National Defence Défense nationale

DEFENCE RESEARCH AND DEVELOPMENT CANADA (DRDC)
RECHERCHE ET DÉVELOPPEMENT POUR LA DÉFENSE CANADA (RDDC)

Methodology to evaluate three-dimensional (3D) body scanning systems in support of the Canadian Armed Forces Anthropometric Program for Soldier System Acquisition (CAPSSA) program

Allan Keefe
DRDC – Toronto Research Centre

Prepared for: Emma Moon, Human Factors Support Cell, Director Soldier Systems Program Management.

Defence Research and Development Canada

Reference Document
DRDC-RDDC-2024-D010
April 2024

Canada

IMPORTANT INFORMATIVE STATEMENTS

This document was reviewed for Controlled Goods by Defence Research and Development Canada (DRDC) using the Schedule to the *Defence Production Act*.

Endorsement statement: This publication has been published by the Editorial Office of Defence Research and Development Canada, an organization of the Department of National Defence of Canada. Inquiries can be sent to: Publications.DRDC-RDDC@drdc-rddc.gc.ca.

Abstract

The Department of National Defence (DND) and the Canadian Armed Forces (CAF) require current and representative anthropometric (body size and shape) data to design, specify, acquire and procure clothing, individual equipment, platforms and workstations to ensure operator safety and optimal performance of a diverse work force. To address this challenge, the Director Soldier System Program Management (DSSPM) has initiated the CAF Anthropometry for Program for Soldier System Acquisition program (CAPSSA), which endeavours to continuously gather, analyze and manage anthropometric data of CAF members from recruitment and periodically throughout their careers. An important aspect of this project is the identification and acquisition of multiple three-dimensional (3D) body scanning systems to acquire body dimensions of CAF personnel.

In response to a request from the Human Factors Support Cell (HFSC) of DSSPM, Defence Research and Development Canada (DRDC) – Toronto Research Centre identified objective methodologies to evaluate the reliability and accuracy of candidate 3D body scanning systems based on international and operational standards. This Reference Document describes three recommended methods, associated statistical analysis and performance standards for evaluating 3D body scanning systems. These methods include: 1) scanning a calibrated test sphere, 2) scanning a manikin/dress form and 3) scanning human participants. Advice on how the evaluation methods can be used for rejecting, accepting, or rating candidate 3D body scanning systems is also provided.

Significance to defence and security

This Document provides advice pertaining to the objective evaluation of candidate 3D body scanning systems in support of the CAF Anthropometry for Program for Soldier System Acquisition program (CAPSSA). The evaluations methods identified are based on a review of literature and international standards and will enable DSSPM to assess candidate 3D body scanning systems against international and operational standards.

Résumé

Le ministère de la Défense nationale (MDN) et les Forces armées canadiennes (FAC) ont besoin de données anthropométriques actuelles et représentatives (taille et forme du corps) pour concevoir, spécifier, acquérir et obtenir des vêtements, de l'équipement individuel, des plateformes et des postes de travail pour assurer la sécurité des opérateurs et le rendement optimal d'un effectif diversifié. Pour cela, le Directeur—Administration du programme de l'équipement du soldat (DAPES) a lancé le Programme anthropométrique des FAC pour l'acquisition du système du soldat des FAC (PAASSF), qui s'efforce de recueillir, d'analyser et de gérer continuellement les données anthropométriques des membres des FAC dès leur enrôlement et de manière périodique tout au long de leur carrière. Un aspect important de ce projet est l'identification et l'acquisition de plusieurs systèmes de balayage corporel tridimensionnel pour obtenir les dimensions physiques du personnel des FAC.

En réponse à une demande de la Cellule de soutien aux facteurs humains (CSFH) du DAPES, le Centre de recherche de Toronto de Recherche et développement pour la défense Canada (RDDC) a défini des méthodologies objectives pour évaluer la fiabilité et l'exactitude des systèmes de balayage corporel tridimensionnel potentiels en fonction des normes internationales et opérationnelles. Le présent rapport décrit trois méthodes recommandées, une analyse statistique qui s'y rattache et les normes de rendement pour évaluer les systèmes de balayage corporel tridimensionnel. Ces méthodes comprennent : 1) le balayage d'une sphère d'essai calibrée, 2) le balayage d'un mannequin ou d'une tenue et 3) le balayage de participants humains. Des conseils sur la manière dont les méthodes d'évaluation peuvent être utilisées pour rejeter, accepter ou noter les systèmes de balayage corporel tridimensionnel potentiels sont aussi offerts.

Importance pour la défense et la sécurité

Le présent rapport offre des conseils sur l'évaluation objective des systèmes de balayage corporel tridimensionnel potentiels à l'appui du Programme anthropométrique des FAC pour l'acquisition du système du soldat des FAC (PAASSF). Les méthodes d'évaluation présentées sont fondées sur un examen de la documentation et des normes internationales et permettront au DAPES d'évaluer les systèmes de balayage corporel tridimensionnel par rapport aux normes internationales et opérationnelles.

Table of contents

List of figures

List of tables

1 Introduction

1.1 Background

The Department of National Defence (DND) and the Canadian Armed Forces (CAF) require current and representative anthropometric (body size and shape) data to design, specify, acquire and procure clothing, individual equipment, platforms and workstations to ensure operator safety and optimal performance of a diverse work force. Currently, the most recent anthropometric data available is the 2012 Canadian Forces Anthropometric Survey (CFAS) (Keefe, Angel, & Mangan, 2015) which contains body measurement and three-dimensional (3D) scan data of 1890 men and 315 women across three services.

Demographics of a population are a key driver of the anthropometric diversity of a population with race, sex, age and occupation being cited as important determinants (Garneau & Parkinson, 2012). Over time, demographic and secular changes within a population result in a change in anthropometry as evident by the differences observed in key body dimensions between the 1997 Canadian Land Forces Survey (Chamberland, Carrier, Forest, & Hachez, 1997) and the 2012 CFAS, with male Australian personnel from 1997 to 2010 (Tomkinson, Daniell, Fulton, & Furnell, 2016) and U.S Army personnel from 1988 to 2012 (Gordon et al., 2014).

Policy changes can also influence anthropometric trends. For example, Canada's Defence Policy, Strong Secure and Engaged (SSE) is committed to ensuring that the DND/CAF better reflects Canada's diversity and sets a target to increase the percentage of women in the CAF from 15% to 25% by 2027. It is anticipated that this will result in a wide spread shift in demographics and concomitant change in the anthropometric distribution of the CAF.

With the intended changes in the composition of the CAF, it is important to monitor the anthropometry of the CAF to identify any potential changes in body size and shape and provide this information early so that decision makers are provided with up-to-date anthropometric data to assess the adequacy of current soldier systems, vehicles and platforms and to inform the development and specification for future system design, development and acquisition.

To support the objectives of SSE, a Joint Chief of Defence Staff / Deputy Minister Directive: Op Generation was issued in May of 2018 with the mission to conduct recruiting operations to meet immediate growth requirements and employment equity goals and to set the conditions for enduring personnel levels. Within Op Generation, Assistant Deputy Minister (Materiel) (ADM[Mat]) is to ensure that sufficient clothing is available for all basic military training and that ADM(Mat) works with Military Personnel Command (MPC) and Assistant Deputy Minister (Defence Research and Development Canada) (ADM[DRDC]) to "systematically evaluate the anthropometric requirements for a wider range of body time to meet the clothing and equipment needs of a more diverse force."

To address this challenge, the Director Soldier System Program Management (DSSPM) has initiated the CAF Anthropometry for Program for Soldier System Acquisition program (CAPSSA), which endeavours to continuously gather, analyze and manage anthropometric data of CAF members from recruitment and periodically throughout their careers. This will be enabled, in part, through the acquisition and deployment of a fleet of automated 3D body scanning systems that will be distributed at key Defence centres across Canada.

To support the evaluation of candidate 3D body scanning systems, the Human Factors Support Cell (HFSC) of DSSPM requested assistance from DRDC – Toronto Research Centre to aid in the development of an objective methodology, or methodologies, to evaluate candidate body scanning systems. In response, this Reference Document provides guidance on the following requirements for evaluating 3D body scanning systems:

a. Objective measures to quantify body scanner accuracy and validity.

b. Accepted body scanner measurements standards.

c. Ability to discriminate between competing systems.

d. Identifying and evaluating key measures of interest to the CAPSSA program.

e. Feasible to implement within time and logistic constraints.

It is recommended that the methodologies and standards described in this Document be considered for implementation as part of the CAPSSA bid evaluation to identify and select the best qualified 3D body scanning system that meets the performance objectives of the program.

2 Methods to evaluate three-dimensional body scanning systems

2.1 Overview

3D body scanning systems provide an alternate to traditional body measurement methods that use tools such as measuring tape and calipers. By providing a touchless, automated, and rapid method of body measuring, many individuals can be quickly assessed while maintaining issues related to privacy and intimate contact by the measurer. This is accomplished by using technologies such as lasers (Kuehnapfel et al., 2016), structured light (Pokorny et al., 2019), infra-red time of flight cameras (Bragança et al., 2016) and photogrammetry (Meunier & Yin, 1999) to capture and model the body shape of the individual.

In the 30+ years since the Loughborough Anthropometric Shadow Scanner (LASS) was first introduced (Jones et al., 1989), 3D body scanning systems have evolved to the point that acceptable body scan images can be obtained from a smartphone (NATO, 2020). The cost of body scanning systems has decreased dramatically as low-cost infra-red sensors are adopted and the adoption of body scanning systems by the fitness and fashion industry increases (Daanen & Ter Haar, 2013).

When assessing a 3D body scanning system, it is important to evaluate both accuracy and reliability in taking anthropometric measurements. Accuracy of a 3D body scanning system is determined by how well its sensors can represent a 3D object and the ability of the software to automatically identify body landmarks and make linear, or geodesic measurements based on these landmarks. In other words, accuracy refers to the agreement of a scan measured value to its "true" value. There are many factors that may affect the accuracy of a 3D body scanning system. For example, error in landmarking, adiposity, posture, poor quality scan models and body hair can all result in measurement errors.

Reliability refers to the consistency of measures across operators and repeated tests. Random error affects the precision of a system by introducing noise into the measurement. A system that shows good reliability, but poor accuracy may be affected by systematic error. Reliability may be affected by changes in posture, breathing cycle, calibration drift and measurement technique. As both accuracy and reliability can be affected by similar factors, it is important to ensure that potential external sources of error are identified and mitigated, where possible.

2.2 Scanning test objects

To evaluate the theoretical reliability of a 3D body scanning systems, it is important to employ a test methodology that eliminates potential sources of error that are due to characteristics of the scanned object. As the human is a key source of measurement error, system reliability can be evaluated by using a stable and rigid object such as a test sphere or rigid human manikin or dress form.

2.2.1 Test sphere

A test sphere is a precision engineered metallic or ceramic object that is used in the field of metrology to validate and calibrate coordinate measuring systems. Its application towards evaluating 3D body scanning systems was proposed by Kouchi et al. (2012) who recommended scanning a 120 mm sphere placed at various locations within the scan volume and reporting on metrics such as the error of diameter measurement and spherical dispersion value (Figure 1).

This methodology has been adopted by the International Standards Organization (ISO) as standard 20685-2 (ISO, 2015). While the application of the sphere-based evaluation process is likely outside the scope of the CAPSSA bid evaluation process, it is possible that the manufacturers of the body scanning systems can provide a certificate to verify that this process was completed. Unfortunately, Kouchi does not provide recommendations on minimum scan accuracy, however, ISO 20685 (2010) does provide maximum allowable error standards. It is recommended that the small circumference maximum allowable error of 4 mm be adopted as the standard for a test sphere evaluation.

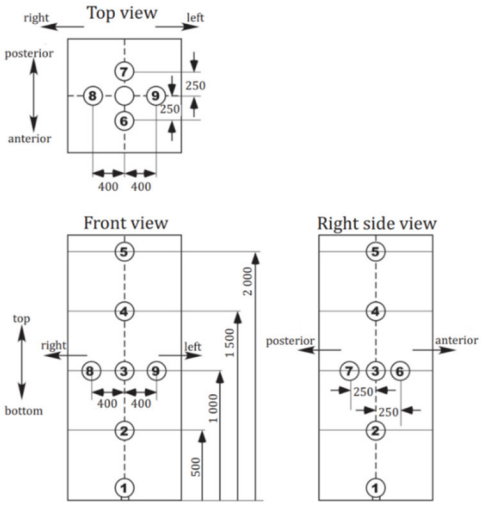

Figure 1: *Location of the calibration sphere when evaluating 3D body scanning systems.*

2.2.2 Manikin/dress form

Evaluations of 3D body scanning systems have been made using a rigid manikin as a surrogate for, or complement to, human testing (Lu & Wang, 2010; Bragança et al., 2017). Advantages of using a manikin or dress form include the elimination of postural sway and torso movement due to breathing, elimination of error due to body hair, and the effect of clothing bulk.

As a manikin is rigid, verification of landmark accuracy can be difficult for bony landmarks that must be palpated for identification (e.g., acromion, trochanterion and 10th rib). Second, many body scanning systems require a standardized pose (e.g., ISO pose "A") to facilitate landmarking and measuring. Hence, for optimal results, it is important to use a rigid manikin that is already formed in this pose (Figure 2).

Figure 2: *Example of ISO pose "A." (Source: Image created by author using Daz Studio Pro v4.12, DAZ3D.com. [Accessed date: 22 January 2024]).*

2.2.3 Human participants

While an evaluation of a 3D body scanning system using a test sphere or manikin is useful for measuring a body scanning system's accuracy, it does not necessarily provide a realistic test of the scanner's operational capability to effectively measure a wide range of human in an operational context. Body features such as arm placement, axillary folds, adiposity, body shape variability and body hair can challenge scan capture and measurement algorithms (Kouchi & Mochimaru, 2011) and requires sophisticated post processing of the scan data to create a suitable model for measurement extraction (Li & Li, 2010). In general, the key steps in generating a finalized mesh model from a raw 3D point cloud scan include:

- Cleaning artifacts and merging scan patches into a point cloud image.
- Conversion to a mesh object and decimation to a standard number of vertices.
- Template fitting or hole filling to create a watertight model.
- Automated landmark detection.
- Measurement extraction.

Thus, when evaluating 3D body scanning systems using humans the accuracy of the measurement is a function of how well complex features of the body are captured and modelled and how well the automated landmark detection and measuring software algorithm's can accurately and reliability record anthropometric measures across a wide range of body types. Reliability of these processing steps must also be evaluated to ensure that these steps are consistently applied within and between participants.

2.3 Measurement accuracy

When applying the test sphere method, it is the accuracy of the scanner sensing technology (i.e., laser, depth sensing infrared radiation [IR] cameras, etc.) to represent a geometric shape that is being evaluated. As the sphere is a smooth, rigid object of simple geometry that is engineered to fine tolerances, simple measurements of circumference and width can be made and compared to the known manufacturing specifications. Unlike measuring a human body, the measurement algorithms used for the test sphere do not need to account for complex surface features or estimate body landmarks. Thus, the measurements can be very accurate and reliable.

2.3.1 Comparing scan derived measurements to manual measurements

Establishing the ground truth measure of a body dimension is difficult to obtain when measuring the human body. Typically, manual measurements taken by experienced anthropometrists have been used as a benchmark reference, and scan derived measures are assessed for their equivalence to manual measurements. Using manual measurements as a benchmark is not without its own inherent problems. For example, manual measures can be affected by diurnal variations, variation in fluid and feeding state, inconsistent measuring technique, soft tissue compression or misreading of measuring tools (Kouchi et al., 1999). Absolute accuracy of manual measurements is also difficult to establish, particularly with measurements that are dependent of palpating deep bony landmarks. Conversely 3D body scanner accuracy may be affected by such effects as the clothing worn, body hair, posture and body sway. However, landmarking accuracy remains a principal challenge for scanning systems (Kouchi & Mochimaru, 2011). Despite these challenges, a systematic review of the accuracy and reliability of 3D body scanners by Rumbo-Rodríguez et al. (2021) reveals that the standard for evaluating 3D scanners, for better or worse, is to compare the measuring accuracy to that of trained measurers.

For clothing sizing and issuing application, an additional consideration that must be considered is the process that the manufacturer uses to develop their sizing scheme. The rigour of the in-house method for determine body measurements is often unknown and may not necessarily agree with measurements taken from a body scanner. Second, variability due to manufacturing tolerances of clothing production means that there can be a slight variation between two garments of the same size. Added to this is the influence of user preference in size selection. All these factors can result in a mismatch between the user's anthropometric measurements and the preferred garment size. This is exemplified by a study by Daanen (2014) that demonstrated most users not selecting the size of military jacket or trouser sizes that agreed with their body measurements.

2.3.2 Objective measures of accuracy

In their review, Rumbo-Rodriquez et al. detail numerous statistical methods that have been proposed to evaluate the accuracy of body scanners to a benchmark reference measurement. A sample of these methods include the following:

- Technical error of measure (TEM) and percent TEM (%TEM),
- Typical error (%),
- Percent coefficient of variation (%CV),
- Root mean square error (RMSE),
- Standard error of measurement (SEM),
- Mean absolute difference (MAD), and
- Paired t-tests.

A challenge in interpreting these standards is that it is difficult to compare studies that use different metrics of accuracy. Second, while measurement accuracy can be straightforward, it is important that the results can be interpreted within an operation context. For example, what do these measures of accuracy mean and what level of accuracy is required by the CAPSSA program? To answer these questions, two approaches have been followed to define meaningful standards for scan accuracy.

2.3.2.1 Accuracy comparison to anthropometric standards

The first method involves utilizing measurement standards used for training and certifying manual measures to set minimum limits of reliability and provide a method of ensuring consistent quality of measures over time. This method is preferred to statistical techniques such as paired t-tests where the significance of differences between measures is affected by the sample size. Additionally, a statistical difference may not be meaningful from an operational perspective. Standards specific to anthropometry measurement error are provided by ISO 20685 (ISO, 2010) and the International Society for the Advancement of Kinanthropometry (ISAK: Norton, 2019). The difference between the two standards are as follows:

- **ISO 20685**—Measure error is based on the standards developed for the 1988 Anthropometric Survey of United States (US) Army Personnel (ANSUR: Gordon et al., 1988). Minimum allowable observer errors (AOEs) are calculated based on the MAD between pairs of individual measures. These can be based on intra- or inter-observer measurements. AOEs differ based on the measurement in question. Annex A provides the formula for calculating MAD and reference AOE tables for several types of body measures.
- **International Society for the Advancement of Kinanthropometry (ISAK)**—Is similar in concept to the ISO 20685 standard but is based on the %TEM. Standards for %TEM for skilled anthropometrists are 1.0% for intra-measurer error and 1.5% for inter-rater reliability. The standard is not as stringent for beginner anthropometrists at 1.5% for intra-measurer error and 2.0% for inter-rater reliability (Periniet et al., 2005). The formulas for the absolute and relative TEM are provided in Annex A.

2.3.2.2 Comparison to garment tolerances

A second method involves comparing the accuracy of a body scanning systems to garment manufacturing tolerances. For example, in an evaluation of the BoSS XI system, Meunier & Yin (2000) proposed the use of garment manufacturing tolerances to establish measurement differences that would have minimum effects on garment fitting. Examples of tolerances for Canadian Forces dress trousers and shirts include ± 13 mm for waist, neck and chest circumferences and sleeve length and ± 3 mm for neck circumference.

Bradtmiller & Gross (1999) consulted expert tailors to determine measurement tolerances for garment measures like those evaluated by Meunier and Yin. Recommended measurement tolerances ranged from ± 6.4 to ± 12.7 mm for all measures except for neck circumference which was determined to be ± 6.4 mm by all tailors. If sizing and fitting of clothing is an objective of CAPSSA, then garment tailoring tolerances may be considered as a potential metric to evaluate minimum acceptable scanner accuracy.

Finally, Vonk & Daanen (2015) utilized SEM to determine the accuracy of the SizeStream body scanning system, setting a limit of ≤ 10 mm as being an acceptable accuracy for use in the garment industry. The authors admit that the SEM limit of 10 mm has not be validated as a suitable threshold for garment sizing and fitting application and that further research is required to confirm its validity.

2.3.3 Repeatability

In addition to system accuracy, repeatability or test-retest reliability is important to quantify how consistent the body scanning system is at measuring a participant. Like measurements for accuracy, several methods have been used to quantify repeatability. An overview of the studies using these following methods may be found in Rumbo-Rodríguez et al. (2021). These methods include:

- Intraclass correlation coefficient (ICC),
- Concordance correlation coefficient (CCC),
- Overall concordance correlation coefficient (OCCC),
- Reliability coefficient,
- Pearson correlation coefficient, and
- Bland-Altman plots.

2.3.3.1 Intraclass correlation

Of these methods, the ICC is the most common and easy to interpret. It is a measure of both reliability and agreement that is often used when comparing scientific instruments and can be used when three or more groups are to be compared. For example, if a participant is measured using different methods or tools or measured more than two times, then the ICC will provide a measure of how consistent the extracted measures are replicated across the three scans and how well the methods compare. A Pearson correlation or Reliability coefficient is limited as they only provide a measure of reliability and does not include a measure of agreement between measurements. Like the Pearson and Reliability coefficients the ICC provides output varies between 0 and 1, with a value of 1 indicating perfect correlation between measures. ICC is difficult to calculate by hand, but it is readily available in most statistics software packages. Annex A provides and example of how to calculate ICC using SPSS.

2.3.3.2 Systematic bias

Systematic bias is observed when manual and scan-based results show good reliability but poor agreement between measures. For example, many scan-based measures have been shown to be slightly larger than equivalent manual measures due to compression of soft tissue or body hair during manual measurements or differences in measurement path between a measuring tape and a digital equivalent (Koepke et al., 2017; Tiwari & Anand, 2021). Other sources of systematic error can be due to calibration errors or methodological errors (e.g., diurnal error due to taking scan measures in the morning and manual measures in the afternoon). Along with verifying the collected data for outliers, skewness or homoscedasticity, an evaluation of systematic bias is important to determine if this bias may affect measures of accuracy. If a bias is noted between manual and scan measures, subtracting (or adding if the scan measures are smaller) the bias from the scan data will bring the two measures into better agreement (Figure 3).

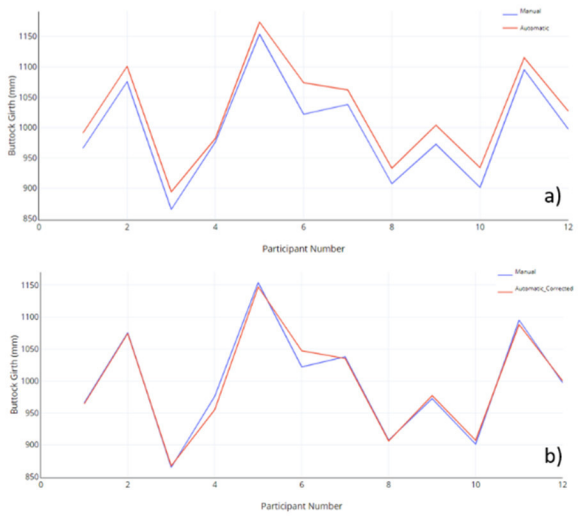

Figure 3: *Example of systematic error in buttock girth scan measurement (a) and result of removing the error (b).*

Several statistical methods are used to identify systematic error (bias) in comparisons of manual and scan extracted data. These methods include:

- Paired T-tests,
- Bland-Altman analysis, and
- Effect Size—e.g., Cohen's d.

Of these tests, the Bland-Altman is recommended for the body scanning evaluation as it provides a clear indication of bias while preserving the units of measurements. Cohen's d can be used as a measure of magnitude of the effect size, or practical significance of the relationship between the compared groups. The formula to calculate Cohen's d, along with its interpretation is provided in Annex A.

2.3.3.3 Bland-Altman analysis

Bland-Altman analysis is commonly used to evaluate the agreement between two measuring systems. Typically, one of the systems is the reference or "gold standard" method of measuring and the differences between the "test" and standard measures and are represented across the range of measurement values. Bland-Altman plots are particularly useful for illustrating systematic differences (bias) between the measurement systems and identifying outliers. The ability to show data bias is helpful in explaining the poor validity of a system even if the reliability is high. It is well documented that 3D body scan measures, particularly circumferences, tend to be slightly larger than manual measures due to noise in the scan model or the inability to compress body hair. If the difference between manual and scan measures is consistent, the bias can be subtracted from the larger measure to improve agreement between the two measuring methods.

2.4 Recommended test protocols for Canadian Armed Forces Anthropometric Program for Soldier System Acquisition test evaluation

Based on the analysis of accuracy and reliability of 3D body scanning systems provided in Section 2.2 and 2.3, three evaluation methods are recommended for assessing candidate 3D body scanning systems as part the CAPSSA bid evaluation process. These methods include the use of a test sphere, evaluation using a manikin or dressing form and evaluation using human participants. The first two methods provide a high degree of control of the test object while testing with human participants evaluate the robustness and operational effectiveness of the scanning system when challenged by the variability of the human form.

2.4.1 Use of test sphere—points rated

As mentioned in Section 2.2.1 above, verification of the absolute accuracy of the 3D scanning system may be accomplished through the evaluation of a calibrated test sphere in accordance with ISO 20865-2.

It is recommended that CAPSSA request bidders to provide a certificate verifying that the scan system can achieve ISO 20685 minimum allowable error standards for small circumferences using a standardized test object. Ideally, this object would be a sphere tested in accordance with ISO 20685-2 methodology. If another object such as a manikin was used, the certificate should indicate which body measures were evaluated and the measuring error provided. Measurements should include a mix of linear and circumferential measures across the top, middle and lower portion of the scan volume.

As not all vendors may be able to provide this certificate of accuracy, it is recommended that this be a rated criterion. For example:

a. Full points awarded if ISO 20685 or equivalent certificate of measurement accuracy using a standardized test device is provided.

b. No points awarded if a certificate of measurement accuracy cannot be provided.

2.4.2 Use of manikin test device—points rated or ranked

An independent evaluation of the test device should be performed using an anthropomorphically accurate manikin or dressing form to verify the validity and reliability of the body scanning system under controlled conditions. Two manikins should be evaluated, representing the mean Canadian Armed Forces male and female body size and shape as represented by 2012 CFAS. For example, a life-size 3D printed manikin of the average CFAS male, combat arms provided to Med-Eng Inc. (Ottawa) (Figure 4). An advantage of this manikin is that it is in the ISO "A Pose" and does not have articulations at the joints which may affect landmark detection. If possible, a small and large manikin for each sex should be included in this evaluation. If representative female or large and small manikins are not available, the test can be done with the single male manikin with the understanding that the robustness of the evaluation will be reduced.

Figure 4: *3D printed whole-body manikin representing the mean CFAS male, combat arms.*

2.4.2.1 Selection of body measures

Most 3D body scanning systems claim to be capable of measuring upwards of two hundred body dimensions. Unfortunately, data reflecting the accuracy and reliability of these measures is not typically available. Information available in published literature is rapidly dated as manufactures frequently update their software algorithms. Additionally, many measures are difficult to measure accurately due to occlusion in the crotch, effect of hair on head measurements, posture and body features (e.g., under bust for large or small breasts, obesity etc.). For these reasons, it is not reasonable to expect a 3D scanning system to effectively measure all body dimensions to the same level of standard. However, it is important that DSSPM identifies a system that can accurately obtain select body dimensions that are critical for operational requirements.

Table 1 is a list of recommended ISO 8559 (ISO, 1989) measures that are important to garment and personal equipment sizing and fitting. They include both upper and lower body measures and are comprised of linear, circumferential, and surface measures. It is recommended that any evaluation of a 3D body scanning system include these measures as a minimum.

Table 1: *Table of recommended measures for scan evaluation.*

ISO 8559 Designation	Measure Name
2.1.11	Waist girth (natural)
2.1.8	Bust girth
2.1.12	Hip girth
2.1.3	Neck-base girth
2.2.1	Height
2.2.27	Inside leg length; Crotch height
2.2.10	Back waist length (cervical to waist)

2.4.2.2 Recommended measuring protocol

- Each participant should be landmarked by a skilled anthropometrist who is familiar with the required body landmarks, measuring tools and technique.
- Following this, each of the body dimensions should be measured the process repeated a minimum of 10 times.
- After each measure, it is recommended to place the manikin in slightly different position within the scan volume to represent the randomness of the human standing position.

- Figure 5 provides an example of recommended placement at the center and ± 5 cm fore/aft and left/right of the center.
- It is suggested that the manikin be measured immediately after each scan and then moved to the next position. After each of the five positions have been scanned/measured, the process should be repeated for a total of 10 scan and 10 manual measurement pairs.

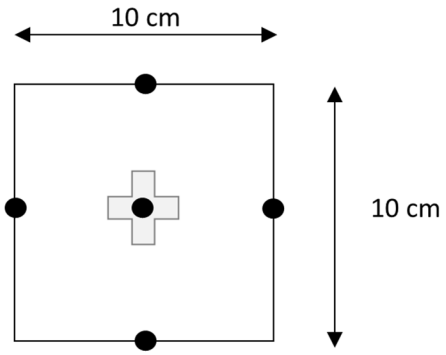

Figure 5: *Recommended placement of manikin with the scan volume (black dots).*

2.4.2.3 Recommended data analysis

1. Transcribe all scan and manual measures to a spreadsheet.

2. Visually inspect the data for any outliers by creating a regression plot of paired manual and scanned measures for each body dimension Figure 6. Examine the studentized residuals for any data points greater than ± 3 standard deviations. Recalculate the regression if any point is deleted as an outlier and reinspect the studentized residuals.

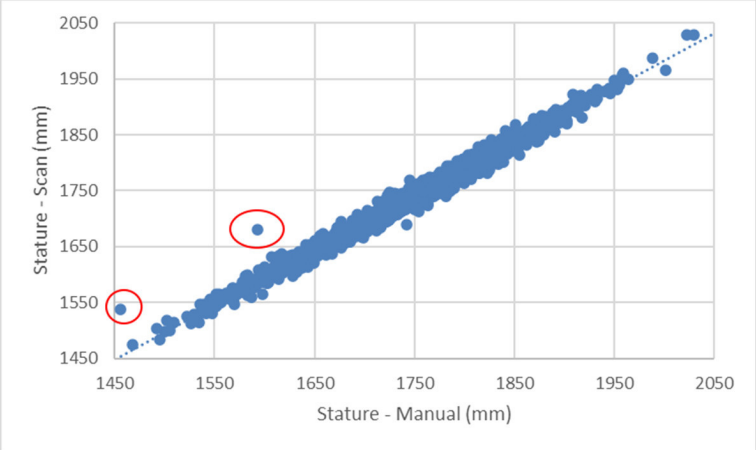

Figure 6: *Bivariate plot of manual and scan extracted stature. Possible outliers indicated in red circles.*

3. Conduct a Bland-Altman analysis comparing the manual and scan measures to determine if there is any bias in the data and verify that the slope of the differences is equal, or close to zero.

4. Calculate Cohen's d to determine the effect size between the two groups.

5. If there is detected bias or if the effect size is > 0.2 then create an "adjusted scan data" column by adding/subtracting the bias from the scan data and redo the analysis.

- Note: Correction for bias will increase the agreement between the manual and scanned data. However, this is only useful if the adjustment is incorporated in the measurement outputs of the scanning system. A question remains as to whether the accuracy of the scan system should be scored as it is presented by the system or with bias adjustment.

6. To evaluate the reliability and agreement of the measurer and the scanning system, calculate the ICC using the two-way mixed effects, absolute agreement (see Annex A). A minimum ICC of .90 is acceptable.

- Note: It is beneficial to have two manual measurers to provide a method to evaluate inter-measurer reliability as well as measurer/scanner reliability.

7. Calculate descriptive statistics (e.g., mean, standard deviation) and %TEM, SEM and MAD ± 95% confidence interval, using the difference between scan and manual measurement pairs for each body dimension in accordance with the instructions in Annex A.

8. Criterion measures for rating/ranking for each body measure, from most to least stringent:

- Level 1—Most stringent criterion
    - MAD within ISO 20685 allowable standards.
- Level 2—Acceptable criterion
    - %TEM is within the 1.5% ISAK standard.
- Level 3—Unacceptable criterion
    - Reject if the SEM is > 10 mm as per Vonk and Daanen.

As the evaluation of a manikin or dress form eliminates much of the random error it is expected that most 3D body scanning systems will be able to meet the Level 1 or Level 2 criterion. If this is not the case, then the measurement error may be due to landmarking or features of the manikin that are not consistent with the requirements of the scan hardware or automated landmarking and measurement software (e.g., pose or surface anomalies). It is therefore important to evaluate the manikin and scanning system by inspecting the landmarks identified by several systems to identify if any measurement errors are consistent across vendors.

2.4.3 Assessment using human participants

The most important evaluation of the accuracy and reliability of a body scanning system is to conduct a comparative analysis using human participants. This evaluation will reveal the robustness of the performance of the system across a range of human body characteristics including postural sway, skin colour, body hair and different body shapes. It is expected that this evaluation will not yield the levels of performance determined using a manikin, but it will provide a measure of the system's capability in an operational setting.

2.4.3.1 Training of measurers

As evaluation of the body scanning systems is referenced to manual measurements it is important that the measurers are trained to a high level of proficiency to minimize inter- and intra-measurer error. It is recommended that measurers are trained on the measurement to be evaluated according to the skilful anthropometric standards of ISAK which are 1.0% TEM for intra-measurer error and 1.5% inter-measurer error (Perini et al., 2005). If this is not possible due to limitations of training time and measurer experience, the beginner ISAK anthropometrist standards of 1.5% TEM for intra-measurer error and 2.0% inter-measurer error should be acceptable. An Excel-based proforma and support materials to guide training and calculation of %TEM can be provided by DRDC – Toronto Research Centre.

2.4.3.2 Selection of measurements

At a minimum, the list of body measures assessed in the human trials should be the same as those indicated as Section 2.4.2.1. Additional measures may be included if desired, but care must be taken to ensure that the method and definition of measurement is consistent with that identified by the system documentation. Typically, 3D body scanning systems comply with ISO 7250-1 (ISO, 2008) or ISO 8559, however, certain systems may use a different standard or their own measurement scheme.

2.4.3.3 Selection of study participants

Sufficient participants should be evaluated to provide the necessary statistical power to determine true differences between the measurement methods. Annex B provides a methodology for calculating the minimum sample size to detect a predetermined difference between means based on ISO 20685 allowable maximum difference. ISO 20685-2 recommends a minimum of 40 participants.

Participants should be distributed evenly by sex (e.g., 20 men and 20 women) and represent the range of body size and shapes found in the CAF. Figure 7 provides a bivariate plot of stature and weight of CFAS men and women that can be used to guide the selection of male and female study participants.

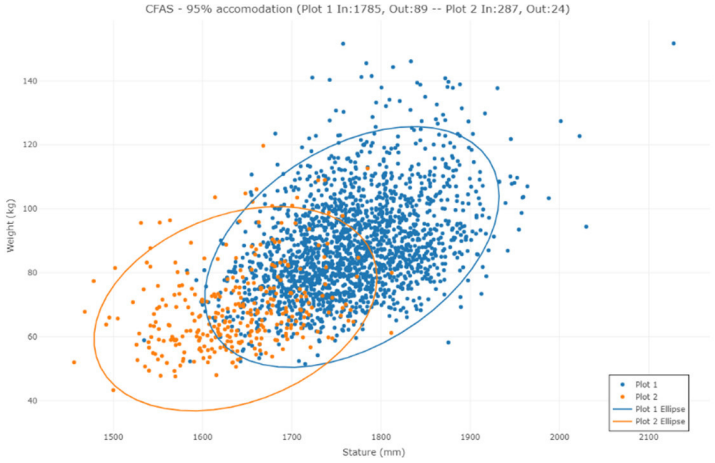

Figure 7: *Bivariate plot with 95% confidence ellipses of stature and weight of CFAS men and women.*

2.4.3.4 Scan clothing

Scan clothing should be standardized to consist of form fitting shorts for both men and women. For best results, as recommended by ISO 20685, these clothing should be form-fitting but not compress the tissue. Lycra or Lycra-blend shorts/briefs for men and women work well. As bra type can affect women's bust measurement, a sport bra or crop top is recommended for women as it supports the breast sufficiently to facilitate under bust measurement and is a common undergarment worn by women under protective armour (Coltman et al., 2021). A Lycra swim cap should be fitted to compress hair. Women's hair should be in a bun placed no higher than the vertex of the head. Ideally, this bun should be placed outside of the lycra cap. This can be facilitated by making a small hole in the lycra cap, allowing the hair to be pulled through as per ISO 20685 (Figure 8). Prior to scanning, any wrinkles in the clothing should be smoothed out and shorts raised so that they are snug around the crotch. The omphalion should not be covered as it is a landmark often used by automated landmarking software and other landmarks may be referenced to this point.

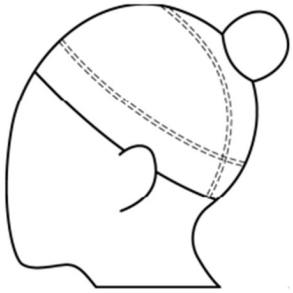

Figure 8: *ISO 20685 recommended configuration of a cap for managing long hair. Note, the hair bun is encapsulated by a small hair net.*

Clothing should be of a light colour to facilitate reflectance by light or laser-based scanning systems.

2.4.3.5 Recommended measuring protocol

Each participant should be landmarked by a skilled anthropometrist who is familiar with the required body landmarks, measuring tools and technique. A hypoallergenic eyebrow pencil may be used to mark the landmarks as it is easily cleaned with soap and water. Sharpen the pencil between uses to ensure a clean pencil between participants.

Following this, each of the body dimensions should be measured three times. Repeated measurements of the same body measure should not be taken in sequence. Instead, divide all measures into logical groups based on measurement type and tool (e.g., circumferences, breadths). Once all measurements of a group have been completed, allow the participant to relax followed by reposing. Return to the first measure and repeat the sequence to obtain the second measurement. Repeat this sequence for the third measure.

An alternate method to obtain the three manual measures is to measure immediately before or after a body scan. This can be done in a counterbalanced way according to the following sequence: participant 1—scan, manual, scan, manual, scan, manual and participant 2—manual, scan, manual, scan, manual, scan.

Note that proper posturing is critical to obtain accurate and reliable scan measurements. When scanning in the ISO “A” pose, be certain that the shoulder area is relaxed and that the arms are not raised more than 20° to avoid movement in the shoulder girdle. Feet should be standardized at 20 cm apart and the head placed in the Frankfort plane. During manual measurements, pay attention to the participants breathing cycle. ISO 7250 recommends measurement be taken during normal, quiet respiration.

Each participant should be scanned for measurement three times in the ISO standing “A” pose. Between each scan, the participant should leave the scan volume and then return to be repositioned for subsequent scans.

Recommended data analysis.

1. Transcribe all scan and manual measures to a spreadsheet.

2. As reliability was established during the manikin test, the repeated measures data can be collapsed into a single measure for manual and scan measurements for each measure according to the following procedure. If the difference in between the first two measurements is within the ISO 20685 allowable measurement error, average the two measurements. If a third measurement is required, take the median of the three measurements as the final value.

3. Visually inspect the data for any outliers by creating a regression plot of paired manual and scanned measures for each body dimension. Examine the studentized residuals for any data points greater than ± 3 standard deviations. Recalculate the regression if any point is deleted as an outlier and reinspect the studentized residuals.

4. Conduct a Bland-Altman analysis comparing the manual and scan measures to determine if there is any bias in the data and verify that the slope of the differences is equal, or close to zero.

5. Calculate Cohen's d to determine the effect size between the two groups.

6. If there is detected bias or if the effect size is > 0.2 then create an "adjusted scan data" column by adding/subtracting the bias from the scan data and redo the analysis.

   - Note: Correction for bias will increase the agreement between the manual and scanned data. However, this is only useful if the adjustment can be incorporated in the measurement output of the scanning system.

7. To evaluate the reliability and agreement of the measurer and the scanning system, *calculate the intraclass correlation coefficient* using the two-way mixed effects, absolute agreement (see Annex A). Since the measurements are collapsed from three repeated measures, refer to the ICC in the "Averaged Measures" row. A minimum ICC of 0.90 is acceptable.

   - Note: It is beneficial to have two manual measurers to provide a method to evaluate inter-measurer reliability as well as measurer/scanner reliability.

8. Calculate descriptive statistics (e.g., mean, standard deviation) and %TEM, SEM and MAD ± 95% confidence intervals (CI), using the difference between scan and manual measurement pairs for each body dimension in accordance with the instructions in Annex B.

   - Note: Normally, the 95% confidence intervals are used when comparing differences between manual and scan measures against a standard. For example, if the measured mean difference between a manual and scan measure of stature is 2.8 mm ± 1.7, where 1.7 is the 95th confidence interval, this gives a range of differences of 1.1 mm to 4.5 mm. Since the maximum allowable error for stature is 4 mm, the system would not be compliant with this standard. A less stringent criterion would be to accept the mean value of 2.8 mm, making the system compliant with the ISO standard.

9. Criterion measures for rating/ranking for each body measure:

   - Level 1—Most stringent criterion
     - MAD within ISO 20685:2010 standards.
   - Level 2—Acceptable criterion
     - %TEM is within the 1.5% ISAK standard.
   - Level 2—Unacceptable criterion
     - Reject if the SEM is > 10 mm as per Vonk and Daanen.

3 Conclusion

This Document provides recommendations on methods to quantify the accuracy and reliability of bidder provided 3D body scanning systems in support of the CAPSSA program. These methods involve increasing complexity to evaluate the scanning systems from a highly controlled to an operationally simulated environment. These methods include:

1. Evaluating measurement accuracy using a calibrated test sphere.

2. Evaluating system accuracy and precision using a test manikin.

3. Evaluating system accuracy using human test participants.

A review of literature has identified test methods to identify outliers and bias in measurements (Bland-Altman and Cohen's d), as well as a method to quantify measurer and scan reliability (ICC). The minimum acceptable standard for reliability is an ICC value of 0.90.

For system accuracy, three criterion measures have been identified which are associated with published standards. Standards range from the most stringent (MAD), based on international anthropometric survey standards, to practitioner level standard (%TEM) to garment sizing standards (SEM).

It is anticipated that the methods described in this Document will provide the Human Factors Support Cell of DSSPM with sufficient knowledge to conduct an objective comparative analysis of 3D body scanning systems to identify a suitable system that will meet the needs of the CAPSSA program.

References

Bradtmiller, B., and Gross, M.E., (1999). 3D whole body scans: Measurement extraction software validation. In *SAE Technical Papers*, vol. 1999-01–18, USA. https://doi.org/10.4271/1999-01-1892.

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

Bragança, S., Arezes, P., Carvalho, M., Ashdown, S.P., Xu, B., and Castellucci, I., (2017). Validation study of a Kinect based body imaging system. *Work*, *57*(1), pp. 9–21. https://doi.org/10.3233/WOR-172532.

Chamberland, A., Carrier, R., Forest, F., and Hachez, G., (1997). Anthropometric Survey of the Land Forces.

Coltman, C.E., Brisbine, B.R., and Steele, J.R., (2021). Bra-body armour integration, breast discomfort and breast injury associated with wearing body armour. *Ergonomics*, *64*(12), pp. 1623–1633. https://doi.org/10.1080/00140139.2021.1951849.

Daanen, H.A.M, Woering, A., Ter Haar, F.B., Kuijpers, A.A.M., Haker, J.F., and Reulink, H.G.B., (2014). Optimization of Military Garment Fit. Ambience 2014, 7–9 September 2014, Tampere, Finland.

Daanen, H.A.M., and Ter Haar, F.B., (2013). 3D whole body scanners revisited. *Displays*, *34*(4), pp. 270–275. https://doi.org/10.1016/j.displa.2013.08.011.

Gordon, C.C., Blackwell, C.L., Bradtmiller, B., Parham, J.L., Barrientos, P., Paquette, S.P., and Kristensen, S., (2014, December). 2012 Anthropometric Survey Of U.S. Army Personnel: Methods And Summary Statistics. NATICK/TR-15/007.

Han, H., Nam, Y., and 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), pp. 530–540. https://doi.org/10.1016/j.ergon.2010.06.002.

ISO, (1989). ISO 8559 Garment Construction. International Standards Organization.

ISO, (2008). ISO 7250-1: Basic human body measurements for technological design—Part 1: Body measurement definitions and landmarks. International Standards Organization.

ISO, (2010). ISO 20685. International Standards Organization, (2nd edition).

ISO, (2015). ISO 20685 3-D scanning methodologies for internationally compatible anthropometric databases—Part 2: Evaluation protocol of surface shape and repeatability of relative landmark positions, vol. 2015.

Jones, P.R.M., West, G.M., Harris, D.H., and Read, J.B., (1989). The loughborough anthropometric shadow scanner (LASS). *Endeavour*, *13*(4), pp. 162–168. https://doi.org/10.1016/S0160-9327(89)80004-3.

Keefe, A., Angel, H., and Mangan, B., (2015). 2012 Canadian Forces Anthropometric Survey. Defence Research and Development Canada, Scientific Report, DRDC-RDDC-2015-R186.

Koepke, N., Zwahlen, M., Wells, J.C., Bender, N., Henneberg, M., Rühli, F.J., and Staub, K., (2017). Comparison of 3D laser-based photonic scans and manual anthropometric measurements of body size and shape in a validation study of 123 young Swiss men. *PeerJ*, *2017*(2). https://doi.org/10.7717/peerj.2980.

Koo, T.K., and Li, M.Y., (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), pp. 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Kouchi, M., Masaaki, M., Tsuzuki, K., and Tokoi, T., (1999). Interobserver errors in anthropometry. *Journal of Human Ergonomics*, *28*, pp. 15–24.

Kouchi, M., and Mochimaru, M., (2011). Errors in landmarking and the evaluation of the accuracy of traditional and 3D anthropometry. *Applied Ergonomics*, *42*(3), pp. 518–527. https://doi.org/10.1016/j.apergo.2010.09.011.

Kouchi, M., Mochimaru, M., Bradtmiller, B., Daanen, H., Li, P., Nacher, B., and Nam, Y., (2012). A protocol for evaluating the accuracy of 3D body scanners. *Work*, *41*(SUPPL.1), pp. 4010–4017. https://doi.org/10.3233/WOR-2012-0064-4010.

Kuehnapfel, A., Ahnert, P., Loeffler, M., Broda, A., and Scholz, M., (2016). Reliability of 3D laser-based anthropometry and comparison with classical anthropometry. *Scientific Reports*, *6*(May), pp. 1–11. https://doi.org/10.1038/srep26672.

Li, X., and Li, X., (2010). Human body dimensions extraction from 3D scan data. *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, *2*, pp. 441–444. https://doi.org/10.1109/ICICTA.2010.849.

Lu, J.M., and Wang, M.J.J., (2010). The evaluation of scan-derived anthropometric measurements. *IEEE Transactions on Instrumentation and Measurement*, *59*(8), pp. 2048–2054. https://doi.org/10.1109/TIM.2009.2031847.

Meunier, P., and Yin, S., (1999). Measurement Accuracy and Precision Assessment of an Image-Based Clothing and Equipment Sizing system, (511502), pp. 1–40.

Meunier, P., and Yin, S., (2000). Performance of a 2D image-based anthropometric measurement and clothing sizing. *Applied Ergonomics*, *1*(31), pp. 445–451.

NATO, (2020). Sizing and fitting of military combat clothing, individual equipment and protection, vol. AEP-4833.

Norton, K.I., (2019). Standards for Anthropometry Assessment. In *Kinanthropometry and Exercise Physiology,* pp. 68–137. Taylor and Francis. https://doi.org/10.4324/9781315385662-4.

Perini, T.A, de Oliveira, G.L., Ornelia, J.S., and de Oliveira, F.P., (2005). Technical error of measurement in anthropometry. *Revista Brasileira de Medicina Do Esporte*, *11*, pp. 81–85. https://doi.org/http://dx.doi.org/10.1590/S1517-86922005000100009.

Pokorny, C., Seifert, E., Griffin, L., Holschuh, B., Juhnke, B., and Savvateev, E., (2019). Validation of the Artec Eva for Hand Anthropometric Data Collection. *Seifert, E*, *2017*, 2017–2020.

Rumbo-Rodríguez, L., Sánchez-Sansegundo, M., Ferrer-Cascales, R., García-D'urso, N., Hurtado-Sánchez, J.A., and Zaragoza-Martí, A., (2021). Comparison of body scanner and manual anthropometric measurements of body shape: A systematic review. *International Journal of Environmental Research and Public Health*, *18*(12). https://doi.org/10.3390/ijerph18126213.

Tiwari, M., and Anand, N., (2021). Validation and Reliability of Sizestream 3D Scanner for Human Body Measurement. *Functional Textiles and Clothing 2020*, pp. 13–23. https://doi.org/10.1007/978-981-15-9376-5_2.

Tomkinson, G.R., Daniell, N., Fulton, A., and Furnell, A., (2016). Time changes in the body dimensions of male Australian Army personnel between 1997 and 2012. *Applied Ergonomics*, *58*, pp. 18–24.

Vonk, T.E., and Daanen, H.A.M., (2015, October). Validity and Repeatability of the Sizestream 3D Scanner and Poikos Modelling System, pp. 293–299. https://doi.org/10.15221/15.293.

Annex A Statistical measures

This annex provides the necessary formulas and methods for computing the recommended statistics to evaluate the reliability and accuracy of 3D body scanning systems. Where recommendations to use statistical software packages are made (e.g., SPSS), guidance on how to perform the analysis and interpretation of the outputs is provided.

A.1 Measures of accuracy

A.1.1 Mean absolute difference

Mean absolute difference is the basis of ISO 20685 maximum allowable error and is calculated using the following formula. Allowable errors are specific to the type of measure and are provided in the ISO 20685 standard.

$$MAD = \frac{\sum|x_i - \bar{x}|}{n} \tag{A.1}$$

Where:

- xi is the data values in the set,
- $\bar{X}$ is the mean of the data set, and
- n is the total number of observations.

A.2 Technical error of measure and relative technical error of measure

Percent technical error of measurement is the standard method measuring of inter- and intra-measurer error recommended by the International Society for the Advancement of Kinesiology (ISAK). The ISAK standard for intra-measurer error is 1.5% and is the recommended value to use when comparing 3D body scanning and human measurer performance.

$$TEM = \sqrt{\frac{\sum(x_{i-\bar{X}})^2}{2n}} \tag{A.2}$$

Where:

- xi is the data values in the set,
- $\bar{X}$ is the mean of the data set, and
- n is the total number of observations.

$$\%TEM = \frac{TEM}{VAV} \tag{A.3}$$

Where:

- VAV is the average value of all measurements (manual and scanned).

The SEM is a metric recommended by Vonk & Daanen (2015). Note that it is important to calculate the intraclass correlation coefficient (ICC) to determine ICC.

A.3 Standard error of measurement

$$\text{SEM} = \text{SD}\sqrt{1\text{-ICC}} \tag{A.4}$$

Where:

- Standard deviation (SD) is the standard deviation of difference between measurements (manual and scanned), and
- ICC is the calculated intraclass coefficient.

A.4 95% Confidence intervals

Confidence intervals provide the range of estimates of the parameters described above. The confidence interval should be used when assessing system accuracy. For example, a MAD of 12 mm with a 95% confidence interval of ± 3 means that the true MAD lies between 9 and 15 mm.

$$CI_{95\%} = \bar{X} \pm 1.96 \times \frac{SD}{\sqrt{n}} \tag{A.5}$$

Where:

- $\bar{X}$ is the mean of the difference between measurements,
- SD is the standard deviation of the difference between measurements, and
- n is the number of observations.

A.5 Measures of reliability

A.5.1 Intraclass correlation coefficient

The intraclass coefficient is a measure of reliability that is often used when comparing scientific instruments and when three or more groups are to be compared. For example, if a participant is measured using different methods or tools or measured more than two times, then the ICC will provide a measure of how well the extracted measures are replicated across the three scans. If only two measures are obtained, then a paired t-test would be preferred. A Pearson correlation is unsuitable for reliability measurement as it only provides correlation and does not include a measure of agreement between measurements. Like the Pearson correlation, the ICC provides output that varies between 0 and 1, with a value of 1 indicating perfect correlation between measures. ICC is difficult to calculate manually but it is readily available in most statistics' software packages.

There are many forms of ICC, so it is important to select the appropriate ICC calculation according to the experimental design. Fortunately, Koo & Li (2016) provide guidance to help determine the appropriate test.

The following is a summary for proposed 3D body scanning testing based on the flowchart presented in Koo et al., with SPSS menu example in Figure A.1.

1. When comparing repeatability of a single body scanning system:
   - Assume a single measurement is conducted three times.
   - ICC—Two-way mixed effects, absolute agreement.
   - ICC output—*Single Measures output*.

For example, if everyone is scanned three times using a single scan system, the ICC would be based on three columns of data for each measure (chest_circ1, chest_circ2 and chest_circ3).

2. When comparing two or more methods of measuring (e.g., manual vs. scan, or Scanner A vs. Scanner B):
   - Assume all subjects are measured by the same systems.
   - Use a specific set of raters (e.g., scanners).
   - ICC—Two-way mixed effects, absolute agreement.
   - ICC output—If the average of multiple measures of each scanner are compared then use "Average Measures" output. Otherwise *use Single Measures output* (see Figure A.2).

Example: Data from the repeatability evaluation of each scanner can be averaged to provide a single data value and used to compare to the associated data obtained from manual measures or another scan system. If averaged data are used, then the "Average Measures" row of the ICC output should be consulted (Figure A.3).

Note that in SPSS, the ICC test is found under the menu item Analyze>Scale>Reliability Statistics. Once ICC is selected, dropdown boxes are enabled which will allow for selecting two-way mixed and absolute agreement (Figure A.1). Note that presentation of the ICC options may be different across statistics packages.

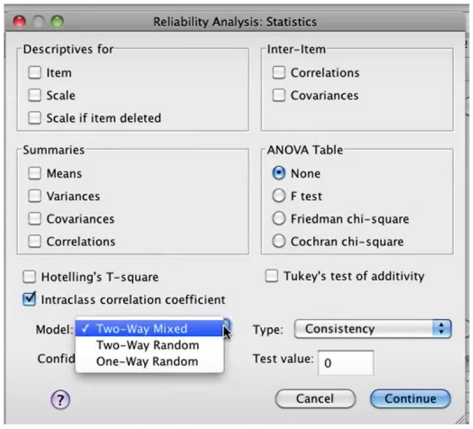

Figure A.1: *Intra class correlation selection in SPSS.*

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.738	.396	.951	12.250	5	18	.000
Average Measures	.918	.724	.987	12.250	5	18	.000

Figure A.2: *Example of ICC output in SPSS showing results for single and average measures.*

The ICC analysis recommended above agrees with Vonk & Daanen (2015) who employed a two-way mixed effects reliability analysis to compute the ICC when assessing repeatability of scan systems. Although it was not mentioned in their report, it is assumed that they used the Type: absolute agreement since both agreement and reliability were assessed.

When reporting ICC, it is important to consult the ICC and 95% confidence interval. These values can be interpreted as follows:

- Values less than 0.5 are indicative of poor reliability,
- Values between 0.5 and 0.75 indicate moderate reliability,
- Values between 0.75 and 0.9 indicate good reliability, and
- Values greater than 0.90 indicate excellent reliability.

In the example provided in Figure A.2, the Single Measures result indicates Poor to Excellent agreement and reliability while the Average Measures results indicates Good to Excellent agreement. A real-world example may be provided by comparing stature collected using manual and scan-extracted methods (Figure A.3). In this example, the upper and lower 95% confidence intervals are similar indicating a high certainty of the point estimate. The reliability of this comparison is rated as excellent.

Intraclass Correlation ▼

Type	Point Estimate	Lower 95% CI	Upper 95% CI
ICC3,1	0.994	0.993	0.994

Note. 2211 subjects and 2 judges/measurements. ICC type as referenced by Shrout & Fleiss (1979).

Figure A.3: *Intraclass coefficient results comparing CFAS manual, and scan extracted stature measurements using JASP v0.16 software.*

A.5.2 Bland-Altman analysis

Figure A.4 and Table A.1 demonstrate the results of a Bland-Altman analysis and plot of CFAS anthropometric data comparing stature obtained by manual and scan-extracted methods. This analysis was conducted using the "blandr" module in the freely available JAMOVI v1.2.27 software. From the plot in Figure A.4, there are two large outliers below the –50 mm Difference. Typically, these outliers would be inspected and removed from the dataset if necessary. This plot also shows the bias (middle dashed line), upper limit of agreement (upper dashed line) and lower limit of agreement (lower dashed line). Upper and lower limits of agreement are defined as the mean ± 1.96 x standard deviation. Numerical representations of these lines are provided in Table A.1.

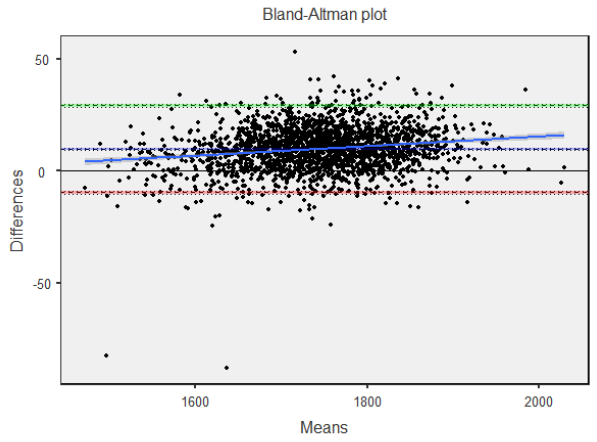

Figure A.4: *Bland-Altman plot of CFAS manual vs. scan extracted stature data, including proportionality bias line.*

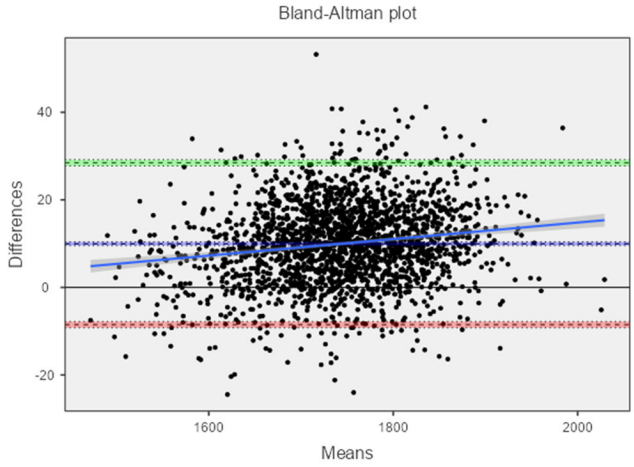

Figure A.5: *Bland-Altman plot of CFAS manual vs. scan extracted stature data with two extreme outliers removed.*

Table A.1: *Bland-Altman analysis of CFAS manual vs. scan extracted stature data.*

Bland-Altman

		95% Confidence Interval	
	Estimate	Lower	Upper
Bias (n = 2045)	9.97	9.56	10.38
Lower limit of agreement	−8.51	−9.21	−7.81
Upper limit of agreement	28.45	27.75	29.15

From the Figure A.5 and Table A.1 above, it is noted that there is a 9.97 mm bias between manual and scan measures with upper and lower limits of agreement approximately ± 20 mm about this value. As the manual measure was defined as the standard, this means that the scan measure tends to be consistently larger than manual measures by this amount. If desired, this bias can be subtracted from each scan extracted measure to bring it closer to the manual measure.

An important note is the solid line in Figure A.4 which is the proportional bias line. This line indicates the proportional bias across the range of data values. If there was zero slope, then the bias would be constant across the data range. In this example, there is a slight positive slope indicated a lower bias for shorter statures and larger bias for taller statures. Closer inspection of the Q-Q plot in Figure B.6 of differences indicates that the observed data conforms to the straight line indicating the bias is close to being normally distributed.

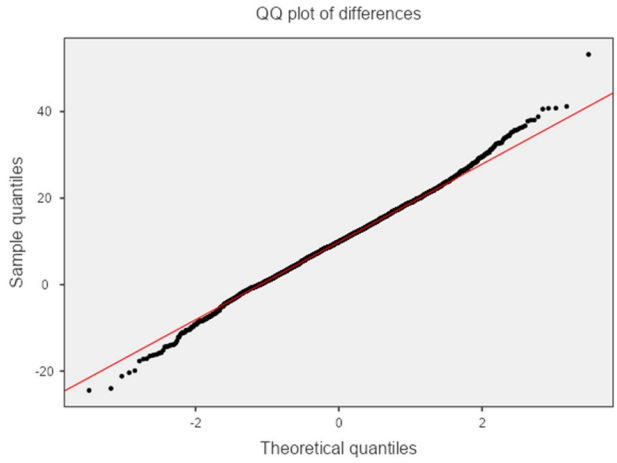

Figure A.6: *Q-Q plot of differences indicating the data is close to normal distribution.*

Most statistics software packages support the creation of Bland-Altman plots and analysis. They can also be easily calculated in Excel by following the instructions provided on the www.statology.org/bland-altman-plot-excel/ website (Accessed date: 22 January 2024).

A.5.3 Cohen's d

To evaluate the systematic error between scanning systems or manual and scan measurements, it is useful to calculate the standardized effect size. This is the ratio of the difference between group means divided by the pooled standard deviation of the two groups. One method to express this is Cohen's d which is calculated as follows:

$$Cohen's\ d = \frac{\bar{x}_s - \bar{x}_m}{s_{pooled}} \tag{A.6}$$

Where:

$$s_{pooled} = \sqrt{\frac{s_s^2 + s_m^2}{2}}$$

- $\bar{x}_s$ and $\bar{x}_m$ are the mean values for manual and scan extracted measures.
- s_s^2 and s_m^2 are the variances of manual and scan extracted measures.

Note that Cohen's d may be positive or negative, depending on which order the means of the measuring methods are calculated. For example, if $\bar{x}_1$ is smaller than $\bar{x}_2$ then Cohen's d will have a negative value. Cohen's d is interpreted in terms of Effect Size and is evaluated according the to the following Table A.2.

Table A.2: *Cohen's d interpretation.*

Effect size	*d*
Very small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.0

An effect size of ≤ 0.20 would be considered negligible meaning the two means are essentially equivalent. If it is larger, then it is possible that there may be a significant bias between the measurements which would affect the accuracy measure. This may be further investigated by using a Bland-Altman plot to identify any systematic bias between measurements as well as possible outliers.

Annex B Estimation of sample size

This annex provided detailed information on the calculation of sample size required for calculating sample size for evaluation of 3D body scanning systems using human participants.

B.1 Sample size requirements for comparing two means

$$n = \frac{s^2}{\varepsilon^2} \times \left(Z_{\left(\frac{\alpha}{2}\right)} - Z_{(\beta)} \right)^2 \quad \text{(B.1)}$$

Where:

n = total sample size of each group.

s = standard deviation of differences of measures between each group (e.g., manual vs. scan measures). This is difficult to estimate without a pilot study to provide test data, however there is sufficient literature available to provide good estimates of this value.

$Z_{\left(\frac{\alpha}{2}\right)}$ = critical Z-value at a chosen α (probability of Type I error) for a two tailed test. Typically, α is equal to .05 and Z = 1.96.

$Z_{(\beta)}$ = critical Z-value of the normal distribution at β (probability of making a Type II error or 1–α). For β, a one tailed value for 95% confidence is Z = 1.65.

ε = minimal detected difference between groups. These values can be derived from the maximal differences provided in ISO 20685 (see Table B.1).

Table B.1: *Table of ISO allowable errors for different body measures (ISO, 2010).*

Measurement type	Maximum mean difference (see 5.4) mm
Segment lengths (e.g. buttock-popliteal length)	5
Body heights (e.g. shoulder height)	4
Large circumferences (e.g. chest circumference)	9
Small circumferences (e.g. neck circumference)	4
Body breadths (e.g. biacromial breadth)	4
Body depths (e.g. chest depth)	5
Head dimensions without hair	1
Head dimensions with hair	2
Hand dimensions	1
Foot dimensions	2

For example, in a comparative study of manual and 3D body scanner measurements Han, Nam, & Choi, (2010) report the standard deviation of mean difference between the two methods for waist circumference as 20.2 mm Table B.2.

Table B.2: *Differences between manual and scan measurements (Han, Nam, & Choi, 2010).*

Statistical differences between manual measurements and scan measurements (unit: nun).

Measurements	Manual measurement (MAM)	Scan measurement (SM)	MAD	MD (SM-MAM) *t*-test	SD of MAD	SD of MD	CV of MD	Percent difference (SM-MAM)	Allowable error-ISO 20685	Repeatability	
										Manual method	Scan method
Stature	1550.0	1554.3	7.5	4.3*	8.1	8.9	2.1	0.277	4	0.55	1.34
Cervical height	1314.9	1308.6	8.2	−6.3**	5.5	7.4	−1.2	−0.480	4	0.55	0.96
Neck base circumference	369.7	369.4	16.7	3.2	11.9	20.2	6.3	−0.081	4	–	–
Chest circumference	852.5	891.3	39.0	38.8***	20.6	20.0	0.5	4.450	9	1.34	1.73
Bust circumference	888.2	904.0	19.7	15.8	15.6	17.7	1.1	1.763	9	0.71	0.24
Under-bust circumference	744.8	763.5	20.8	18.5***	17.1	19.3	1.0	2.480	9	0.84	0.45
Waist circumference	726.4	741.6	20.3	15.2***	15.3	20.2	1.3	2.071	9	1.1	0.08
Abdominal circumference	829.1	850.9	23.7	21.4***	18.3	20.6	1.0	2.595	9	1.48	0.66
Hip circumference	904.1	911.6	11.2	7.5***	10.3	10.5	1.4	0.826	9	0.55	0.30
Armscye circumference	362.6	340.4	23.3	−22.1***	13.6	15.5	−0.7	−6.316	4	0.55	1.04
Lateral shoulder length	378.9	390.0	13.1	11.1***	10.3	12.3	1.1	2.887	5	0.55	1.30
Ann length	525.4	503.1	22.4	−22.1***	10.1	9.8	−0.4	−4.336	5	0.55	1.58
Waist back length	371.7	373.7	6.2	2.0	5.0	7.6	3.8	0.537	5	–	–
Foot length	224.5	225.9	7.2	1.3***	5.0	8.5	6.5	0.622	2	–	–

*$p < .05$, **$p < .01$, ***$p < .001$.
Repeatability is from study of Park SM (2004).

Inserting the ISO maximum mean difference of 9 mm results in the following formula:

$$n = \frac{20.2^2}{9^2} \times (1.96 + 1.65)^2 = 65 \text{ participants} \quad \text{(B.2)}$$

Thus, the required sample size for comparing manual and scan measures for waist circumference is 65 participants. This process is repeated for each measure and the largest sample size requirement is selected. For example, using the table above, predicted sample sizes for statue, hip circumference and waist back length is 65, 18 and 8 respectively. Thus, a sample size of 65 participants would be sufficient for all measures.

Pragmatically, a sample size of 65 is large for a bid evaluation and meeting the ISO allowable measurement standards can be quite challenging for humans and automated systems to achieve. ISO 20685-2 suggests a sample size of at least 40 participants.

Review of relevant literature suggests that the ISO maximum mean difference can be difficult to achieve for most scanning systems, however, other reliability and accuracy measures are detailed in Annex A.

List of symbols/abbreviations/acronyms/initialisms

3D	three-dimensional
ADM(DRDC)	Assistant Deputy Minister (Defence Research and Development Canada)
ADM(Mat)	Assistant Deputy Minister (Materiel)
AOE	allowable observer error
CAF	Canadian Armed Forces
CAPSSA	CAF Anthropometric Program for Soldier System Acquisition
CCC	concordance correlation coefficient
CFAS	Canadian Forces Anthropometric Survey
CI	confidence interval
CV	coefficient of variation
DND	Department of National Defence
DRDC	Defence Research and Development Canada
DSSPM	Directorate Soldier System Project Management
HFSC	Human Factors Support Cell
ICC	intraclass correlation coefficient
IR	infrared radiation
ISAK	International Society for the Advancement of Kinanthropometry
ISO	International Standards Organization
LASS	Loughborough Anthropometric Shadow Scanner
MAD	mean absolute difference
MPC	Military Personnel Command
OCCC	overall concordance correlation coefficient
RMSE	root mean square error
SD	standard deviation
SEM	standard error of measurement
SSE	strong secure and engaged
TEM	technical error of measurement
US	United States
VAV	is the average value of all measurements (manual and scanned)

DOCUMENT CONTROL DATA

*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive

1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
Toronto, Ontario M3K 2C9
Canada

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

CAN UNCLASSIFIED

2b. CONTROLLED GOODS

NON-CONTROLLED GOODS
DMC A

3. TITLE (The document title and sub-title as indicated on the title page.)

Methodology to evaluate three-dimensional (3D) body scanning systems in support of the Canadian Armed Forces Anthropometric Program for Soldier System Acquisition (CAPSSA) program

4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used)

Keefe, A.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total pages, including Annexes, excluding DCD, covering and verso pages.)	6b. NO. OF REFS (Total references cited.)
April 2024	36	33

7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.)

Reference Document

8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
Toronto, Ontario M3K 2C9
Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
CVPE_015	

10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC-RDDC-2024-D010	

11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

Public release

11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)

12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.)

Anthropometry; 3D Body Scanning; Bid Evaluation; Procurement

13a. ABSTRACT (When available in the document, the French version of the abstract must be included here.)

The Department of National Defence (DND) and the Canadian Armed Forces (CAF) require current and representative anthropometric (body size and shape) data to design, specify, acquire and procure clothing, individual equipment, platforms and workstations to ensure operator safety and optimal performance of a diverse work force. To address this challenge, the Director Soldier System Program Management (DSSPM) has initiated the CAF Anthropometry for Program for Soldier System Acquisition program (CAPSSA), which endeavours to continuously gather, analyze and manage anthropometric data of CAF members from recruitment and periodically throughout their careers. An important aspect of this project is the identification and acquisition of multiple three-dimensional (3D) body scanning systems to acquire body dimensions of CAF personnel.

In response to a request from the Human Factors Support Cell (HFSC) of DSSPM, Defence Research and Development Canada (DRDC) – Toronto Research Centre identified objective methodologies to evaluate the reliability and accuracy of candidate 3D body scanning systems based on international and operational standards. This Reference Document describes three recommended methods, associated statistical analysis and performance standards for evaluating 3D body scanning systems. These methods include: 1) scanning a calibrated test sphere, 2) scanning a manikin/dress form and 3) scanning human participants. Advice on how the evaluation methods can be used for rejecting, accepting, or rating candidate 3D body scanning systems is also provided.

13b. Résumé (when available in the document, the French version of the abstract must be included here)

Le ministère de la Défense nationale (MDN) et les Forces armées canadiennes (FAC) ont besoin de données anthropométriques actuelles et représentatives (taille et forme du corps) pour concevoir, spécifier, acquérir et obtenir des vêtements, de l'équipement individuel, des plateformes et des postes de travail pour assurer la sécurité des opérateurs et le rendement optimal d'un effectif diversifié. Pour cela, le Directeur—Administration du programme de l'équipement du soldat (DAPES) a lancé le Programme anthropométrique des FAC pour l'acquisition du système du soldat des FAC (PAASSF), qui s'efforce de recueillir, d'analyser et de gérer continuellement les données anthropométriques des membres des FAC dès leur enrôlement et de manière périodique tout au long de leur carrière. Un aspect important de ce projet est l'identification et l'acquisition de plusieurs systèmes de balayage corporel tridimensionnel pour obtenir les dimensions physiques du personnel des FAC.

En réponse à une demande de la Cellule de soutien aux facteurs humains (CSFH) du DAPES, le Centre de recherche de Toronto de Recherche et développement pour la défense Canada (RDDC) a défini des méthodologies objectives pour évaluer la fiabilité et l'exactitude des systèmes de balayage corporel tridimensionnel potentiels en fonction des normes internationales et opérationnelles. Le présent rapport décrit trois méthodes recommandées, une analyse statistique qui s'y rattache et les normes de rendement pour évaluer les systèmes de balayage corporel tridimensionnel. Ces méthodes comprennent : 1) le balayage d'une sphère d'essai calibrée, 2) le balayage d'un mannequin ou d'une tenue et 3) le balayage de participants humains. Des conseils sur la manière dont les méthodes d'évaluation peuvent être utilisées pour rejeter, accepter ou noter les systèmes de balayage corporel tridimensionnel potentiels sont aussi offerts.