

Man vs. Machine: Measuring People for the Apparel Industry

Warren P. WRIGHT*
Size Stream LLC, Cary (NC), USA

<https://doi.org/10.15221/19.033>

Abstract

The Institute of Electrical and Electronics Engineers *3D Body Processing Industry Connections* group is conducting a Comparative Analysis of Measurement Methods of 3D Body Scans (project details are given in Working Group Progress for IEEE P3141 - Standard for 3D Body Processing, 2018-2019). This article reports on the results of the first phase of the project. Over 60 subjects were scanned and manually measured at the Portland, OR site during November 2018. Here we report on the measurements acquired manually and those acquired by a Size Stream SS20 3D body scanner. The project goal in focus here is *to understand the reliability and compatibility of measurements obtained through traditional 1D and advanced 3D methods*. Scanner reliability was shown to have more than double the precision of manual measurements (via Coefficient of Variation analysis). Furthermore, we find that the variability of the two measuring techniques individually is greater than any bias between them (difference in the mean). After accounting for this bias, manual and scanner measurement techniques are compatible and can be used interchangeably.

Keywords: 3d body scanning, Measurement analysis.

1. Introduction

Scanner and manual measurement data from a survey of 61 people has been analyzed in order to compare scanner and manual measurements. These two measurement techniques both had very few outliers that needed to be removed according to ISO 15535. These two techniques were compared in terms of precision/repeatability and accuracy/compatibility. Precision/repeatability was analyzed in terms of Mean Absolute Deviation (MAD) and used the ANSUR[1] tolerances for comparison. Accuracy/compatibility was analyzed in terms of ISO 20685 and standard ANOVA techniques.

A centralized Comparative Analysis of both the 2018 and 2019 scan events is planned to be compiled and distributed by early 2020.

2. Data Processing

2.1. Preparation and Cleaning

The manual measurements were received in an Excel spreadsheet. Some minor reshaping of the data was performed to make it amenable for import into python for data analysis. Two values were amended by hand: for subject 908035, a measurement of 4525 was changed to 45.25, and there was one other measurement where the decimal point was mistyped as a space. The scanner mesh files were batch processed into extracted measurements using the same settings as those used at the scan event (thus would give the same results shown at scan time). This was needed because the measurement files saved at scan time did not include all the measurements necessary for comparison to manual measurements.

The ISO 15535 standard was used to clean outliers from both the manual and scanner measurement outputs. For each measurement technique separately, the following was done. For each measurement (chest, waist, ...), all measurements were pooled (size of pool is number of subjects, times the number of measurements per subject). Any single measurement that fell outside of 3 standard deviations (σ) was flagged for investigation. Each flagged measurement was manually analyzed via correlation scatter plots to determine if these represented good or bad data.

* wwright@sizestream.com

The following manual measurement data were removed as per this outlier removal standard:

- subject = 908040, Upper arm girth. Manual measurer 4 measured 26 and 91. The 91 is suspect, and correlation plots (vs any other measurement) show it as an outlier.
- subject = 908014, Back neck point to wrist. Manual measurer 4 measured 79.5 and 57.5. The 57.5 is suspect (maybe 77.5?), and correlation plots (vs any other measurement) show it as an outlier.
- subject = 908020, Across back shoulder width. Manual measurer 1 measured 27 and 43.5. The 27 is suspect (maybe 47?), and correlation plots (vs any other measurement) show it as an outlier.
- subject = 908059, Inside Leg Length. Manual measurer 4 measured 62.5 and 82.5. This is within 3σ , but correlation plots do show it as an outlier.

The following scanner measurement data were removed as per this outlier removal standard:

- subject = 908026, enum=241, Right sleeve length. Correlation plots (vs any other measurement) show it as an outlier.
- subject = 908026, enum=90, Right bicep. This was within 3σ but correlation plots (and further analysis) reveal this to be an outlier.
- subject = 908086, enum=241, Right sleeve length. Correlation plots (vs any other measurement) show it as an outlier.
- subject = 908039, enum=241, Right sleeve length. This was within 3σ but correlation plots (and further analysis) reveal this to be an outlier.
- subject = 908039, enum=90, Right bicep. This was within 3σ but correlation plots (and further analysis) reveal this to be an outlier.

Summary: 4 data points were removed from the manual measurement set (out of >5000 points) and 5 data points were removed from the scanner measurement set (out of >1300 points).

2.2. Comparing measurement techniques

There are multiple ways to display data and throughout this paper, a few different options will be presented. The hope is that, amongst the options, at least one will be useful to professionals from a variety of different backgrounds. One data representation that has found recent appeal is the "charts on fire" representation (Figs. 1 and 2). These Fire Charts distill a common dilemma. The dilemma is that any single quality measurer, using a tape measure, can achieve a precision that is only a little worse than the scanner (comparing the pass rates between Fig. 1 and Fig. 2(top)). This is the practical evidence that is often put forward anecdotally (since getting the statistical comparison involves a large effort). However, the problem is that even high-quality measurers differ one to another, and thus the true situation when multiple manual measurers is what is seen in Fig. 2(bottom). Now it is clear that the scanner's precision is much better than a group of manual measurers. Another use of these Fire Charts is to show that any measuring system will occasionally have errors that on the surface seem extremely large. For manual measurement, this could be a simple human error. For scanner measurement, these large errors are often attributable to some defect in the scan, and there is active development to try and minimize these occurrences.

One question that needs care is: which scanner measurement should we compare to each manual measurement? For some, it is easy, based on the definitions. For others, say the waist, there are multiple scanner measurements that could conceivably be compared to the manual measurements. For these cases we will select the scanner measurement that best matches the manual measurements. However, for the manual measurements on arms and legs, we include both left and right arm and leg measurements where the manual measurement only has one number. Therefore, the results for scanner measurement below include more measurements than the manual measurements.

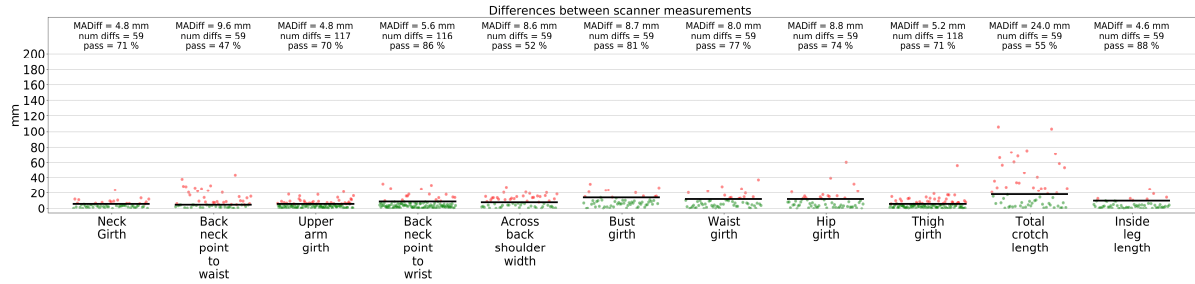


Fig. 1. Differences between scanner measurements. For two scans per person the number of differences is roughly equal to the number of scan subjects. For those measurement types that have a left and right (like Thigh Girth), both are displayed and hence there is twice the data. Red means the difference was greater than the ANSUR[1] tolerance. Green means the difference was less than the ANSUR[1] tolerance. The black line is the tolerance for each type of measurement. The pass percentage is the percentage of green dots to the number of differences (num diffs). The MADiff is given and is equivalent to the MADiff displayed in Fig. 6 (the larger bar).

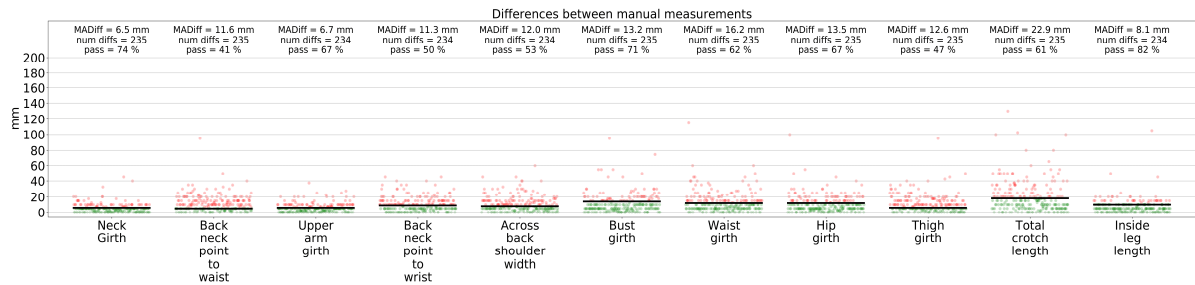


Fig. 2. Differences between manual measurements. The structure is the same as Fig. 1. The top plot shows intra-measurer precision (differences between measurements by the same measurer). For 4 measurers each measuring twice we get 4 differences per scan subject (~240 differences). The bottom plot shows all differences (intra and inter measurer). For 8 measurements per scan subject we get ~1680 differences ((8 choose 2) * 60). That the plots contain a little less data is indicative of missing data or the effect of outlier removal.

3. Precision or Repeatability or Consistency

The goal here is to see how precise each measurement technique is. That is, when measuring the same thing twice, how close are the two numbers. Put another way, we are looking for the uncertainty attached to each measurement. This has nothing to do with finding the 'correct' answer or if the measurement definitions are the same.

3.1. Analyzing the Variances Separately

Figure 3 shows the boxplots of the difference-from-mean distributions for manual measurements per measurement type (chest, waist, ...). This is calculated by taking, for each person and measurement, all the repeated measurement values and subtracting the mean of those measurement values. This results in an array of error terms for each person. All those arrays are gathered into one array of deviations and this forms the difference-from-mean distribution. There is a lot one can glean from this plot. Primarily this gives an intuitive feel for the spread of measurements for each type of measurement. We see that roughly 95% of manual measurements were within about an inch of each other. When comparing the measurements to the Orange ANSUR[1] tolerance, the variability seen in Fig. 3 seems high. We should remember that the ANSUR[1] tolerances were obtained in a context of specialist anthropometrists using specialized equipment, undergoing standardized training, and with the benefit

of expert land-marking prior to measurements. However, manual measurers represented by Fig. 3 operated under a somewhat typical retail environment. This meant no land-marking, different training, and only having a tape measure with which to obtain the measurements.

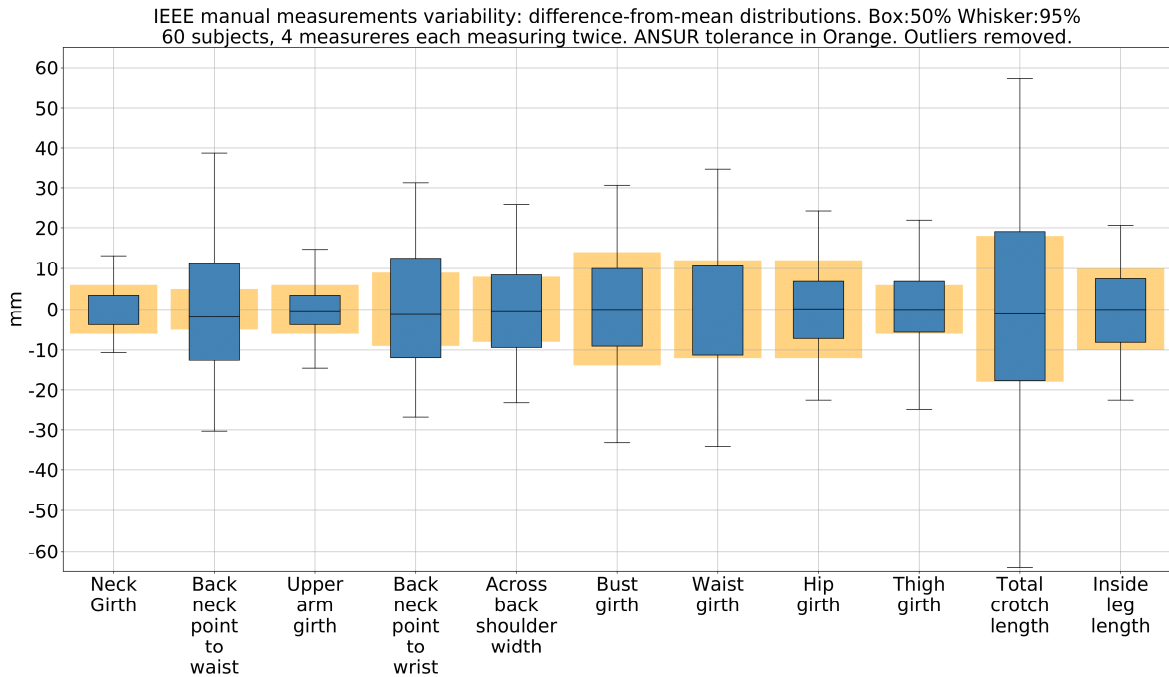


Fig. 3. Manual measurement variability: difference-from-mean distributions. The blue boxes contain 50% of the distribution and the whiskers bound 95% of the distribution. 60 subjects, 4 measurers each measuring twice. ANSUR[1] tolerance given in Orange. Number of differences-from-median within ANSUR[1] is given as percentages for each measurement.

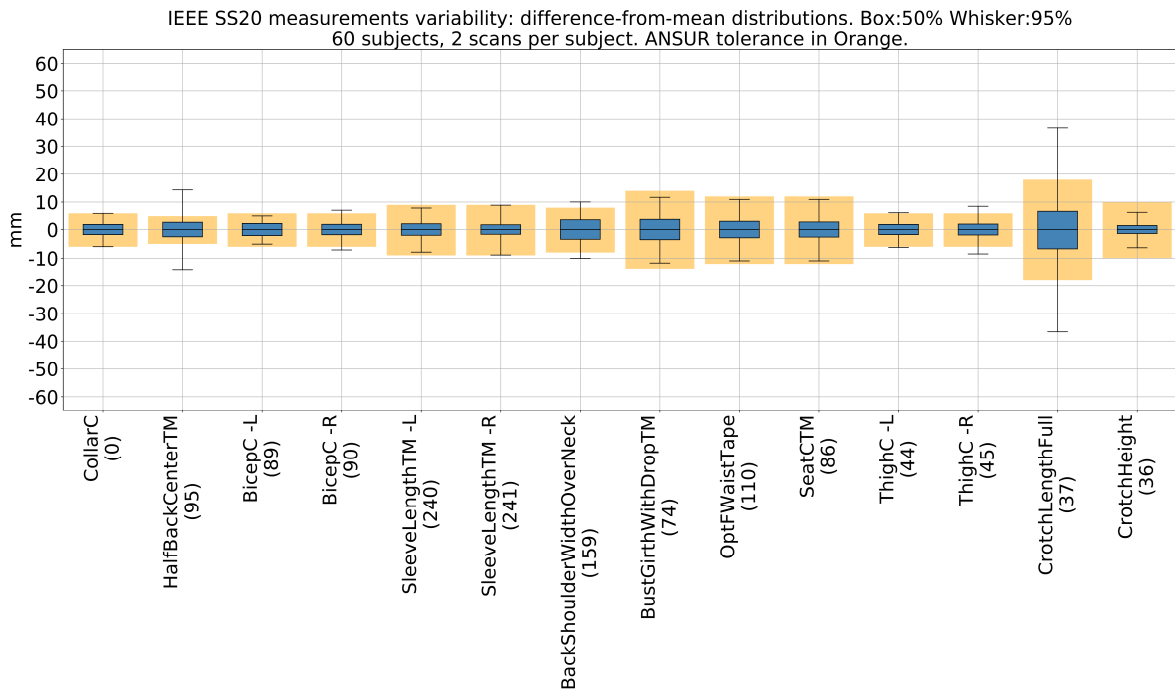


Fig. 4. Scanner measurement variability: difference-from-mean distributions. The blue boxes contain 50% of the distribution and the whiskers bound 95% of the distribution. 60 subjects, 2 scans each. ANSUR[1] tolerance given in Orange. Number of differences-from-median within ANSUR[1] is given as percentages for each measurement.

Figure 4 shows the boxplots of the difference-from-mean distributions for scanner measurements per measurement type (chest, waist, ...). Visually, the improvement over manual variability is quite striking. We see that roughly 95% of scanner measurements were within about a quarter of an inch of each other. When comparing the measurements to the Orange ANSUR[1] tolerance, the variability seen in Fig. 4 seems low. This gives us our first indication that scanner measurement reliability is greater than manual measurement reliability.

Figures 5 and 6 show the Coefficient of Variation (CV) and the Mean Absolute Deviations/Differences (MAD). Behind the blue bars representing the MAD numbers are grey bars that show the ANSUR[1] tolerance for that measurement. For each measurement, the left blue bar is the Mean Absolute Deviation (MADev) from mean, and the right blue bar is the Mean Absolute Difference (MADiff) between measurers. The MADev measures the deviation from the 'true' measurement, while the MADiff measures how different any two measurers could be. In Fig. 5, we see that 5 of the 11 manual measurements have about 70% of all deviations (MADev) within ANSUR[1] tolerance. However, 6 of the 11 measurements have MADev numbers higher than ANSUR[1] tolerances. None of the manual MADiff statistics are within the ANSUR[1] tolerances. The ANSUR[1] tolerances were crafted using a MADiff methodology. The CV shows that the amount of variation in manual measurements is between 1% and 4% with an average of 2.3%. This helps show variability without the need for units or scale. Figure 6 shows that all the scanner MADev values are within ANSUR[1] tolerances and only 3 of the MADiff values are outside of the ANSUR[1] tolerances. This is an indication that, for the majority of the measurements presented, scanner reliability is better than the 'best' manual variability realistically achievable. Here we see that the CV for scanner measurements ranges from 0.5% to 1.5% with an average of 0.9%. This is less than half that of the manual measurements.

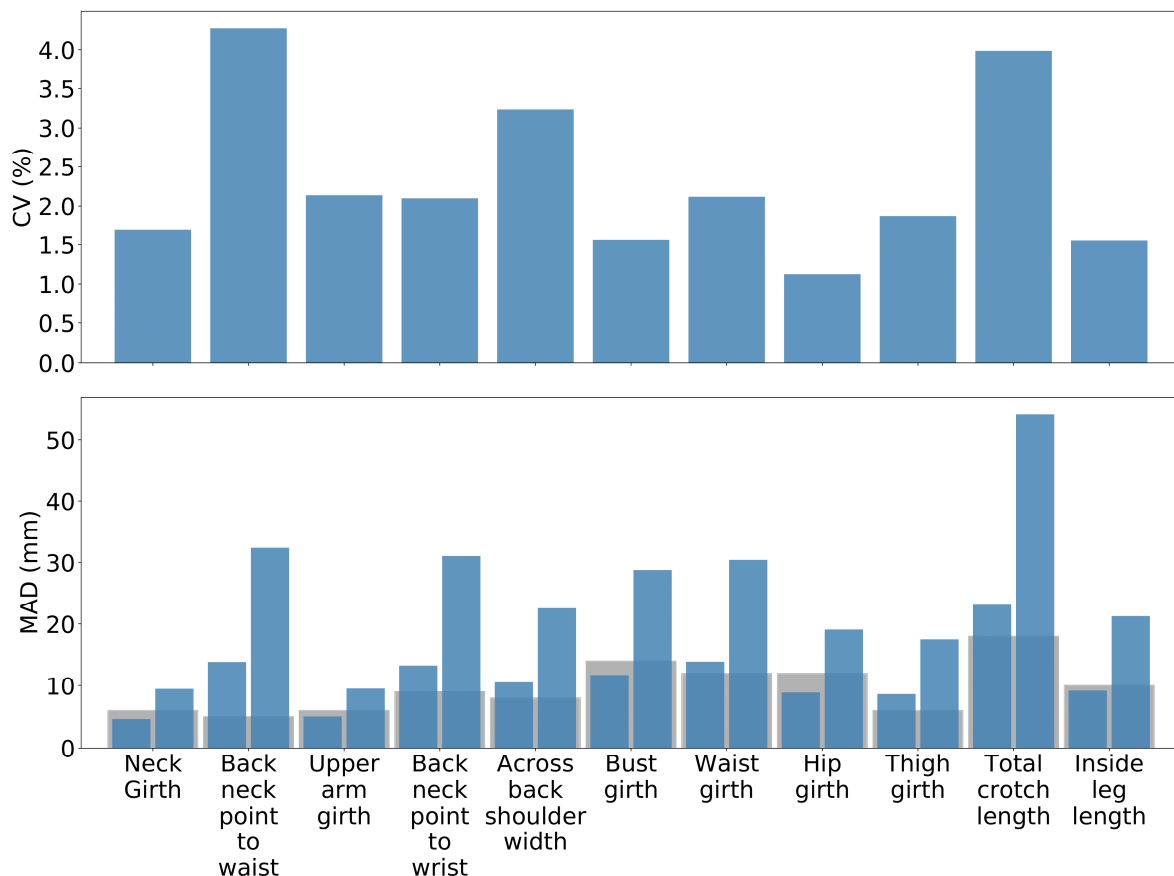


Fig. 5. Manual measurement variability statistics. Top: Coefficient of Variation given as a percentage. Bottom: For each measurement, the left blue bar is the Mean Absolute Deviation from mean and the right blue bar is the Mean Absolute Difference between measurers. Grey background is the ANSUR[1] tolerance. 60 subjects, 4 measurers each measuring twice.

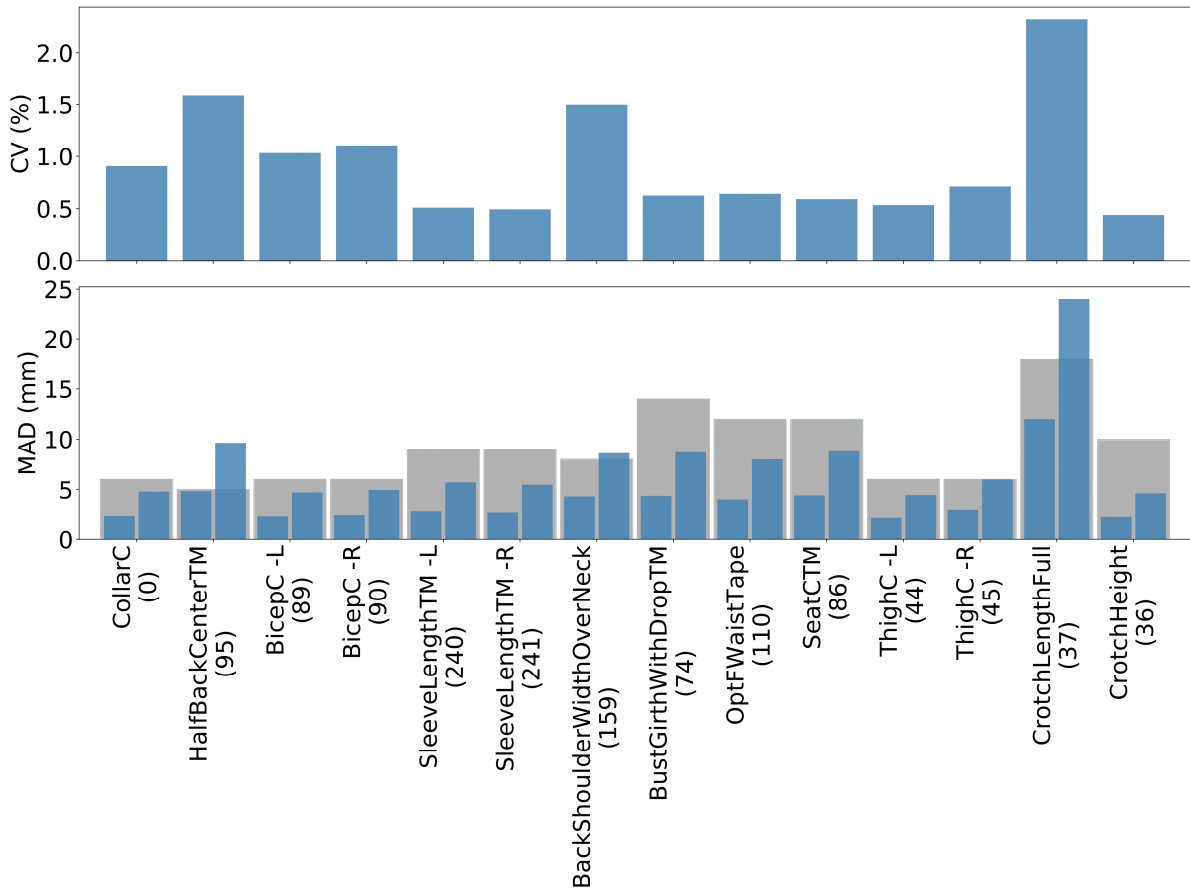


Fig. 6. Scanner measurement variability statistics. Top: Coefficient of Variation given as a percentage. Bottom: For each measurement, the left blue bar is the Mean Absolute Deviation from mean and the right blue bar is the Mean Absolute Difference between measurers. Grey background is the ANSUR[1] tolerance. 60 subjects, 4 measurers each measuring twice.

When compared to ANSUR[1] standards, the manual reliability is low. However, given that ANSUR[1] standards are probably too stringent to apply to a normal tailoring context, I would say that the manual reliability seen was rather high. Since the manual measurers chosen for this event were of high caliber, this sets a good benchmark for what the manual variability in the apparel world looks like. That said, we see that scanner reliability is significantly higher than the already high manual reliability. This means that the precision of scanners is certainly good enough to be of use in any industry that currently relies on manual measurements (of the type analyzed here).

3.2. Comparing Variances

We have seen that scanner measurements are more precise than the manual measurements for all the measurements under consideration. This statement can be verified by using standard statistical techniques for comparing the variance of two distributions. When using the F-test, for a particular person and a particular measurement, comparing the variance of the 6 to 8 manual measurements to the variance of the 2 scanner measurements just about always reveals that the hypothesis of equal variance *cannot* be rejected. However, we can also, for a particular measurement, use the average variance (across all people) and compare the variances of the two measurement techniques. This result shows that, for every measurement, the hypothesis of equal variance between manual and scanner measurements is rejected at greater than 99% confidence. In every case, the scanner precision or repeatability is better than that of manual measurement.

Table 1 gives the standard deviations for each measurement (in mm). For example, to get a 95% confidence interval for the scanners ability to measure the Hip Circumference, we obtain $x \pm 1.96 * table_value$, which is $x \pm 1.96 * 9.51$ resulting in $x \pm 18.64$ mm.

manual name	scanner name (enum)	sqrt(ave_var_scanner)	sqrt(ave_var_manual)	sqrt(ave_var_scanner+ave_var_manual)
Neck Girth	CollarC(0)	4.43	6.52	7.88
Back neck point to waist	HalfBackCenterTM(95)	9.76	19.01	21.37
Upper arm girth	BicepC -L(89)	4.18	7.62	8.69
Upper arm girth	BicepC -R(90)	4.68	7.55	8.89
Back neck point to wrist	SleeveLengthTM -L(240)	5.7	16.93	17.87
Back neck point to wrist	SleeveLengthTM -R(241)	5.74	16.96	17.9
Across back shoulder width	BackShoulderWidthOverNeck(159)	7.62	14.32	16.22
Bust girth	BustGirthWithDropTM(74)	7.79	16.8	18.52
Waist girth	OptFWaistTape(110)	7.73	19.63	21.09
Hip girth	SeatCTM(86)	9.51	13.25	16.31
Thigh girth	ThighC -L(44)	4.06	12.61	13.25
Thigh girth	ThighC -R(45)	7.17	12.61	14.51
Total crotch length	CrotchLengthFull(37)	25.28	32.72	41.35
Inside leg length	CrotchHeight(36)	4.75	12.8	13.65

Table 1. Scanner measurement variability statistics. Top: Coefficient of Variation given as a percentage. Bottom: For each measurement, the left blue bar is the Mean Absolute Deviation from mean and the right blue bar is the Mean Absolute Difference between measurers. Grey background is the ANSUR[1] tolerance. 60 subjects, 4 measurers each measuring twice.

4. Accuracy or Definitions or Compatibility

The goal here is to see how accurate each measurement technique is. That is, we examine if the two techniques measure the same thing.

4.1. ISO 20685

Figures 7 and 8 show the results of an ISO 20685 analysis of the manual and scanner measurements (in inches). The below plot records all combinations of scanner minus manual measurements for each person. That means that if we had 8 manual measurements and 2 scanner measurements for each subject (this is true for almost all subjects), then we would record 16 differences between scanner and manual measurement.

The plot shows the mean of these differences as points and the 95% confidence interval for the mean as the error bars. The grey bars indicate the ISO tolerance. For two measurements to be considered equivalent from an ISO 20685 perspective, the entirety of the confidence intervals needs to be within the grey boxes. Thus, we can see that only Chest Girth has a scanner measurement (BustGirthWithDropTM) that is equivalent.

Note that:

- These plots only report the average difference between scanner and manual measurements. The spread of these differences is rather large (see Table 1). This means that even though, on average, scanners (without any bias-shift) measure neck circumference larger than manual measurements, any particular measurement still has a high probability of scanners measuring smaller than manual measurements.
- These differences can sometimes be explained by the scanner algorithms and the manual measurers using different measurement definitions. For example, manual waist measurements involve palpation to find the lower rib and upper thigh bones. This is impossible for a scanner and thus the scanner algorithms look for different markers.
- Even when measurement definitions agree between manual and scanner, the differences can have many explanations. For example:
 - Compression: for circumferences, manual measuring techniques require some compression to counter forces wanting to pull the tape measure off the desired measurement line. 2mm of reduction in radius due to compression results in circumference change of 12.6mm, which is 1/2 an inch.
 - Gravity: Even though compression tries to prevent gravity from pulling the tape measure off the desired measurement line, in some cases it doesn't work. For example, chest measurements: the chest is shaped like an inverted cone (big side up) and it is pretty much impossible with only one point of pinning to keep a tape measure horizontal. This is why the scanner's 'WithDrop' measurement is more successful at matching manual measurements than a horizontal measurement.

Psychological: The shape of the subject changes when faced with someone with a tape measure versus when standing in a booth. This effect is mostly seen in the waist area (sucking-in).

- We could have chosen to just compare the mean scanner measurement against the mean manual measurement. While that doesn't really change the location of the points in Fig. 7, it increases the size of the confidence interval because there is less data to predict from. This is an incorrect approach since it is saying, 'let's compare a measurement technique involving 8 manual measurers against a measurement technique involving 2 scanners'. This is not the test we are trying to do.
- In the next section, we will report on a standard 2-way ANOVA (with replication) analysis. These results could also be used to create Fig. 7}. This has been done and the results are equivalent (only small variations in location and confidence interval size). The only difference is that number of points used is sometimes less than the number required by the ISO standard. This consistency can be seen as confirmation of the validity of these results.

When the variability of the scanner is lower, then it is common practice to allow the scanner to adjust for a bias-shift. This could be accomplished using something similar to the custom scripting feature included with SS20 scanners. This feature was designed for this type of situation, and can be thought of as fine-tuning our scanner for each particular use-case. A caveat is that before applying bias-shifts, one should first ensure that the scanner measurement-lines are close to the corresponding manual measurement line. The results of this survey (or one like it) would give us the mean measurement difference (for each measurement like chest, waist, ...) and then we could use to modify scanner output to account for the bias. This process is sort of like a tare function on a scale. Which would mean that, going forward, the differences between scanner and manual measurement would look more like what is seen in the Fig. 8. Figure 8 has the means (green points) at zero by construction, but the important observation is that the green standard-error bars are wholly within the grey ISO tolerance boxes. Now we see that we can expect all these scanner and manual measurements to be considered equivalent or compatible from an ISO 20685 perspective.

IEEE SS20 scanner minus manual accuracy: Mean (and 95% SE) of all measurement differences. ISO20685 tolerance in grey. X-axis: Scanner names below, manual names above.

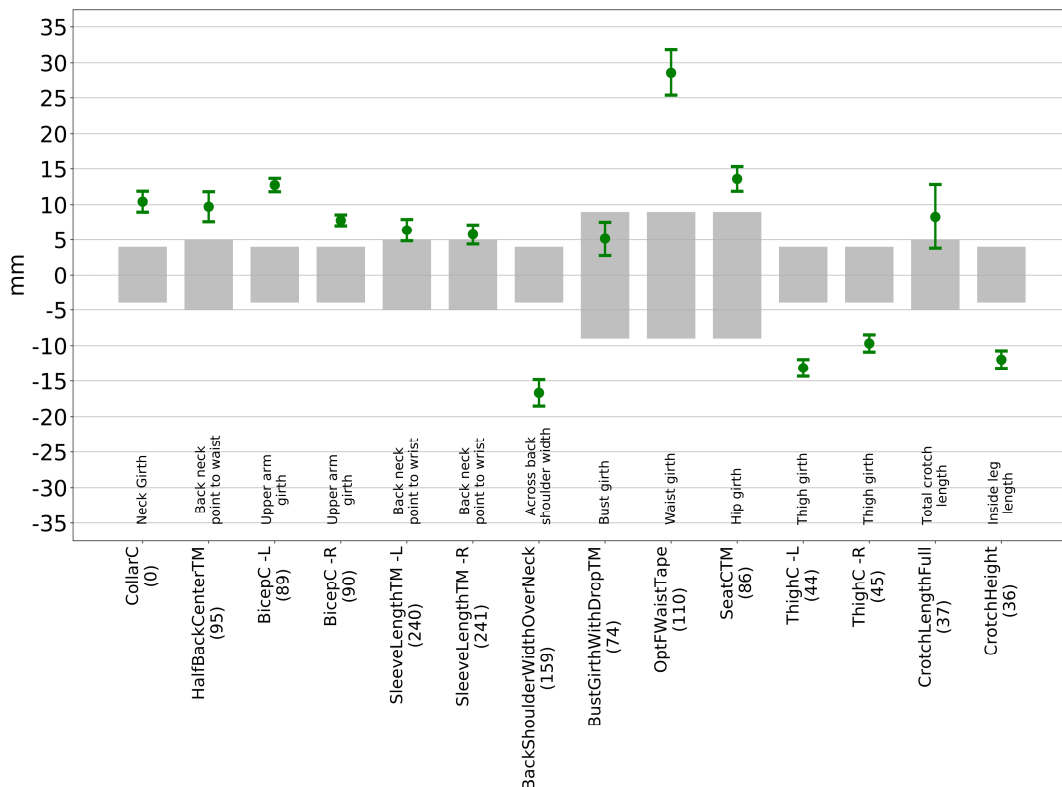


Fig. 7. Scanner minus manual accuracy. Data points are mean difference and the error bars represent the 95% confidence in the mean (standard error). The ISO 20685 tolerance is in grey. Labels above x-axis are manual measurement names and labels below x-axis are scanner measurement names.

IEEE SS20 scanner minus manual accuracy: Mean (and 95% SE) of all measurement differences.
 ISO20685 tolerance in grey. X-axis: Scanner names below, manual names above.
 Using scripting to remove bias.

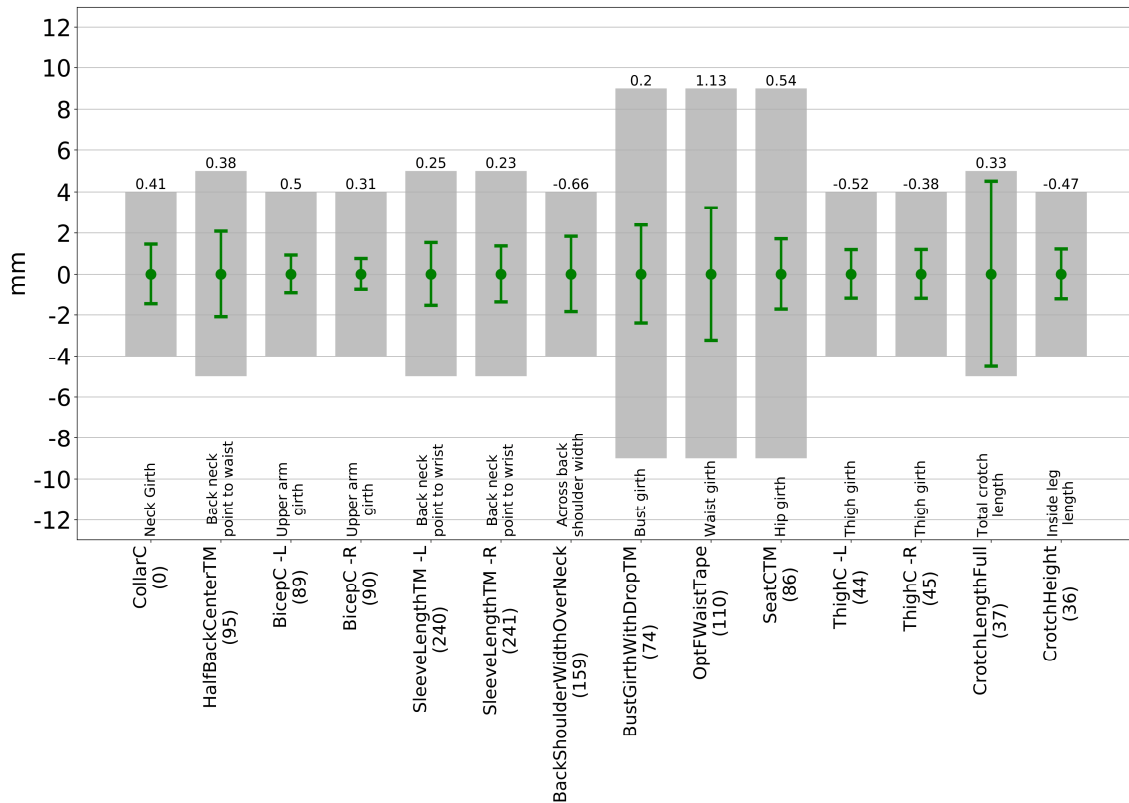


Fig. 8. Scanner minus manual accuracy. Data points are mean difference and the error bars represent the 95% confidence in the mean (standard error). The ISO 20685 tolerance is in grey. Labels above x-axis are manual measurement names and labels below x-axis are scanner measurement names.

By ISO 20685 standards, the raw scanner measurements and manual are not measuring the same measurements (with the exception of Chest). This motivates the need for a layer in between scanner output and downstream applications that are expecting manual measurements. This layer would mean that all the scanner and manual measurements can be considered equivalent or compatible by ISO 20685 standards.

4.2. Two-Way ANOVA (with replication)

This set of data is exactly of the form the 2-way ANOVA analysis is designed to process. The replication is uneven because the two measurement techniques (called treatments by ANOVA) have a different amount of measurements per person.

4.2.1. Straightforward Implementation

A simple execution of this method on the measurement data shows the following hypothesis decisions:

- The hypothesis that each person has the same mean is rejected (at extremely high confidence). This is an easy check since it is quite obvious that different people have different shapes.
- The hypothesis that the two measurement techniques have the same mean is rejected (at extremely high confidence). This is a verification of the findings of the ISO test above.
- The hypothesis that there is no 'interference' between technique and subject is rejected (at extremely high confidence). This is basically a statement that there is a lot of variability still in the system even after subject differences and technique differences are accounted for. In other words, yes there is a lot of variance inherent in measuring people that has nothing to do with a person's shape or the measuring technique (as independent factors).

An analysis of the power or significance of these effects show that: The amount of variability coming from subject differences vastly exceeds the variability from measurement technique. This is good because if measuring techniques had more variability than people, we could never use them for measuring people.

4.2.2. Implementation with people variability removed

We can also do a reduced analysis where we subtract off the average measurement (across manual and scanner measurements for each person) from each measurement. This would remove all variability due to subject. This gives:

- The hypothesis about subject means being the same can now not be rejected. We just removed this variability.
- The hypothesis that the two measurement techniques have the same mean is still rejected (at extremely high confidence).
- The hypothesis that there is no 'interference' between technique and subject is still rejected (at extremely high confidence).

The analysis of the power or significance of these effects show that, while the variance due to subject is zero (as expected), the variance due to technique is smaller (mostly vastly smaller) than the variance due to other factors.

4.2.2. Implementation with people variability removed and bias adjustment

An additional modification is if we implemented the bias-shift discussed above. This is a modification of the scanner measurements in addition to the modifications needed to remove the people variability. This gives:

- The hypothesis about subject means being the same cannot be rejected. We have removed this variability.
- The hypothesis that the two measurement techniques have the same mean cannot be rejected. This shows that the bias-shift works as advertised.
- The hypothesis that there is no 'interference' between technique and subject is still rejected (at extremely high confidence).

The analysis of the power or significance of these effects show that the variance due to subject and technique is zero (as expected). Leaving the variances inherent in measuring people as the only remaining contributor of variance.

There is enough statistical evidence to reject the hypothesis that the mean of the manual measurements is the same as the mean of the scanner measurements. However, when the bias is accounted for, then the hypothesis that the mean of the manual measurements is the same as the mean of the scanner measurements remains with very high statistical confidence. In other words, these two techniques measure the same thing. If manual measurement is the definition of truth, despite its imperfections, then these results show that scanner measurements are highly accurate.

5. Conclusion

The goals of this study were to examine the reliability and compatibility of measurements obtained by two different measurement techniques. The two techniques were firstly, manual measurement by skilled apparel professionals using just a tape measure, and secondly, measurements extracted from 3D body scans performed by a Size Stream SS20 scanner. The measurements used for the comparison were those typical to the retail apparel industry. The study measured over 60 people and each measurement technique was performed multiple times on each subject.

The reliability of both techniques was evaluated using CV, MADev, and MADiff. Manual measurements were shown to have an average CV of 2.3%, which is quite good. Each individual scanner measurement was more reliable than the corresponding manual measurement and the average CV of them all was 0.9%, less than half that of manual measurements. Furthermore, 12 out of 14 of the evaluated scanner measurements were shown to be very near, or well below, the ANSUR[1] tolerance for these measurements.

The compatibility of these measurement techniques was evaluated according to ISO 20685 and by using a 2-way ANOVA (with replication). The ISO 20685 metric showed that there is enough evidence

to conclude that the two techniques, when a bias-shift is permitted, are compatible. The ANOVA came to the same conclusion and furthermore revealed that the variance due to technique is smaller, mostly vastly smaller, than the variance due to other factors.

Appendix A: Statistical Definitions

Consider any particular measurement, like chest girth for example. Given I scan subjects and J measurers, if x_{ij} is a measurement for the i -th subject by the j -th measurer then:

- The mean measurement for the i -th individual is $\bar{x}_i = \frac{1}{J} \sum_{j=1}^J x_{ij}$
- The measurement standard deviation for the i -th individual is $s_i^2 = \frac{1}{J-1} \sum_{j=1}^J (x_{ij} - \bar{x}_i)^2$
- Coefficient of Variation (CV): the standard deviation divided by the mean.
- MADev: The deviations are $|x_{ij} - \bar{x}_i| \forall i, j$. The mean of these deviations is the MADev.
- MADiff: Let $\Delta_{lm} x_i = |x_{il} - x_{im}| \forall l > m$. Also let $\bar{\Delta}_{lm} = \frac{1}{I} \sum_{i=1}^I \Delta_{lm} x_i$. Then the MADiff is $\max(\bar{\Delta}_{lm} \forall l > m)$. This definition follows the ANSUR[1] standard.

Appendix B: Two-way ANOVA

The two "ways" or, two hypotheses, of this ANOVA are

- H1 = There is no difference beyond random fluctuation between people.
- H2 = There is no difference beyond random fluctuation between scanner measurements and manual measurements.

The table below gives the average p-values (over the 14 measurement sites) associated with the two ANOVA hypothesis tests:

Data Used	P1	P2
<i>Raw</i>	0	0
<i>Bias Shift</i>	0	0.89
<i>People Shift</i>	1	0
<i>People & Bias Shift</i>	1	0.97

References

[1] 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., and Kristensen, S. 2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics. 2015. Report No. NATICK/TR-15/007.