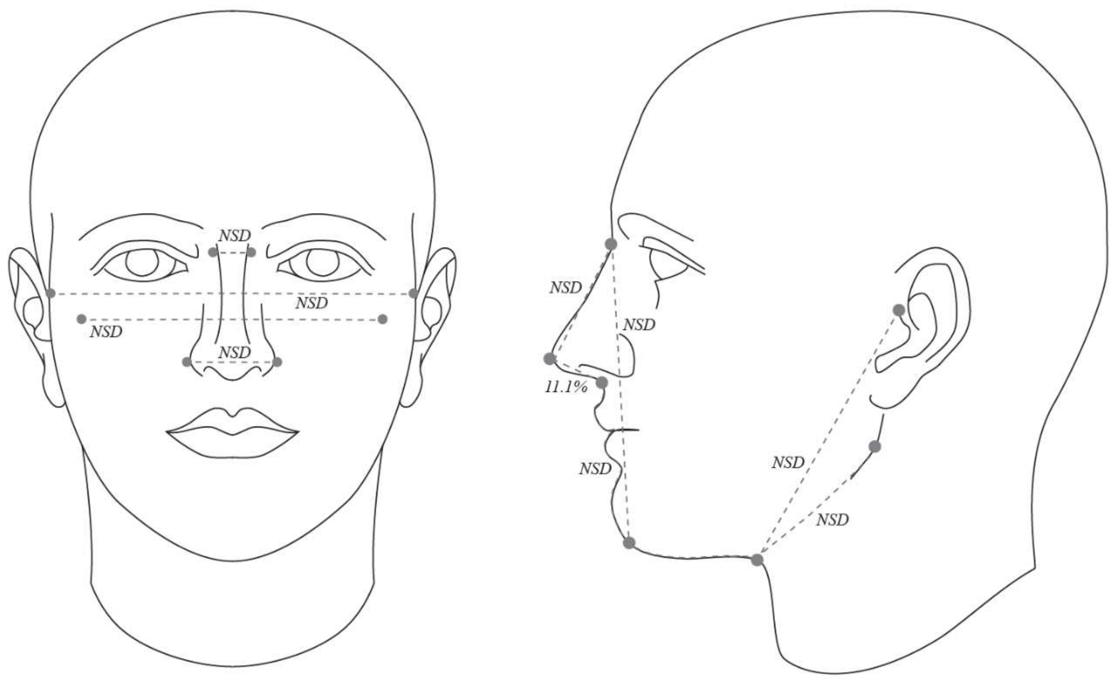

Figure 5.23: Significant differences in estimated marginal means between White and Other race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, $p<0.05$).

White – LatinX. Between White and LatinX race/ethnicity groups, Tukey's testing indicated that 1 out of 12 measurements had a significant difference in estimated marginal means p<0.05). The significant measurement difference between groups was for the Pronasale to Subnasale Linear measurement, which was different by 1.09mm. The difference in Pronasale to Subnasale Linear estimated marginal means between White and LatinX race/ethnicity groups, as a percentage of the mean of the measurement for the entire sample population, is presented in Figure 5.24.

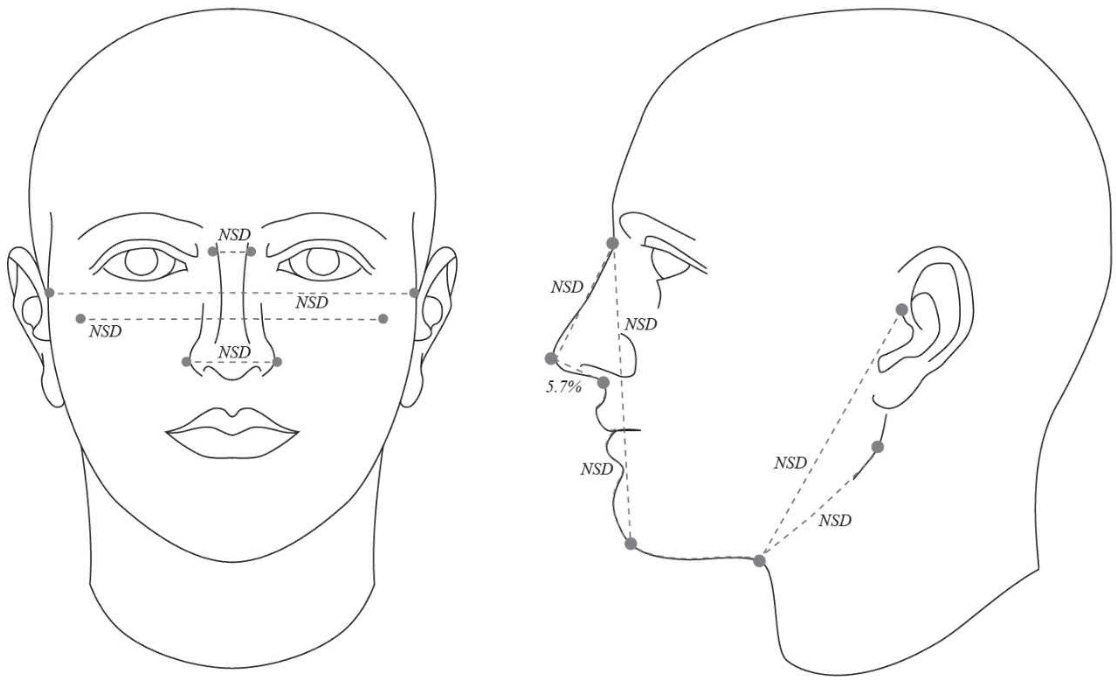

Figure 5.24: Significant differences in estimated marginal means between White and LatinX race/ethnicity groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Age Group

Based on ANOVA testing in Chapter 4, all 9 out of 12 measurements were found to be significantly different for age group (11 out of 12 using imputed data). Quantified differences between each age group, as found to be significant ($p<0.05$) by Tukey's testing, are described below in order of highest to lowest number of found differences. The age group demographic category was comprised of three groups: 18-36 (referred to as the youngest), 37 to 54 (referred to as mid-age), and 55-72 (referred to as the oldest). Sample sizes for these groups are presented in Table 5.4.

Table 5.4.
Age makeup of 3D face scan dataset.

Age	Term	n (total n=1677)
18-34	Youngest	826
35-54	Mid-age	777
55-72	Oldest	74

*Age was given by participants as exact numeric and subsequently divided into three groups for data analysis. Group limits were developed based on the oldest participant's age and approximately equal age spacing in each group.

Mid-age - Youngest. Between mid-age and the youngest age groups, Tukey's testing indicated that 8 out of 9 tested (based on ANOVA significance) and 12 total measurements had significant differences in estimated marginal means ($p<0.05$). Based on the principal components analysis (PCA) score plotting in Chapter 4, differences in variability between the mid-age and youngest age groups were indicated to be less significant than those between the oldest and youngest groups. However, mid-age and youngest age groups had highest number of significantly different measurements between any of the age groups using Tukey's testing. The significant measurement differences between the mid-age and youngest age groups are presented in the context of millimeters in Figure 5.25. Significant differences in estimated marginal means between the mid-age and youngest age groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.26 and 5.27.

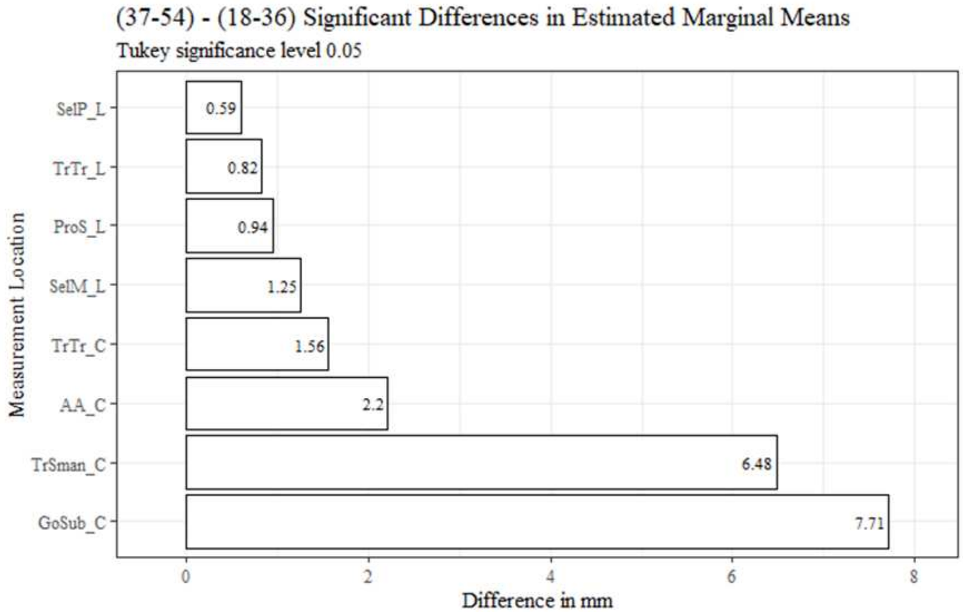

Figure 5.25: Significant differences in estimated marginal means between Mid-age (36-54) and Youngest (18-36) age groups. Difference estimates are presented in millimeters.

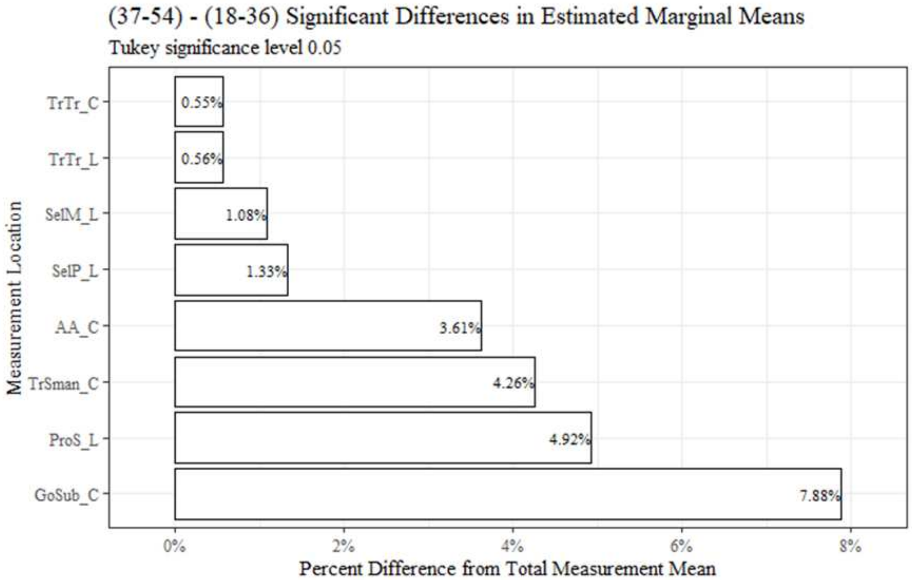

Figure 5.26: Significant differences in estimated marginal means between Mid-age (36-54) and Youngest (18-36) age groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

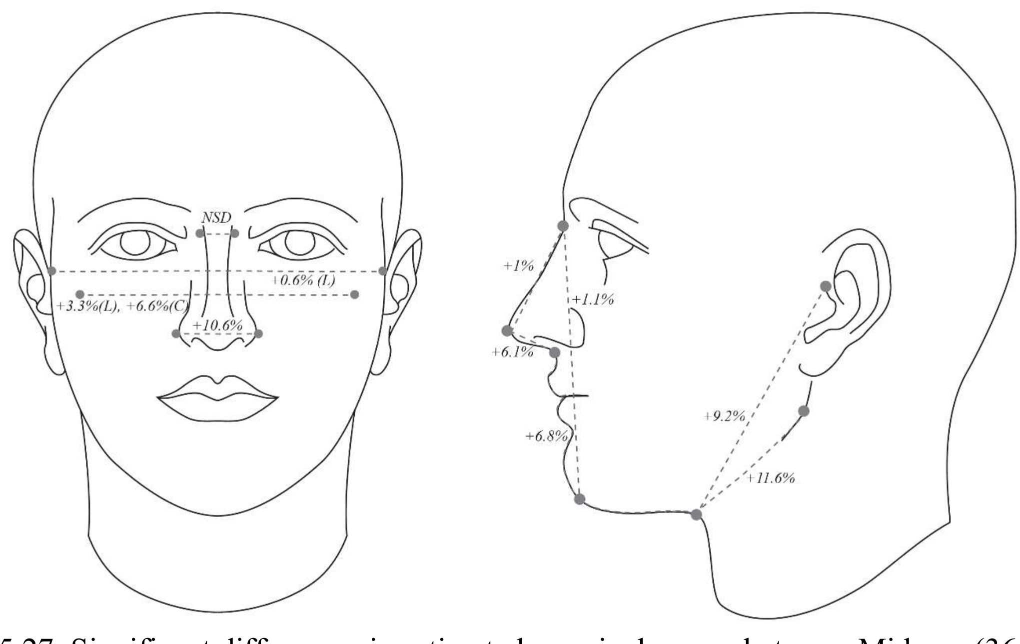

Figure 5.27: Significant differences in estimated marginal means between Mid-age (36-54) and Youngest (18-36) age groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, p<0.05).

Oldest - Youngest. Between the oldest and youngest age groups, Tukey's testing indicated that 7 out of 9 tested (based on ANOVA significance) and 12 total measurements had significant differences in estimated marginal means ($p<0.05$). In Chapter 4, PCA score plotting indicated that differences in variability were highest between the oldest and youngest age groups. The significant measurement differences between the oldest and youngest age groups are presented in the context of millimeters in Figure 5.28. Significant differences in estimated marginal means between the oldest and youngest age groups, as a percentage of the mean of the measurement for the entire sample population, are presented in Figures 5.29 and 5.30.

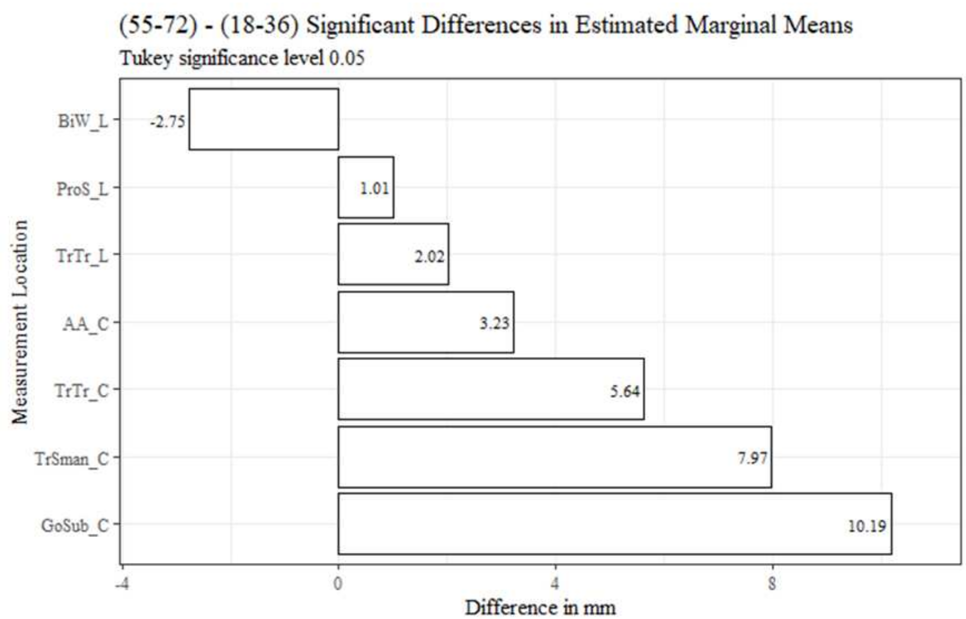

Figure 5.28: Significant differences in estimated marginal means between the oldest (55-72) and youngest (18-36) age groups. Difference estimates are presented in millimeters.

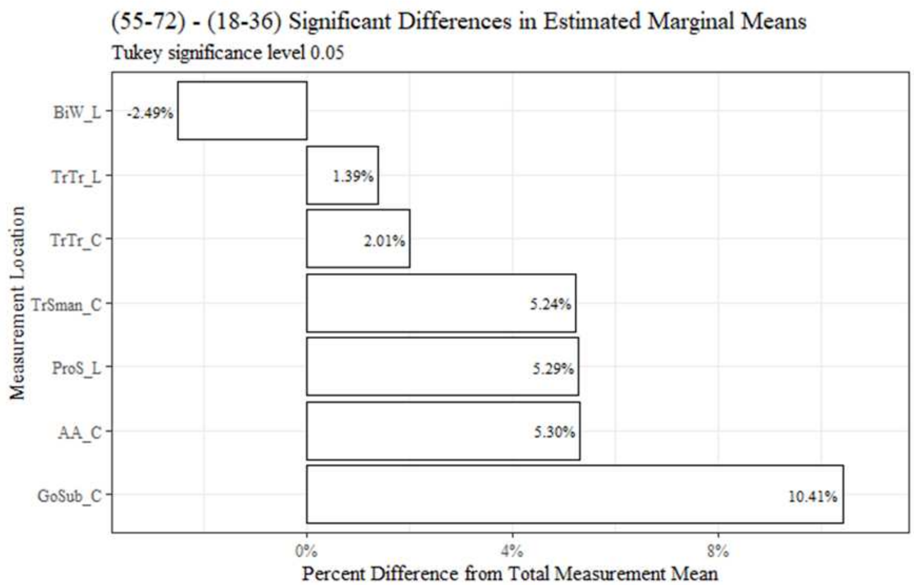

Figure 5.29: Significant differences in estimated marginal means between Oldest (55-72) and Youngest (18-36) age groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location.

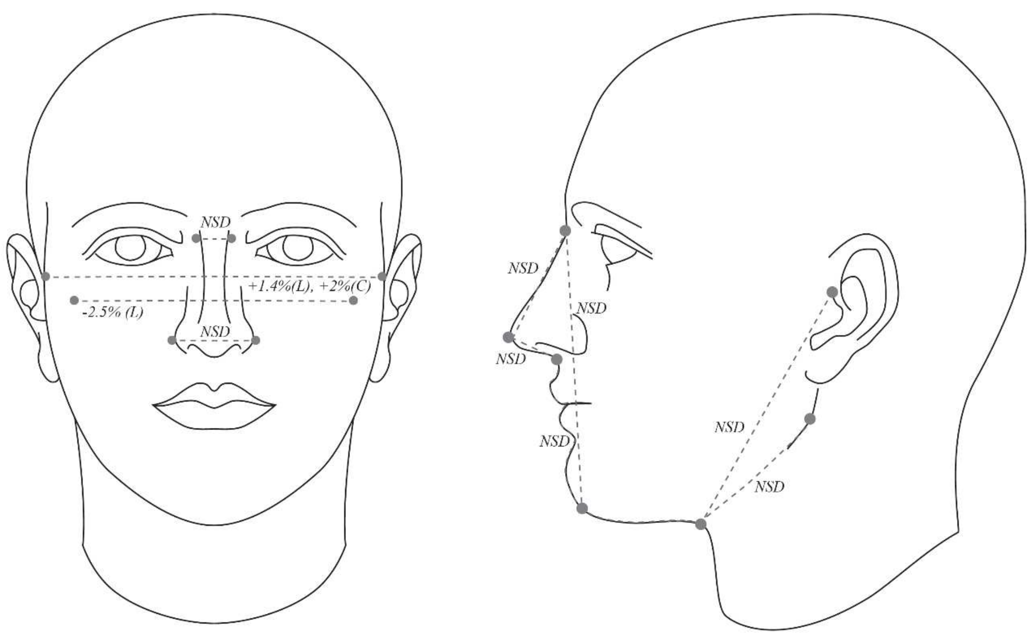

Figure 5.30: Significant differences in estimated marginal means between Oldest (55-72) and Youngest (18-36) age groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, $p<0.05$).

Oldest – Mid-age. Between the oldest and mid-age age groups, Tukey's testing indicated that 1 out of 9 tested (based on ANOVA significance) and 12 total measurements had significant differences in estimated marginal means ($p<0.05$). The significant measurement difference between groups was for the Tragion to Tragion Contour measurement, which was different by 4.08mm. The difference in Tragion to Tragion Contour estimated marginal means between the oldest and mid-age age groups, as a percentage of the mean of the measurement for the entire sample population, is presented in Figure 5.31.

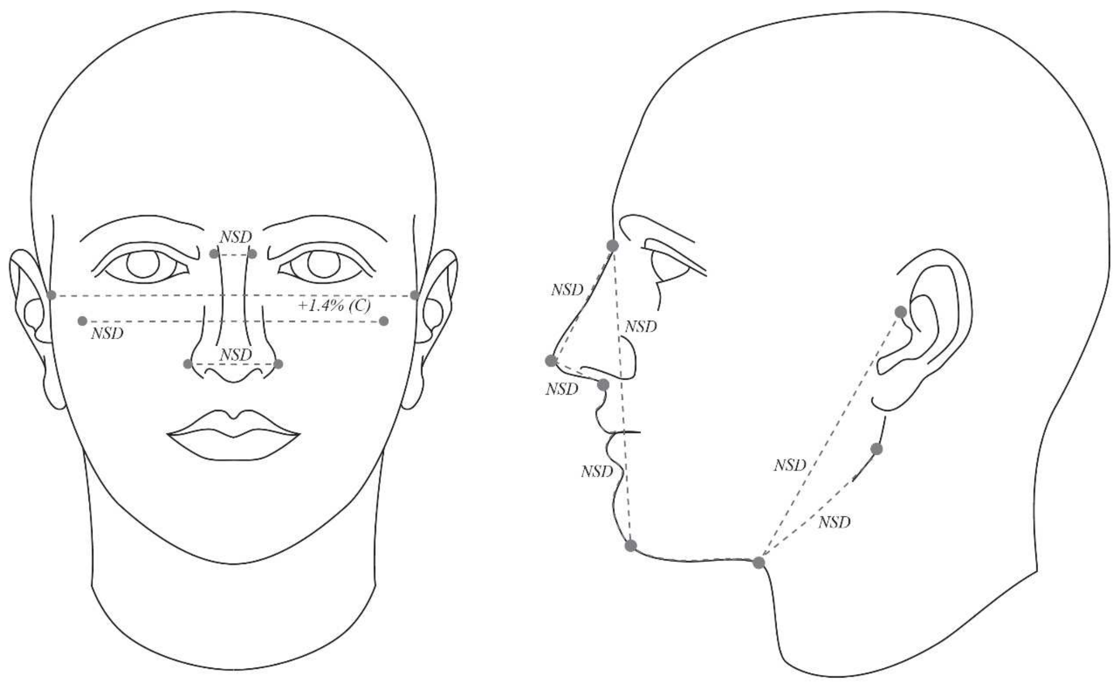

Figure 5.31: Significant differences in estimated marginal means between Oldest (18-36) and Mid-age (36-54) age groups. Difference estimates are presented in percentages. Percentage calculated as Tukey's test estimate divided by total measurement mean for the entire sample population per each measurement location. NSD indicates that location had no significant difference in measurements between the two groups (Tukey's test, $p<0.05$).

Conclusions

Within the gender and age demographic categories tested in these post hoc analyses, all possible group combinations had at least one significant difference for estimated marginal means in 12 tested 3D facial measurements based on Tukey's testing. Within the race/ethnicity demographic category tested in these post hoc analyses, 8 out of a possible 10 combinations had at least one significant difference for estimated marginal means in 12 tested 3D facial measurements based on Tukey's testing. Though positive and negative differences were demonstrated often, the majority of significant differences in estimated marginal means showed one demographic group as mostly ubiquitously larger than another. Similar to the practical implications of providing population summary statistics (Chapter 3), is expected that the results of these analyses will have practical implications for designers and manufacturers of facial products such as respirators.

CHAPTER 6: CONCLUSION

Limitations

3D facial measurement data, collected from 3D facial scans of 2022 participants, were used to complete three separate studies for this dissertation research. Each study had its own specific aim (*SA*) and research questions (*RQ*) but use the same dataset for analyses. In addition to the limitations discussed within each chapter, there are collective limitations of this dissertation research and the 3D facial measurement dataset that must be addressed. 3D facial scans were collected from a sample population recruited by Human Solutions within the Raleigh, North Carolina area. It is unclear as to if the demographic representation seen in this population was reflective of the demographic representation of the United States or the respirator-wearing population in the United States. It is also unclear whether the groups within each demographic category allowed for full reporting of the sample population's demographic status. This is particularly a limitation in the case of a participant identifying as more than one gender or race/ethnicity, given that participants were able to report their demographic identity as within only one group per category. Limitations related to the demographic representation of the sample population recruited and reported on by Human Solutions are outside the scope of this study but are nonetheless important to reflect upon.

The 3D facial measurement data were collected using 3D scan processing, Anthroscan software, and anthropometric knowledge that were completely new to the raters and thus required training. Though the assessment of intra- and inter-rater reliability helped to understand the quality of the data collected in this research, 43.01% of the 3D measurement data was collected by one rater (the author of the dissertation research). In this way, bias in data collection could have been introduced, and the 3D measurement data may have been affected. Figure 6.1

illustrates the amount of 3D measurement data collected by each coder, with rater code letters (Rater A, Rater B, Rater C, Rater D) corresponding to findings from Chapter 2.

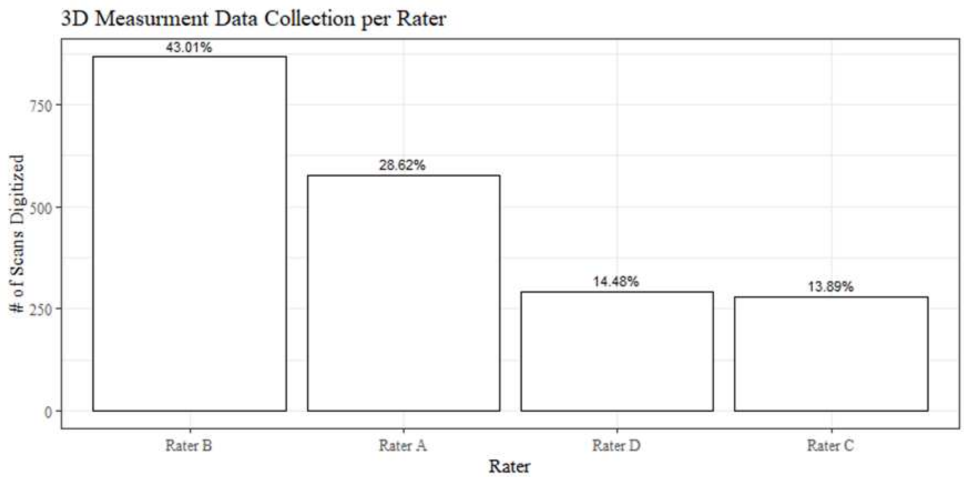

Figure 6.1: Amount of 3D data collected by each coder (y-axis: count, bar text label: percent). Rater code letters (Rater A, Rater B, Rater C, Rater D) correspond to reliability findings in Chapter 2.

Perhaps the most notable limitation present in this dissertation work is the lack of control measurements against which 3D measurements could be tested. If manual measurements of similar locations to 3D measurements were collected from the sample population of 2022 at the time of 3D scanning, each study would have a standard by which to assess the accuracy of the 3D measurements. Furthermore, a comparison between manual measurements as well as 3D measurements using both 3D and manual landmarking (as opposed to only 3D landmarking) would have allowed for the best possible understanding of data accuracy in this work. However, because Human Solutions did not collect manual measurement data or place manual landmarks prior to 3D scanning, this limitation was outside of the scope of this research. Again, this limitation is important to address in this dissertation, but could not have been remedied by the dissertation researcher.

Strengths

Like the limitations, the collective strengths of the three studies conducted in this dissertation work are found within the methods of data collection used. This dissertation research presented the largest collection of 3D facial measurement data analyzed in the relevant field of literature to date. Furthermore, the research utilized 3D landmarking to collect 3D facial measurements. This is a process that has not been used often in the literature, and by extension not used on a sample population of this size. As 3D scanning and 3D measurement data collection become more popular methods of anthropometric data collection due to strengths regarding less time required and reduced researcher-participant contact, so too may 3D landmarking become favored over manual landmarking of the subject prior to 3D scanning processes. Overall, 3D methods provide opportunities for many benefits over manual methods. This dissertation research utilizes 3D methods in a way that may inform future researchers, developers, designers, and many other types of professionals who seek to incorporate 3D methods in their anthropometric work.

Summary

This dissertation research assessed the largest sample of 3D facial measurement data seen in the literature to date, both a) in the context of 3D anthropometric data collection reliability and b) for practical purposes such as the design, sizing, and fit of wearable facial products such as respirators. A summary of findings for each specific aim as well as a summary of thematic findings from dissertation research overall are discussed in the subsequent sections.

Specific Aim Summaries

Specific Aim 1: Intra-Rater and Intra-Rater Reliability of 3D Facial Measurements.

Specific Aim 1 of this research sought to assess the intra- and inter-rater reliability of 3D facial measurements gathered by four novice anthropometric raters. In this dissertation, intra-rater reliability (intraRR) is the degree of agreement among collections of a 3D measurement performed on the same subject by a single rater, and inter-rater reliability (interRR) is the degree of agreement among all raters who collect the same 3D measurement on the same subjects. The results of the present study indicate 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. Future studies are needed to confirm that 3D landmarking and 3D measurement data collection are reliable processes for anthropometric data collection.

Specific Aim 1 (SA1): Assess the intra- and inter-rater reliability of 3D facial measurements gathered by four novice anthropometric raters.

> *SA1.RQ1*: What percentage of good to excellent (>0.75 ICC statistic) intra-rater reliability (on average across four raters) can be achieved by the final phase of data collection?
>
> > *ANSWER:* 90.74% good to excellent (>0.75 ICC) intra-rater reliability was achieved by Phase 3 on average across all four raters.
>
> *SA1.RQ2:* What percentage of good to excellent (>0.75 ICC statistic) inter-rater reliability can be achieved by the final phase of data collection?
>
> > *ANSWER:* 74.07% good to excellent (>0.75 ICC) inter-rater reliability was achieved by Phase 3.
>
> *SA1.RQ3:* In percentage terms and averaging across four raters, how much does intra-rater reliability improve over two phases of data collection?

ANSWER: ICC scores for intra-rater reliability improved by 58.34% (averaged across four raters) over the two phases in which intraRR was assessed (Phase 1 and Phase 3).

SA1.RQ4: In percentage terms, how much does inter-rater reliability improve on average over three phases of data collection?

ANSWER: ICC scores for inter-rater reliability improved by 42.59% averaging over the three phases of data collection.

Specific Aim 2: Facial Anthropometrics: A Comparison of Measurements from 3D and Manual Methods. Specific Aim 2 of this research sought to compare the 3D facial anthropometric summary statistics from the present study to relevant summary statistics from manual facial measurements found in the literature (Zhuang et al., 2007; Zhuang & Bradtmiller, 2005). The most notable result of SA2 was that of differences in summary statistics between 3D and manual measurement methods, but not always in the direction predicted by previous literature. Researchers and practitioners deciding between 3D and manual methods should consider what types of measurements will be best suited to inform their specific goals. Future studies are needed to compare 3D and manual measurements collected from the same population and identical measurement locations from each subject.

Specific Aim 2 (SA2): Compare measurements collected using 3D and manual methods.

SA2.RQ1: What are the summary statistics for the 3D measurement data collected?

ANSWER: The summary statistic data provided in Tables 3.3, 3.4, and 3.5 address and answer SA2.RQ1. Based on summary statistics, male faces were generally larger than female faces.

SA2.RQ2: How do the 3D measurement summary statistics compare to manual measurement summary statistics found in the literature?

ANSWER: In terms of precision and measurement means, 3D measurements may differ from manual measurements in ways that a) contradict previous literature and b) have implications for those collecting anthropometric data. Measurement method differences (between manual and 3D) generally had the expected effects on summary statistics.

Specific Aim 3: Demographic Differences in 3D Facial Anthropometrics Related to Respirator Fit. Specific Aim 3 of this research sought to assess the presence of differences in 3D facial anthropometric measurements related to respirator fit, based on demographic factors of gender, race/ethnicity, and age in a sample of 2022 3D scans. The most notable result of SA3 was that of significant differences in facial measurements between different groups within the demographic categories of gender (Male and Female/Other), race/ethnicity (White, Black, LatinX, Asian, and Other), and age (18-34, 35-54, and 55-72). Furthermore, the large majority of the 12 tested measurement locations were significantly different between different groups within each demographic category. Future studies are needed to confirm these findings with more demographically diverse populations.

Specific Aim 3 (SA3): Assess the presence of significant differences in 3D facial anthropometric measurements related to respirator fit.

SA3.RQ1: Are differences in 3D measurements present between gender groups?

ANSWER: Based on PCA, MANOVA, and ANOVA, differences in 3D measurements were present between gender groups.

SA3.RQ2: Are differences in 3D measurements present between race/ethnicity groups?

ANSWER: Based on PCA, MANOVA, and ANOVA, differences in 3D measurements were present between race/ethnicity groups.

SA3.RQ3: Are differences in 3D measurements present between age groups?

ANSWER: Based on PCA, MANOVA and ANOVA, differences in 3D measurements were present between age groups.

Thematic Findings about Nose Shape

The primary thematic findings throughout this dissertation work were regarding nose shape. These findings are new to the field of 3D facial anthropometrics for respirator fit. Nose shape has not previously been considered by other research as excessively important in understanding respirator-related or 3D measurement differences between demographic groups.

Four 3D measurements indicating two prominent aspects of nose shape were present throughout the notable results of this research on a large scale. These measurements included Pronasale to Subnasale Length and Alare to Alare Contour indicating nose tip shape, and Nasal Root Breadth and Sellion to Pronasale Length indicating nose bridge shape.

Nose Tip 3D Measurements. Pronasale to Subnasale Length (ProS_L) measured the straight-line distance between the pronasale and subnasale landmarks, and Alare to Alare

Contour (AA_C) measured the contour distance between right and left alare landmarks on the sides of the nostrils (definitions from Table 3.2). Together, ProS_L and AA_C provide anthropometric information regarding the shape of the tip of the nose. Because the tip of the nose tends to be the most protrusive area of the face on the frontal plane, understanding the anthropometrics of this area is essential for respirator fit. Poor understanding of this area could cause a respirator to not fit on the face properly (i.e., a large nose could cause poor respirator fit and therefore reduced respiratory protection). IntraRR and interRR tested as excellent for ProS_L and AA_C by the final phase of the reliability study described in Chapter 2. Using standard deviation F-testing (Table 3.8) as a measure of precision in Chapter 3, AA_C (3D measurement) was found to have significantly lower precision than Nose Breadth (manual measurement from Zhuang & Bradtmiller, 2005). ProS_L was the only measurement for which standard deviations had equal variances between manual and 3D methods. In Chapter 4, ProS_L and AA_C negatively affected variance (factor loads -0.4223793 and -0.377925, respectively) in PC2 (19.51% variance described). Previous relevant literature has not found nose-related measurements to inform PCA in a notable way. Also in Chapter 4, ProS_L and AA_C were found to be significantly different for all demographic categories in ANOVA. Overall, 3D nose tip measurements such as ProS_L and AA_C may be important measurements to consider when collecting facial anthropometric data for products (such as respirators) that different demographic groups may wear.

Nose Bridge 3D Measurements. Nasal Root Breadth (NRB_L) measured the straight-line distance between nasal root landmarks (left and right) (definition from Table 3.2), and SelP_L measured the straight-line distance between the sellion and pronasale landmarks. Together, NRB_L and SelP_L provide anthropometric information regarding the upper areas of the nose such as the nose bridge. Because the nose bridge is where respirators tend to meet the face, understanding the anthropometrics of this area is essential for respirator fit and seal. Poor understanding of the anthropometry of this area could cause a respirator to not fit or seal on the face properly, especially for the many types of respirators that use molded plastic to provide shape in this area. IntraRR tested as excellent for NRB_L and SelP_L by the final phase of the reliability study described in Chapter 2. InterRR tested as excellent for SelP_L, but poor for NRB_L by the final phase of the reliability study. Using standard deviation F-testing (Table 3.8) as a measure of precision in Chapter 3, NRB_L (3D measurement) was found to have significantly lower precision than Nasal Root Breadth (manual measurement from Zhuang & Bradtmiller, 2005). SelP_L did not have a comparable manual measurement to test precision against. In Chapter 4, NRB_L did not largely affect any PC, but SelP_L largely negatively affected variance (factor load -0.63600471) in PC3 (11.04% variance described). Previous relevant literature has not found nose-related measurements to inform PCA in a notable way. Also in Chapter 4, SelP_was found to be significantly different for all demographic categories in ANOVA and NRB_L was found to be significantly different for gender and race/ethnicity categories (but not age). Overall, 3D nose bridge measurements such as NRB_L and SelP_L may be important measurements to consider when collecting facial anthropometric data for products (such as respirators) that different demographic groups may wear. Given the more impactful thematic findings regarding nose-tip (AA_C and ProS_L) 3D measurements, more research regarding the effects of nose bridge 3D measurements on respirator fit is needed.

Though these nose-shape findings have implications for the field of 3D facial anthropometrics for respirator fit, it is possible that thematic findings around nose shape were due to these measurements being examined for all specific aims. Of the 27 measurements collected, only five measurements a) were part of the 9 3D measurements comparable to manual methods for SA2.RQ2 and b) were chosen as part of the 12 3D measurements used to evaluate differences in respirator-related measurements in SA3. Three out of these five measurements were related to nose shape (AA_C, NRB_L, and ProS_L). Therefore, these thematic findings should be interpreted as having potential bias regarding variable sampling. Future research is needed to evaluate if nose shape measurements are truly of high importance in this line of research regarding 3D anthropometrics, diverse demographics, and respirator fit.

Implications

This dissertation research presents theoretical and practical implications of 3D measurement data collection as well as the design of facial wearable products. It is expected that a wide range of researchers and professionals may find this dissertation work helpful in advancing the field of literature and providing tangible measurement findings surrounding 3D facial anthropometry and design, sizing, and fit of wearable facial products like respirators.

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 Endocrine Disorders*, *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. *The Angle Orthodontist*, *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

Ball, R., Shu, C., Xi, P., Rioux, M., Luximon, Y., & Molenbroek, J. (2010). A comparison between Chinese and Caucasian head shapes. *Applied Ergonomics*, *41*(6), 832–839. https://doi.org/10.1016/j.apergo.2010.02.002

ben Azouz, Z., Rioux, M., Shu, C., & Lepage, R. (2006). Characterizing human shape variation using 3D anthropometric data. *The Visual Computer 2006 22:5*, *22*(5), 302–314. https://doi.org/10.1007/S00371-006-0006-6

Berkelaar, M., & others. (2022). *lpSolve: Interface to Lp_solve v. 5.5 to Solve Linear/Integer Programs*. https://github.com/gaborcsardi/lpSolve

Brazile, W. J., Buchan, R. M., Sandfort, D. R., Melvin, W., Johnson, J. A., & Charney, M. (1998). Respirator Fit and Facial Dimensions of Two Minority Groups. *Applied Occupational and Environmental Hygiene*, *13*(4), 233–237. https://doi.org/10.1080/1047322X.1998.10390073

Centers for Disease Control and Prevention, & The National Institute of Occupational Safety and Health. (n.d.). *DHHS (NIOSH) Publication No. 2018-130, Filtering out Confusion: Frequently Asked Questions about Respiratory Protection, User Seal Check*. https://doi.org/10.26616/NIOSHPUB2018130

Chang, W. (2022). *extrafont: Tools for Using Fonts*. https://github.com/wch/extrafont

Clauser, C., Tebbetts, I., Bredtmiller, B., McConville, J., & Gordon, C. C. (1988). *Measurer's handbook: U.S. Army anthropometric survey 1987-1988*.

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. *International Journal of Obesity*, *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. *Journal of Craniofacial Surgery*, *29*(5), 1245–1251. https://doi.org/10.1097/SCS.0000000000004580

Ebrahim, M. (2014). *3D Laser Scanners: History, Applications, and Future*. https://doi.org/10.13140/2.1.3331.3284

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. In *BMC Medical Imaging* (Vol. 14). http://www.biomedcentral.com/1471-2342/14/35

Fox, J., Weisberg, S., & Price, B. (2022). *car: Companion to Applied Regression*. https://CRAN.R-project.org/package=car

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. *International Journal of Oral and Maxillofacial Surgery*, *49*(8), 1092–1106. https://doi.org/10.1016/j.ijom.2019.10.019

Gohel, D. (2022). *flextable: Functions for Tabular Reporting*. https://CRAN.R-project.org/package=flextable

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.*

Gross, S. F., & Horstman, S. W. (1990). Half-Mask Respirator Selection for a Mixed Worker Group. *Applied Occupational and Environmental Hygiene*, *5*(4), 229–235. https://doi.org/10.1080/1047322X.1990.10389628

Hack, A., Hyatt, E. C., Held, B. J., Moore, T. 0, Richards, C. P., & Mcconville, J. T. (1973). *Selection of Respirator Test Panels Representative of U.S. Adult Facial Sizes (LA-5488)*.

Han, H., Nam, Y., & Choi, K. (2010). Comparative analysis of 3D body scan measurements and manual measurements of size Korea adult females. *International Journal of Industrial Ergonomics*, *40*(5), 530–540. https://doi.org/10.1016/J.ERGON.2010.06.002

Heymsfield, S. B., Bourgeois, B., Ng, B. K., Sommer, M. J., Li, X., & Shepherd, J. A. (2018). Digital anthropometry: a critical review. *European Journal of Clinical Nutrition*, *72*, 680–687. https://doi.org/10.1038/s41430-018-0145-7

Holmes, S., & Huber, W. (2022). *Modern Statistics for Modern Biology*. Creative Commons.

Horikoshi, M., & Tang, Y. (2018). *ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://rpkgs.datanovia.com/rstatix/

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. *Orthodontics & Craniofacial Research*, *21*(3), 132–139. https://doi.org/10.1111/ocr.12228

Kim, H., Han, D.-H., Roh, Y.-M., Kim, K., & Park, Y.-G. (2003). Facial anthropometric dimensions of Koreans and their associations with fit of quarter-mask respirators. In *Industrial Health* (Vol. 41).

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

Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A., & Scholz, M. (2016). Reliability of 3D laser-based anthropometry and comparison with classical anthropometry. *Scientific Reports*, *6*. https://doi.org/10.1038/srep26672

Lee, W., Park, J., Jeong, J., Jeon, E., Kim, H. E., Park, S., & You, H. (2012). Analysis of the facial anthropometric data of Korean pilots for oxygen mask design. *Proceedings of the Human Factors and Ergonomics Society*, 1927–1931. https://doi.org/10.1177/1071181312561282

Leigh, J. D. (1975). *RFP-2358 Quantitative respirator man-testing and anthropometric survey*.

Lenth, R. v. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://github.com/rvlenth/emmeans

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

Luximon, Y., Ball, R., & Justice, L. (2010). *The Chinese face: A 3D anthropometric analysis*.

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

Mei, Z., & Grummer-Strawn, L. M. (2007). Standard deviation of anthropometric Z-scores as a data quality assessment tool using the 2006 WHO growth standards: a cross country analysis. *Bulletin of the World Health Organization*, *85*(6), 441–448. https://doi.org/10.2471/BLT.06.034421

Menditto, A., Patriarca, M., & Magnusson, B. (2007). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*, *12*(1), 45–47. https://doi.org/10.1007/s00769-006-0191-z

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

Occupational Safety and Health Administration. (2009). *Assigned Protection Factors for the Revised Respiratory Protection Standard: 3352-02*.

Oestenstad, R. K., & Perkins, L. L. (1992). An assessment of critical anthropometric dimensions for predicting the fit of a half-mask respirator. *American Industrial Hygiene Association Journal*, *53*(10), 639–644. https://doi.org/10.1080/15298669291360283

Ooms, J. (2021). *writexl: Export Data Frames to Excel xlsx Format*. https://CRAN.R-project.org/package=writexl

R Core Team. (2022a). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

R Core Team. (2022b). *R: A language and environment for statistical computing*.

R Core Team. (2022c). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

R Core Team. (2022d). *R: The R Project for Statistical Computing*. https://www.r-project.org/

R in Action (Third Edition). (2021). Manning Publications.

Revopoint 3D Technologies Inc. (n.d.). *How does Light Influence 3D Scanning?* . Retrieved April 11, 2022, from https://www.revopoint3d.com/how-does-light-influence-3d-scanning/

Robinson, D., Hayes, A., & Couch, S. (2022). *broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom

Rodríguez, A. A., Escanilla, D. E., Caroca, L. A., Albornoz, C. E., Marshall, P. A., Molenbroek, J. F. M., & Castellucci, H. I. (2020). Level of match between facial dimensions of Chilean workers and respirator fit test panels proposed by LANL and NIOSH. *International Journal of Industrial Ergonomics*, *80*, 103015. https://doi.org/10.1016/J.ERGON.2020.103015

Sandler, A. (2021). Against Standard Deviation as a Quality Control Maxim in Anthropometry. *Econ Journal Watch*, *18*(1), 95–126. https://journaltalk.net/articles/6030/

Spies, A., Wilson, K. S., & Ferrie, R. (2011). Respirator fit of a medium mask on a group of South Africans: A cross-sectional study. *Environmental Health: A Global Access Science Source*, *10*(1), 1–7. https://doi.org/10.1186/1476-069X-10-17/TABLES/5

Taylor, P. J. (2009). *1 An Introduction to Intraclass Correlation that Resolves Some Common Confusions*.

The National Institute for Occupational Health and Safety. (2022). *Respirators*. Centers for Disease Control and Prevention. https://www.cdc.gov/niosh/topics/respirators/default.html

The National Institute for Occupational Safety and Health, & Centers for Disease Control and Prevention. (n.d.). *Anthropometry*. Retrieved April 3, 2022, from https://www.cdc.gov/niosh/topics/anthropometry/default.html

Wickham, H. (2021). *forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats

Wickham, H. (2022). *tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse

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., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2022). *readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl

Wickham, H., & Seidel, D. (2022). *scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales

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. *The Cleft Palate-Craniofacial Journal*, *45*(3), 232–239. https://doi.org/10.1597/06-175

Zhang, X., Jia, N., & Wang, Z. (2020). *The relationship between the filtering facepiece respirator fit and the facial anthropometric dimensions among Chinese people*.

Zhuang, Z., & Bradtmiller, B. (2005). Head-and-Face Anthropometric Survey of U.S. Respirator Users. *Journal of Occupational and Environmental Hygiene*, *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. *Journal of Occupational and Environmental Hygiene*, *4*(9), 647–659. https://doi.org/10.1080/15459620701497538

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

APPENDIX

Phase 1 IntraRR

3D Measurement	Rater A			Rater B			Rater C			Rater D		
	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI
AA_C	0.9440	0.835	0.985	0.8170	0.466	0.950	0.4020	-0.656	0.835	0.8160	0.485	0.949
BGl_C	0.6750	-0.012	0.914	0.7700	0.331	0.938	-0.201	-4.682	0.787	0.8860	0.677	0.969
BiW_C	0.7450	0.275	0.930	0.2670	-0.612	0.774	0.7340	0.246	0.927	0.0629	-0.370	0.592
BiW_L	0.8590	0.546	0.963	0.6870	0.153	0.912	0.8660	0.623	0.963	0.5940	-0.097	0.886
ChCh_C	0.9680	0.907	0.992	0.9160	0.758	0.977	0.9220	0.779	0.979	0.6870	0.105	0.921
GoSub_C	0.4180	-0.596	0.850	0.8700	0.609	0.968	0.6650	-0.071	0.926	0.9120	0.723	0.978
NRB_L	0.8910	0.693	0.970	0.8860	0.655	0.969	0.8650	0.617	0.963	0.8240	0.409	0.954
ProA_L	0.8240	0.486	0.952	0.8430	0.558	0.957	0.7180	0.229	0.922	0.8720	0.617	0.966
ProA_C	0.7150	0.150	0.924	0.7510	0.302	0.932	0.6120	-0.048	0.892	0.8270	0.479	0.954
ProS_C	0.9070	0.730	0.975	0.5500	-0.322	0.879	0.7520	0.244	0.934	0.8930	0.697	0.971
ProS_L	0.9380	0.817	0.983	0.6070	-0.085	0.892	0.8140	0.435	0.950	0.8610	0.601	0.962
SelP_C	0.8990	0.716	0.972	0.9320	0.807	0.982	0.8100	0.432	0.949	0.8900	0.679	0.970
SelP_L	0.9060	0.733	0.975	0.9450	0.843	0.985	0.8200	0.460	0.952	0.8880	0.668	0.970
SelDH_C	0.6980	0.155	0.916	0.6920	0.134	0.914	0.5420	-0.463	0.880	-0.277	-1.040	0.727
SelM_L	0.9580	0.872	0.989	0.9340	0.803	0.984	0.8790	0.576	0.977	0.6970	-0.031	0.941
SnasM_L	0.9330	0.798	0.983	0.9170	0.713	0.980	0.8700	0.533	0.975	0.5000	-0.136	0.884
SmanM_C	0.6610	0.041	0.914	0.8580	0.579	0.965	0.3690	-2.046	0.890	0.6410	-0.092	0.927
SmanM_L	0.6080	-0.086	0.899	0.8100	0.438	0.953	0.4640	-1.507	0.906	0.6910	-0.022	0.939
SnasM_C	0.9470	0.838	0.987	0.9210	0.755	0.981	0.8890	0.591	0.979	0.6260	-0.090	0.923
TrHO_C	0.9260	0.778	0.982	0.7810	0.336	0.945	0.9280	0.717	0.989	0.8940	0.593	0.984
TrEJ_C	0.9180	0.665	0.981	0.8760	0.601	0.970	0.9390	0.810	0.985	0.7360	0.222	0.934
TrGo_C	0.8690	0.547	0.972	0.7620	0.259	0.947	0.8220	0.403	0.961	0.7690	0.248	0.949
TrSel_C	0.9870	0.960	0.997	0.9670	0.869	0.992	0.9370	0.810	0.984	0.9410	0.822	0.985
TrSman_C	0.7110	0.100	0.936	0.9650	0.888	0.992	0.7650	0.162	0.956	0.9110	0.701	0.981
TrSnas_C	0.9950	0.983	0.999	0.0408	-1.913	0.762	0.9660	0.885	0.992	0.9580	0.869	0.990
TrTr_C	0.9880	0.963	0.997	0.9880	0.944	0.997	0.9710	0.910	0.993	0.9810	0.939	0.995
TrTr_L	0.9780	0.935	0.995	0.9510	0.829	0.988	0.3900	-0.402	0.828	0.7440	0.260	0.935

Phase 3 IntraRR

3D Measurement	Rater A			Rater B			Rater C			Rater D		
	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI
AA_C	0.987	0.963	0.996	0.974	0.927	0.993	0.974	0.925	0.993	0.969	0.885	0.992
BGl_C	0.960	0.871	0.991	0.974	0.927	0.993	0.856	0.588	0.961	0.947	0.832	0.987
BiW_C	0.917	0.765	0.977	0.944	0.832	0.985	0.849	0.523	0.960	0.962	0.861	0.990
BiW_L	0.924	0.784	0.979	0.915	0.740	0.977	0.692	0.115	0.915	0.926	0.791	0.980
ChCh_C	0.984	0.949	0.997	0.968	0.909	0.991	0.924	0.783	0.979	0.949	0.855	0.986
GoSub_C	0.937	0.811	0.984	0.950	0.850	0.988	0.919	0.771	0.978	0.944	0.817	0.988
NRB_L	0.830	0.525	0.953	0.757	0.291	0.935	0.384	-0.654	0.828	0.915	0.754	0.977
ProA_L	0.986	0.957	0.996	0.974	0.924	0.993	0.961	0.889	0.989	0.982	0.943	0.995
ProA_C	0.975	0.927	0.993	0.960	0.885	0.989	0.945	0.839	0.985	0.964	0.887	0.990
ProS_C	0.969	0.909	0.992	0.941	0.828	0.984	0.686	0.047	0.916	0.967	0.908	0.991
ProS_L	0.961	0.889	0.989	0.931	0.799	0.981	0.955	0.868	0.988	0.975	0.927	0.993
SelP_C	0.991	0.973	0.997	0.977	0.936	0.994	0.962	0.891	0.990	0.992	0.976	0.998
SelP_L	0.991	0.975	0.998	0.978	0.936	0.994	0.970	0.914	0.992	0.992	0.975	0.998
SelDH_C	0.926	0.787	0.980	0.920	0.719	0.979	0.503	-0.278	0.859	0.943	0.836	0.985
SelM_L	0.995	0.979	0.999	0.969	0.899	0.993	0.977	0.924	0.995	0.997	0.990	0.999
SnasM_L	0.980	0.904	0.997	0.921	0.731	0.983	0.963	0.876	0.992	0.989	0.962	0.998
SmanM_C	0.457	-1.100	0.900	0.740	0.093	0.944	0.885	0.632	0.975	0.970	0.891	0.994
SmanM_L	0.522	-0.812	0.912	0.692	-0.089	0.934	0.879	0.613	0.973	0.964	0.875	0.993
SnasM_C	0.954	0.807	0.992	0.849	0.474	0.967	0.899	0.669	0.978	0.986	0.951	0.997
TrHO_C	0.981	0.934	0.997	0.970	0.911	0.993	0.857	0.567	0.965	0.964	0.799	0.998
TrEJ_C	0.979	0.940	0.994	0.935	0.814	0.982	0.946	0.847	0.985	0.823	0.475	0.956
TrGo_C	0.971	0.913	0.993	0.969	0.895	0.993	0.871	0.624	0.965	0.777	0.220	0.952
TrSel_C	0.993	0.979	0.998	0.990	0.972	0.997	0.993	0.980	0.998	0.732	0.204	0.933
TrSman_C	0.993	0.978	0.998	0.997	0.989	0.999	0.965	0.900	0.991	0.907	0.700	0.980
TrSnas_C	0.971	0.917	0.992	0.988	0.964	0.997	0.983	0.950	0.995	0.740	0.235	0.935
TrTr_C	0.997	0.990	0.999	0.989	0.965	0.997	0.993	0.979	0.998	0.932	0.781	0.985
TrTr_L	0.998	0.995	1.000	0.992	0.976	0.998	0.995	0.986	0.999	0.995	0.985	0.999

InterRR, all phases

3D Measurement	Phase 1			Phase 2			Phase 3		
	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI	ICC	Lower 95% CI	Upper 95% CI
AA_C	0.874	0.673	0.964	0.938	0.826	0.983	0.9520	0.838	0.987
BGl_C	0.827	0.476	0.966	0.923	0.607	0.995	0.0432	-0.002	0.224
BiW_C	-0.201	-0.718	0.575	0.641	0.069	0.898	0.6890	0.116	0.915
BiW_L	0.413	-0.158	0.803	0.629	0.069	0.895	0.5380	0.016	0.856
ChCh_C	0.886	0.598	0.973	0.885	0.607	0.970	0.9660	0.871	0.993
GoSub_C	0.602	0.005	0.902	0.921	0.755	0.980	0.9230	0.779	0.983
NRB_L	0.757	0.253	0.935	0.365	-0.051	0.759	0.2450	-0.130	0.678
ProA_L	0.825	0.543	0.951	0.917	0.741	0.978	0.9500	0.833	0.987
ProA_C	0.750	0.370	0.928	0.879	0.573	0.969	0.9310	0.778	0.982
ProS_C	0.715	0.279	0.919	0.674	0.152	0.915	0.8500	0.616	0.958
ProS_L	0.775	0.435	0.936	0.811	0.491	0.951	0.9380	0.825	0.983
SelP_C	0.860	0.641	0.961	0.959	0.851	0.990	0.9670	0.892	0.991
SelP_L	0.865	0.654	0.962	0.960	0.861	0.990	0.9680	0.892	0.992
SelDH_C	0.269	-0.244	0.726	0.828	0.493	0.953	0.8250	0.497	0.952
SelM_L	0.864	0.434	0.975	0.666	0.099	0.929	0.8730	0.373	0.978
SnasM_L	0.841	0.390	0.970	0.493	0.007	0.869	0.8090	0.239	0.970
SmanM_C	0.148	-2.588	0.849	0.619	0.009	0.917	0.3290	-0.383	0.836
SmanM_L	0.285	-1.895	0.872	0.656	0.063	0.927	0.2990	-0.443	0.830
SnasM_C	0.860	0.443	0.974	0.437	0.022	0.840	0.5210	0.009	0.881
TrHO_C	0.906	0.644	0.989	0.400	-11.069	0.988	0.8710	0.452	0.991
TrEJ_C	0.753	0.338	0.936	0.908	0.652	0.980	0.7760	0.346	0.943
TrGo_C	0.783	0.317	0.951	0.850	0.531	0.966	0.8800	0.631	0.977
TrSel_C	0.954	0.931	0.989	0.936	0.802	0.986	0.8660	0.641	0.966
TrSman_C	0.568	-0.097	0.907	0.984	0.936	0.997	0.9490	0.849	0.989
TrSnas_C	0.973	0.921	0.993	0.958	0.858	0.992	0.9910	0.972	0.998
TrTr_C	0.976	0.900	0.994	0.982	0.938	0.996	0.9580	0.861	0.991
TrTr_L	0.659	0.120	0.909	0.997	0.991	0.999	0.9890	0.963	0.997