Diplomarbeit

Reproducibility of circumferential leg and knee joint flexion measurements and clinical course of recovery after total knee arthroplasty

eingereicht von

Dipl.-Ing. Daniela Hirzberger

Geb.Dat.: 25.09.1982

zur Erlangung des akademischen Grades

Doktor(in) der gesamten Heilkunde
(Dr. med. univ.)

an der

Medizinischen Universität Graz

ausgeführt an der

Universitätsklinik für Orthopädie und orthopädische Chirurgie

unter der Anleitung von

Ass.-Prof. Priv.-Doz. Dr. Mathias Glehr
Priv.-Doz. Dr. Patrick Sadoghi

Graz, am 14. November 2013

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am …… *Unterschrift*

Acknowledgements

Writing my diploma thesis would not have been possible without the assistance and support of many kind people around me, only some of whom I can give particular mention here.

First and foremost, I extend my sincerest gratitude to all participants and observers who have willingly shared their precious time. Kathi, Mathias, Michi, Rosaleen, Sandra, Anna, Angela, Claudia, Gerhard, Gottfried, Irene, Ivo, Johannes, Joseph, Margit, Marie-Therese, Markus, Oli, Patricia, Peter, Regina, Simone, Thomas, Wolfgang - this study would not have been possible without you.

I owe my deepest gratitude to my supervisor Ass.-Prof. Priv.-Doz. Dr. Mathias Glehr, who suggested this topic to me and supported me throughout the study with his continuous optimism and knowledge, whilst allowing me the room to work in my own way. I also express my warmest gratitude to my other supervisor Priv.-Doz. Dr. Patrick Sadoghi whose useful comments, remarks and engagement helped me through the learning process.

I am deeply grateful to Univ. Prof. Dr. Andreas Leithner, Head of the Department of Orthopaedics and Orthopaedic Surgery, for giving me the opportunity to conduct my study at his department and introducing me to scientific working and writing.

I wish to thank Univ.-Ass. Mag. Dr. Alexander Avian and Dr. Gerwin Bernhardt, who patiently answered all my questions concerning statistics. I also thank Dr. Lukas Holzer for his interest in my thesis and for his encouraging words time and again.

I sincerely thank all of the medical and nursing staff of the Department of Orthopaedics and Orthopaedic Surgery, whose daily routine I disturbed with my measurements time and again.

Finally, I am deeply grateful to my parents and my sister Sandra for their love and continuous support – both spiritually and materially.

Abstract

Background: Knee swelling after total knee arthroplasty (TKA) may lead to pain and loss of motion. In clinical practice, postsurgical swelling is recorded by means of circumferential measurements using tape measures, while the knee range of motion (ROM) may be assessed by goniometric measurements. The first objective of this study was to determine inter-observer and intra-observer reproducibility (in terms of reliability and agreement) of circumferential leg and knee flexion measurements. The second aim was to evaluate recovery of in-patients following TKA with respect to postoperative swelling, passive knee ROM and pain intensity, and to determine a possible influence of postsurgical swelling on passive ROM.

Methods: For reproducibility of circumferential and goniometric measurement, two observers examined 40 legs of 20 healthy adults. Circumferential measurements were obtained at three measurement sites (mid-patella, 7 cm proximal of mid-patella, 7 cm distal of mid-patella), using four different types of tape measures (standard tape measure, circumference tape measure, Gulick I tape measure, Gulick II plus tape measure). Knee flexion was measured with a short-arm universal goniometer in two standardised knee joint positions. Agreement was quantified by calculation of the smallest detectable difference (SDD), and reliability by means of the intraclass correlation coefficient ($ICC_{2,1}$).
Clinical assessment was undertaken in 29 patients undergoing TKA. This included lower limb girths at the three measurement sites mentioned above, passive knee ROM, and knee pain intensity using a Numerical Rating Scale (NRS). Pearson´s correlation coefficient was computed to establish a possible relationship between circumference change and passive ROM.

Results: In the circumference study, inter-tester agreement (SDD) ranged from 0.7 to 2.1 cm and intra-tester agreement from 0.9 to 1.7 cm. The circumference measurements were generally reliable ($ICC_{2,1} > 0.93$). Considering the different measurement sites, agreement was lowest at 7 cm proximal of mid-patella. Comparing the different tape measures used, the Gulick II plus tape measure showed the lowest level of agreement (SDD range, 0.8-2.1 cm). Agreement was highest for the circumference tape measure (SDD range, 0.7-1.2 cm).

In knee joint flexion measurements, the level of agreement (SDD) ranged from 5.9 to 9.0° for inter-observer and 7.1 to 8.1° for intra-observer comparisons. Reliability ($ICC_{2,1}$) ranged from 0.85 to 0.99.

Lower limb swelling occurred in all patients after TKA surgery. The circumference change was higher above the knee (mean 5.1 cm, range, 2.3-7.6 cm) than at mid-patella (mean 3.8 cm, range, 1.9-9.8 cm) and below the knee (mean 2.8 cm, range, 1.7-7.2 cm). Maximum swelling was reached on the third to fourth postoperative day. Passive ROM increased continuously after TKA. On sixth postoperative day, the mean passive ROM was 79.0° (range, 55-100°). Circumference change at the three measurement positions did not show any significant correlation with passive ROM on third and sixth postoperative day ($P \geq 0.1375$). Postsurgical pain intensity reported by the patients was highest preoperatively (mean NRS_{max} 7.0, range 4-9), which might be explained by the patient-controlled analgesia for 72 hours postoperatively. After surgery, pain intensity decreased continuously until dismissal day (mean NRS_{max} 3.0, range 1-9).

Conclusion: In circumferential measurements, the level of reproducibility differed substantially depending on the measuring position and tape measure used. With the circumference tape measure, differences in girth exceeding 1.2 cm can be considered a real change above measurement error. Measuring knee flexion with a short-arm universal goniometer, differences of less than 9° cannot be distinguished from measurement error. After TKA, swelling in the knee region was observed in all patients, but did not seem to influence passive ROM after TKA surgery.

Zusammenfassung

Einleitung: Nach der Implantation einer Knietotalendoprothese (KTEP) kann eine Schwellung im Bereich des Kniegelenks auftreten, die zu Schmerzen und einem verminderten Bewegungsumfang (range of motion/ROM) führen kann. Die postoperative Schwellung kann durch Umfangsmessungen mit einem Maßband erfasst und der Bewegungsumfang durch Messungen mit einem Goniometer beurteilt werden. Ein Ziel dieser Arbeit war, die Reproduzierbarkeit (im Sinne von Übereinstimmung und Reliabilität) von Umfangsmessungen und Messungen der Knieflexion zu bestimmen. Weiters sollte der klinische Verlauf nach KTEP-Operation in Bezug auf die postoperative Schwellung, den passiven ROM und die Schmerzintensität beurteilt werden.

Methoden: Zur Beurteilung der Reproduzierbarkeit von Umfangs- und Flexionsmessungen haben zwei Prüfer die Beine von 20 gesunden Probanden untersucht. Die Umfangsmessungen erfolgten mit vier verschiedenen Maßbändern (Standard-, Umfangs-, Gulick I-, Gulick II plus-Maßband) an drei Messpositionen im Bereich des Kniegelenks (Patella-Mitte, 7 cm proximal der Patella-Mitte und 7 cm distal der Patella-Mitte). Die Knieflexionsmessungen wurden mit einem Universal-Goniometer in zwei standardisierten Kniepositionen durchgeführt. Die Übereinstimmung wurde mittels kleinster erfassbarer Messwertdifferenz (smallest detectable difference/SDD) und die Reliabilität mittels Intraklassen-Korrelations-Koeffizienten (ICC) erfasst.
Zur Evaluierung des postoperativen Verlaufs wurden Umfangsmessungen an den oben genannten Messpositionen durchgeführt und der passive Bewegungsumfang mit einem Universal-Goniometer erfasst. Die Schmerzen im Bereich des Kniegelenks wurden mittels numerischer Bewertungsskala (Numerical Rating Scale/NRS) erhoben. Um eine mögliche Korrelation zwischen Umfangsänderung und passivem ROM festzustellen, wurde der Pearson´s Korrelationskoeffizient berechnet.

Ergebnisse: Die Interrater-Übereinstimmung (SDD) der Umfangsmessungen umfasste einen Bereich von 0.7-2.1 cm und die Interrater-Übereinstimmung einen Bereich von 0.9-1.7 cm. Die Umfangsmessungen waren generell zuverlässig (ICC>0.93). Die Reproduzierbarkeit zeigte eine Abhängigkeit von der Messposition und war 7 cm proximal der Patella-Mitte am niedrigsten. Von den getesteten Maßbändern zeigte das Gulick II plus-Maßband die geringste Übereinstimmung (SDD 0.8-2.1 cm). Die Übereinstimmung war für das Waegener Maßband am höchsten (SDD 0.7-1.2 cm).

Die Interrater-Übereinstimmung der Knieflexionsmessungen umfasste einen Bereich von 5.9- 9.0° und die Intrarater-Übereinstimmung einen Bereich von 7.1-8.1°. Die Reliabilität (ICC) reichte von 0.85-0.98.
Postoperative wurde bei allen KTEP-Patienten eine Schwellung im Kniebereich beobachtet, wobei die Umfangszunahme 7 cm proximal der Patella-Mitte am höchsten war und die maximale Schwellung zwischen dem dritten und vierten postoperativen Tag erreicht wurde. Der passive ROM nahm im Rahmen des stationären Aufenthaltes kontinuierlich zu und erreichte einen mittleren Wert von 79.0° (55-100°) am sechsten postoperativen Tag. Die Umfangsänderungen zeigten keine signifikante Korrelation mit dem passive ROM ($P \geq 0.1375$). Die anhand der NRS ermittelten Schmerzen waren präoperativ höher als postoperativ und nahmen postoperativ bis zum Tag der Entlassung kontinuierliche ab.

Diskussion: Die Reproduzierbarkeit von Umfangsmessungen mit Maßbändern ist von der Messposition und dem verwendeten Maßband abhängig. Unterschiede im Beinumfang über 1.2 cm können als tatsächliche Veränderung über dem Messfehlerbereich beurteilt werden, wenn die Messung mit dem Umfangs-Maßband durchgeführt wird. Wenn die Knieflexion mit einem Universalgoniometer gemessen wird, sind gemessene Unterschiede unter 9° nicht vom Messfehler zu unterscheiden. Postoperativ wurde bei allen Patienten eine Schwellung im Kniebereich beobachtet, die jedoch keinen Einfluss auf den passiven Bewegungsumfang nach TKA hatte.

Table of contents

List of abbreviations and glossary

ANOVA	Analysis of variance
B-A	Bland and Altman
BMI	Body mass index, unit: kg/m^2
CI	Confidence interval
CD	Critical difference
cm	Centimetre, 1 cm = 10^{-2} m
CR	Coefficients of repeatability
CV	Coefficient of variation
D	Difference
DP	Measurement site 7 cm distal of mid-patella
e.g.	Exempli gratia
et al.	Et alii/ et aliae/ et alia
g	Gram, = unit of weight (SI)
GI	Gulick I tape measure
GII	Gulick II plus tape measure
ICC	Intraclass correlation coefficient
in	Inch, 1 in = 2.54 cm
kg	Kilogram, 1 kg = 10^3 g
KROM	Knee range of motion
i.e.	Id est
l	Left leg
L	Length (of goniometer arms)
LOA	Limits of agreement
m	Meter, = unit of length (SI)
m^2	Square meter, = unit of area (SI)
mD	Mean difference
MP	Measurement site at mid-patella
n	Sample size
NRS	Numeric rating scale
O	Observer
oz	Ounce, non-metric unit of mass, 1 oz = 28,349523125 g
r	Right leg
ROM	Range of motion

PP	Measurement site 7 cm proximal of mid-patella
RP	Relative precision
S	Standard tape measure
SD	Standard Deviation
SDD	Smallest detectable difference
SEM	Standard error of measurement
TKA	Total knee arthroplasty
VAS	Visual analogue scale
VRS	Verbal rating scale
vs.	Versus
W	Waegener tape measure
°	Degree, = unit of measurement for angles

List of figures

List of tables

1 Introduction

1.1 Aim

Total knee arthroplasty (TKA) is currently the international standard of care for treating severe degenerative and rheumatic knee joint disease, as well as certain knee joint fractures [1]. Swelling of the involved lower limb after TKA due to intraarticular bleeding and inflammation of the periarticular tissues is common and can cause pain, extended bed-rest, and a delay in rehabilitation [2,3]. Thus, reduction in swelling and restoration of knee joint range of motion (ROM) are important components in the overall postoperative rehabilitation [4]. In order to monitor and guide rehabilitation, tape measures and universal goniometers are frequently used instruments in the assessment of swelling and ROM of the knee, respectively [4]. However, a basic prerequisite for any measurement to be of value is a sufficient degree of repeatability in order to allow an observer to recognize some change from normal [5]. Thus, the first aim of this study was to determine the reproducibility of circumference measurements in the knee region using four different types of tape measures. Besides that, the reproducibility of goniometric measurements using a standard goniometer with special attention to clinical application should be evaluated. In particular, the analysis aimed at answering the following questions

- How reproducible are girth/goniometric measurements?
- Does the measurement site/test position have an influence on reproducibility?
- Which tape measure is most accurate?
- Does tester experience matter?

The second aim of this study was to evaluate the course of in-patient recovery after total knee arthroplasty and to establish a possible relationship between postsurgical swelling and passive ROM. For this purpose, the course of rehabilitation for patients undergoing TKA at the Department of Orthopaedic Surgery, Medical University of Graz, was recorded by means of circumferential measurements in the knee region, measurements of passive ROM, and a pain intensity score (NRS).

1.2 Total knee arthroplasty

In the year 2011, the most frequent surgical procedures in Austria were cataract surgery, followed by dermatological operations, vaginal delivery and knee joint surgeries [6]. Among the ten most frequently performed surgical procedures, total knee arthroplasty (TKA) ranked in eighth place (Figure 1). Thus, TKA is a very commonly performed, major orthopaedic procedure.

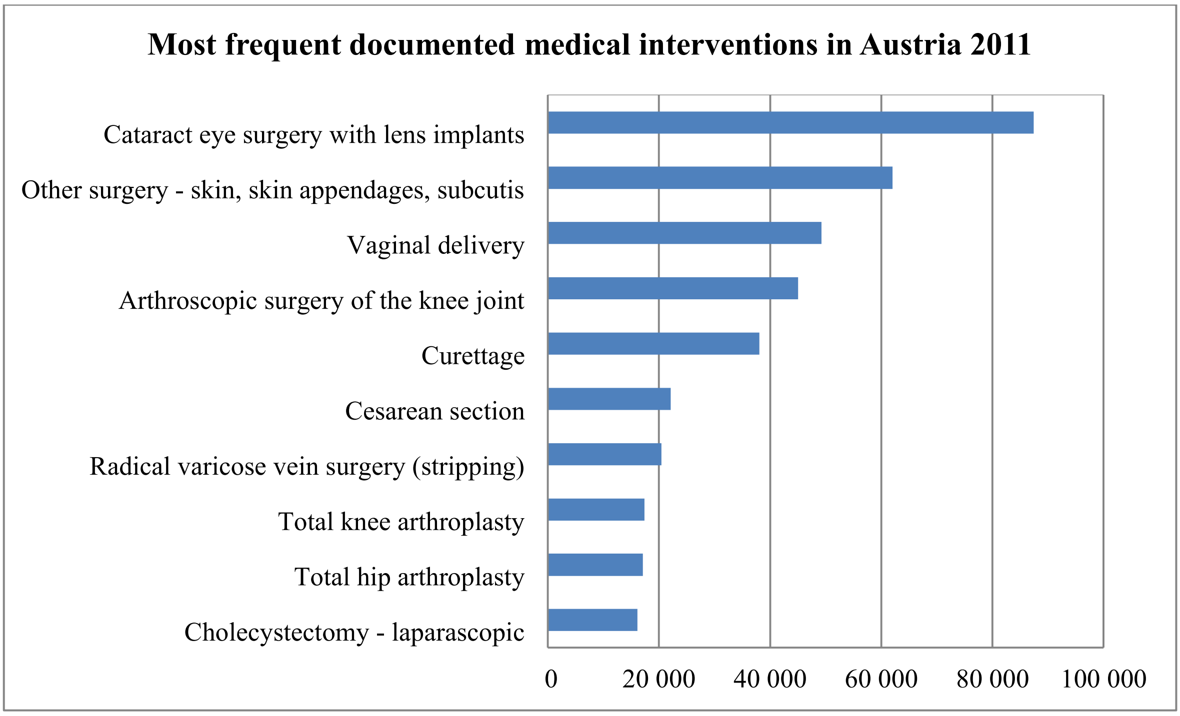


Figure 1: The 10 most frequent medical services in Austria in the year 2011
Source: STATISTIK AUSTRIA [6]

Over the past 10 years, there has been a considerable increase in the amount of primary arthroplastic knee joint replacement (Figure 2). Between 1999 and 2011, the incidence of total knee replacement surgery in Austria has more than doubled. At present, approximately 17 400 total knee joint arthroplasties are performed each year in Austria [6]. This might be due to the fact, that, knee replacement surgery is most commonly performed in people with advanced osteoarthritis. Osteoarthritis of the knee is one of the most common causes of disability. As the older adult and obese populations grow, osteoarthritis of the knee continues to increase in prevalence [7].

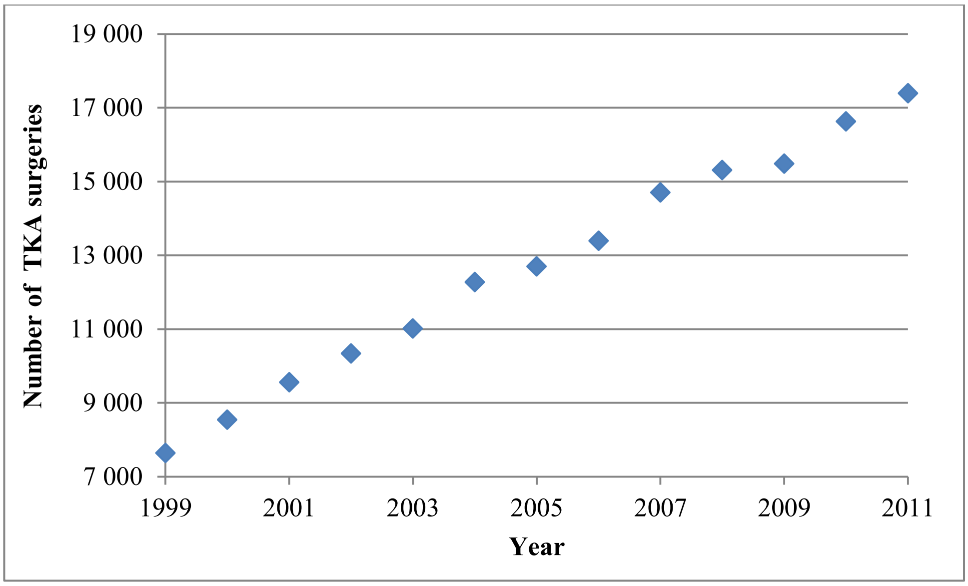


Figure 2: Number of TKA surgeries performed in Austria from 2004 to 2011
Source: STATISTIK AUSTRIA [6,8-18]

1.2.1 Indications and contraindications for TKA

The primary indication for primary TKA is pain caused by severe osteoarthritis with impairment of daily function, detoriating health-related quality of life and radiological signs of osteoarthritis [19,20]. Further, in patients with moderate arthritis and variable levels of pain, deformity can become the principal indication for arthroplasty when the progression of deformity begins to threaten the expected outcome of an anticipated arthroplasty. An occasional indication for arthroplasty in the absence of complete cartilage space loss is severe pain from chondrocalcinosis and pseudogout in an elderly patient. Rarely, arthroplasty may be justified in severe patellofemoral arthritis [20].

"The primary goals of joint arthroplasty in the treatment of arthritic disease are pain relief and restoration of function and health-related quality of life" [19].

Before surgery is considered, conservative treatment measures should be exhausted. Conservative, nonsurgical interventions for pain associated with osteoarthritis of the knee include non-steroidal anti-inflammatory drugs (NSAIDs), certain non-narcotic analgesics, and intra-articular injections with corticosteroids or viscosupplements [21]. Furthermore, surgical arthroscopy may be considered. Other important modalities in early management include physical training and weight loss [19,22].

TKA surgery should be considered in patients with persistent, moderate-to-severe pain associated with activity despite nonsurgical interventions that are medically fit and are willing to accept the risks associated with the operation. There should be radiographic evidence of significant joint damage. If there appears to be inconsistency between the

radiographic image and symptoms, other explanations for the patient's pain should be considered [21].
The patient's ability and willingness to participate in an aggressive regimen of postoperative physical therapy is an essential factor to take into account in considering surgery. For a good postsurgical result, vigorous physical rehabilitation, including exercises specifically intended to require early and repetitive motion of the affected knee despite substantial pain, is necessary. The outcome can permanently be prejudiced due to failures of rehabilitation, which often stem from problems in managing postoperative pain early on [21].
Contraindications to TKA include recent or current infection, a remote source of ongoing infection, extensor mechanism discontinuity or severe dysfunction, recurvatum deformity secondary to muscular weakness, untreated thrombophilias and bleeding disorders, severe vascular disease or neurologic disease affecting sensory or motor function in the affected leg, and inadequate soft tissue to cover the joint [20,21].

1.2.2 Technique

The TKA operation consists of removal of the damaged cartilage, correction of joint deformities, and replacement of the worn cartilaginous bearing surfaces on the femur, tibia, and, optionally, patella, with an artificial bearing [21].
The most widely accepted approach uses an anterior, medial parapatellar capsular, longitudinal incision. It is started at the medial one-third of the patellar tendon three centimetres above the superior pole of patella, curves round the medial aspect of patella with at least 1 cm of the medial retinaculum attached to the patella and ends 1 cm medial to the tibial tubercle [23]. The Hoffa´s fat pad is partly resected to improve sight and reduce volume [24]. Then, retractors are placed in a fixed position for maximal exposure and the tibiofemoral joint is dislocated. The patella is everted laterally, allowing exposure of the distal end of the femur, and the proximal end of the tibia and the knee hyperflexed. The cartilages and the anterior cruciate ligament are removed. The posterior cruciate ligament may also be removed but the tibial and fibular collateral ligaments are preserved. The ends of femur and tibia are then accurately cut to shape using cutting guides oriented to the long axis of the bones. Cutting jigs and anatomic landmarks are used to determine the depth and orientation of tibial and femoral bone resections. Precise resections are made in the distal end of the femur, the proximal end of the tibia, and, optionally, the posterior surface of the patella to fit the corresponding surfaces of the three arthroplasty components (Figure 3). A

round ended implant is used for the femur, mimicking the natural shape of the joint. The femoral component is typically made of metal (most commonly, a cobalt–chromium alloy). On the tibia the component is flat, although it sometimes has a stem which goes down inside the bone for further stability. The tibial implant is usually made of metal (either a titanium or a cobalt–chromium alloy). A flattened or slightly dished high density polyethylene surface is then inserted onto the tibial component so that the weight is transferred metal to plastic, not metal to metal. There is an exchangeable polyethylene bearing on the tibia, which makes it possible to replace the plastic articular surface without replacing the metal component if wear of the bearing surface occurs. During the operation any deformities must be corrected, and the ligaments balanced so that the knee has a good range of movement and is stable and aligned. Careful attention to ligament balancing and protecting neurovascular structures must be maintained. Trial implants then are placed over the resected bone surfaces; joint stability, ligament balance, and range of motion then are assessed [21]. If satisfactory, final components are inserted (Figure 3), hemostasis is obtained and the joint is irrigated and closed [21].

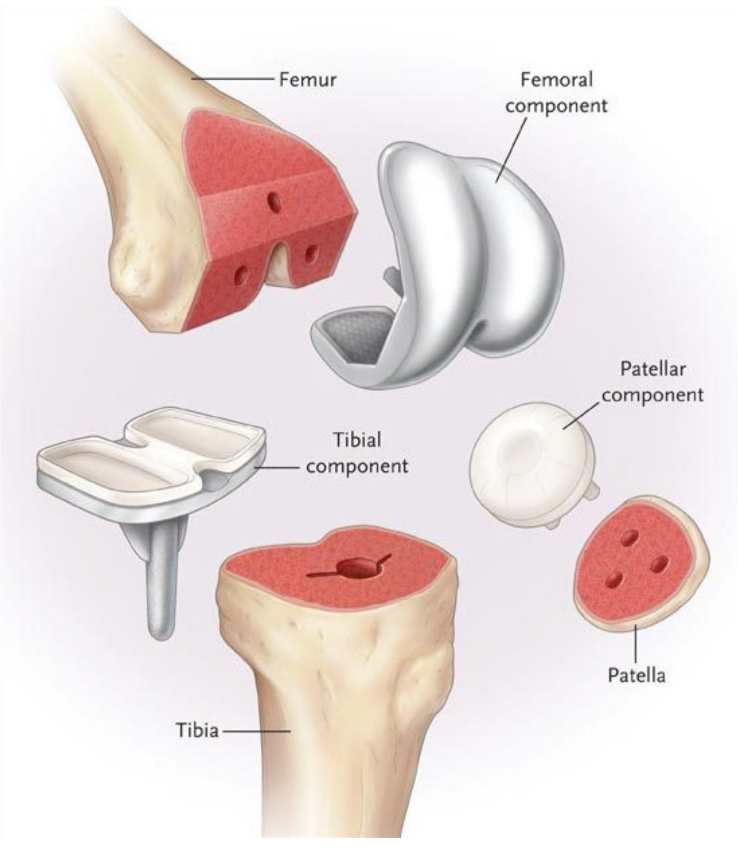


Figure 3: Scheme of total knee arthroplasty
Source: Leopold SS. N Engl J Med 2009 [21]

In some cases the articular surface of the patella is also removed and replaced by a polyethylene button cemented to the posterior surface of the patella.
In TKA surgery, surgical management and successful clinical outcomes rely on accurate soft tissue balancing, well positioned components and a neutrally aligned axis [22].

1.2.3 Risks and complications

As with other major surgery, complications may occur during and after TKA operation. These include general postoperative complications, infection, venous thromboembolism, stiffness, patellofemoral complications, peripheral nerve injury, vascular complications, ligament injury, periprosthetic fracture, knee joint dislocation and prosthesis failure. Obesity, increasing age, and medical comorbidities increase the risk of postoperative complications in patients undergoing TKA [25].

1.2.3.1 Infection

The most serious complication is infection of the joint, which occurs in 1-2% of patients within one year after surgery [21,25,26]. While it is a relatively infrequent, periprosthetic infection remains one of the most challenging complications of joint arthroplasty [27]. The underlying diagnosis leading to TKA seems to have an influence on the incidence of postoperative infections. Arthritic diseases other than osteoarthritis, such as posttraumatic osteoarthritis, seropositive rheumatoid arthritis, as well as fractures around the knee, showed increased rates of infections [28].
Such infections are likely to arise from bacterial contamination at the time of surgery [25]. Infection should be considered in patients with a consistently painful TKA and especially in patients with a previously pain-free arthroplasty. A history of swelling, erythema, or prolonged wound drainage is suggestive of infection, although these signs are not uniformly present.
Basic treatment options include antibiotic suppression, debridement with prosthesis retention, resection arthroplasty, knee arthrodesis, one-stage or two-stage reimplantation, and knee-above amputation as last option in the case of life-threatening infection or persistent local infection with massive bone loss [20,26]. Considering antibiotic suppression, the choice of prophylactic antibiotics is a first generation cephalosporin, such as cefazolin, because the most common organisms causing postoperative infection are Staphylococcus aureus, Staphylococcus epidermidis, and Streptococcus species [20].

1.2.3.2 Venous thromboembolism

Venous thromboembolism is a potentially fatal complication after TKA [25,29]. There is a substantial risk of developing deep vein thrombosis (DVT) and pulmonary embolism (PE) after TKA surgery [30]. Deep vein thrombosis is common even with appropriate thromboprophylaxis and occurs in up to 15% of patients, being symptomatic in 2 to 3% [21]. The overall incidence of DVT after TKA without prophylaxis has been reported to range from 40% to 88%. The risk of asymptomatic PE may be as high as 10% to 20%, with symptomatic PE reported in 0.5% to 3% of patients and a mortality rate up to 2% [20].
During and after TKA, factors significant in the development of venous thrombosis like venous stasis, injury to the vascular endothelium, and release of tissue thromboplastin commonly occur [30]. Further, age over 40 years, female gender, obesity, varicose veins, smoking, hypertension, diabetes mellitus, and coronary disease are factors that have been correlated with an increased risk of DVT [20].
Considering the location of the DVT, proximal thrombi, in the popliteal vein and above, are more common than thrombi in calf veins and have a higher potential to cause PE than thrombi in calf veins [20,30].
As late sequelae, 50% to 60% of patients with symptomatic proximal vein thrombosis and 30% of those with symptomatic calf vein thrombosis develop chronic venous insuffiency [30].
Methods of DVT prophylaxis include mechanical compression stockings or foot pumps and pharmaceutical agents including low-dose warfarin and low-molecular-weight heparin. Pharmacological prophylaxis of DVT in TKA decreases the frequency of fatal pulmonary embolism and is thus strongly indicated [20].

1.2.3.3 Patellofemoral complications

Patellofemoral complications include patellofemoral instability, patellar fracture, component failure, patellar component loosening, patellar clunk syndrome, and extensor mechanism tendon rupture and have been cited as the most common reasons for reoperation [20].
"Patellofemoral instability can be caused by a number of factors, including extensor mechanism imbalance with a tight lateral retinaculum associated with preoperative valgus deformity, excessive lateral patellar facet resection, lateral placement of the patellar component with failure to reproduce the normal position of the median eminence, and early postoperative rupture of the medial capsular repair" [20].

Patellar fracture after TKA is uncommon and has been correlated with multiple factors, including excessive patellar resection, vascular compromise secondary to lateral release, patellar maltracking secondary to component malposition, excessive joint line elevation, knee flexion of more than 115 degrees, trauma, thermal necrosis from PMMA polymerization, and revision TKA [20].

"Patellar component loosening occurs in approximately 0.6% to 2% of arthroplasties" [20]. Deficient bone stock, component malposition and subluxation, patellar fracture, avascular necrosis of the patella, and loosening of other knee components are predisposing factors.

In patellar clunk syndrome, a fibrous nodule forms on the posterior surface of the quadriceps tendon just above the superior pole of the patella. This nodule can become entrapped in the intercondylar notch of the femoral prosthesis and cause the knee to pop or "clunk" at approximately 30 to 45° of knee flexion as the knee is actively extended. The recommended treatment for patellar clunk syndrome is open or arthroscopic debridement of the nodule, with possible revision of the patellar component [20].

"Rupture of the quadriceps or patellar tendon is an infrequent but severe complication of TKA" [20]. In part, quadriceps rupture may be related to lateral release because of vascular comprise of the tendon and possibly with extension of the release anteriorly that weakens the tendon. "*Patellar tendon rupture is associated with previous knee surgery, knee manipulation, and distal realignment procedures of the extensor mechanism"* [20].

1.2.3.4 Vascular complications

Acute occlusive vascular problems after TKA are rare (0.03% to 0.2%). Patients with a history of vascular disease, vascular calcification, previous vascular reconstruction, and possibly popliteal aneurysms are at risk of an acute vascular event [25].

Direct vascular damage may occur intraoperatively, because the popliteal artery and vein, and the tibial nerve lie close to the posterior aspect of the knee. Further, the popliteal artery bifurcates below the level of the joint. At this point, the anterior tibial artery becomes most at risk [25].

Arterial thrombosis after TKA is very rare but a devastating complication that frequently results in amputation [20].

1.2.3.5 Peripheral nerve complications

Nerve injuries, especially peroneal nerve palsy, occur in 1% to 2% of patients [21]. Risk factors for nerve injury include rheumatoid arthritis, preoperative deformity, postoperative epidural anaesthesia, prolonged tourniquet time, and pre-existing peripheral neuropathy [25]. To avoid nerve injury, attention to surgical detail such as careful retractor placement and avoidance of over releasing of soft tissue and excessive polyethylene insert thickness and care when applying dressings should be exercised [25]. When a peroneal nerve palsy is discovered postoperatively, expectant treatment is recommended with immediate local pressure relief and flexion of the affected knee to 20 to 30° advised [20,25].

1.2.3.6 Fractures

Periprosthetic fractures may occur around the femoral component, the tibial component, or the patella. Reported risk factors include anterior femoral notching, osteoporosis, rheumatoid arthritis, stiff knees, revision arthroplasty, and neurological disorders [20,31]. Femoral periprosthetic fractures are more common and occur in 2%, while the incidence of tibial fractures is much less [32].

Function of the TKA after fracture healing depends on restoration of alignment, adequate patellofemoral function, maintenance of prosthesis fixation, and adequate residual motion [20].

1.2.3.7 Loss of motion

The primary goals of joint arthroplasty include pain relief and restoration of function. Activities of daily living need a certain range of motion. For example, 70° of knee flexion are needed to walk normally on level ground, 90° to go up most stairs, 100° to come down those stairs, 105° to get up from most chairs, and 115° to get up from a low sofa [33,34]. Satisfactory postoperative range of motion is thus an important component of successful total knee replacement [34]. Further, range of motion is an extremely important determinant of patient satisfaction [35].

Postoperative stiffness is often defined by a range of motion less than 90° at six weeks after operation. A number of factors may lead to a loss of motion after TKA, including patient factors, preoperative ROM, prosthetic geometry, intraoperative technical errors, knee kinematics, postoperative rehabilitation, and perioperative complications [25,36]. Patient factors predictive of stiffness include high Body Mass Index (BMI), previous knee surgery, patients on disabilities, diabetes, depression, and pulmonary disease. Stiffness is more common in women, and women of a younger age with a low BMI, high femoral flexion angle, and a patella baja. Further, preoperative range of motion (ROM) influences postoperative range [25].

Perioperative complications, which may influence postoperative ROM, are infection, periprosthetic fracture, component failure, complex regional pain syndrome, heterotopic ossification, and postoperative medical complications.

Postoperative ROM may be improved by positive preoperative reinforcement, multimodal pain management, and physiotherapy [25].

1.2.4 Postoperative swelling

Lower limb swelling occurs in most patients who undergo TKA, which can cause pain, extended bed-rest, and a delay in rehabilitation [2]. Further, lower limb swelling may be a warning sign of severe complications, such as DVT and infection [2]. Post-TKA limb swelling is related to damage to blood and lymph vessels, their increased permeability, extravasation into tissue, and the release of inflammatory factors [2].

Total knee arthroplasty requires sequential osteotomy and soft tissue release, which lead to significant blood loss. The total blood loss is composed of ‘visible’ blood loss from the surgical field and wound drainage, and “hidden” blood loss into the tissues [37]. Blood retained in the joint cavity, extravasation of blood into the tissue, and haemoglobin loss due to haemolysis contribute to hidden blood loss [2,38]. Hidden blood loss may be one of the causes or aggravating factors of limb swelling after TKA [2].

After surgery, blood collects in the joint cavity and penetrates into the surrounding soft tissue, increasing the circumference of the extremity in the knee region. Further, fat, bone cement, and bone fragments may enter the circulation, leading to abnormal permeability in the capillaries and subsequently extravasation of blood into the tissue. Extravasation into the tissues can aggravate limb swelling [38]. As a consequence, the soft tissue around the joint is exposed to a higher tension, which causes local pain due to pressure [2].

Current methods of reducing post-operative swelling include elevation of the affected limb and the use of medication. Measures taken to minimize HBL could also reduce limb swelling. These include using ice-packs or elastic bandages, intra-capsule injection of epinephrine, restricting the drainage tube, and restoring blood volume [2].

1.3 Methods to evaluate knee swelling

A variety of methods are used to measure leg swelling. These can be subdivided into direct, indirect and dynamic measurement methods [39]. Direct volume assessment methods include optoelectronic measurements, computed tomography, magnetic resonance imaging scans, and volume displacement. Indirect methods, which are most frequently used, are based on leg circumference measurements. Dynamic measurement methods are based on dynamic manoevres or compression for a short period of time when using volume plethysmography.

Currently, water displacement leg volumetry is considered to be the gold standard or reference method [40,41]. Water displacement volumetry is based on the "*Archimedes' principle, which states that any object that is completely or partially submerged in a fluid at rest is acted on by an upward force. The magnitude of this force is equal to the weight of the fluid displaced by the object. The volume of fluid displaced is equal to the volume of the portion of the object submerged"* [42]. There are two major subtypes of water displacement volumeters. The first and more common variant uses a container with an overflow spout. Water is filled into the container until water flows from the spout. Thereafter the patient immerses the limb into the container. The water flowing from the spout representing the volume of the limb is weighted or its volume is measured in a calibrated container. The second variant of water displacement volumeters measures the level of the water in a container before and after the patient lowered the limb. The rise of water levels is translated into volume change from a calibration curve established with bodies of known volume [41]. Although the water displacement method is regarded as the gold standard, it is not suitable for patients with postoperative wounds [39,43].

In general, direct and dynamic methods are expensive, cause inconvenience to patients and are difficult to apply due to wounds and relative immobilization [39]. Preferably, a simple and fast, but nevertheless reliable and reproducible measurement method is required for patients who experience pain after knee surgery. Circumference measurements with measuring tapes are easy to perform, cheap and applicable in clinical practice [39].

1.4 Methods to evaluate knee flexion

Goniometry for measuring knee range of motion (ROM) is well entrenched in the orthopaedic field. As severe restriction in ROM has ramifications for gait, function, and the need for manipulation, goniometry is a measure of particular importance. Furthermore, knee flexion and extension ROMs are incorporated into orthopaedic knee scoring tools to assess disease severity, and frequently used to assess recovery after various knee surgeries [44].

Knee range of motion can be determined by visual estimation, universal goniometers, digital gravity goniometers, or measurement of joint angles after X-ray visualization [45]. The choice of the appropriate instrument is based upon the purpose of the measurement (i.e., clinical or research), the motion being measured, and the instrument´s accuracy, availability, cost, ease of use, and size [46].

The measured radiographic angulation between the long axis of the femur and the long axis of the tibia is considered the gold standard for measurement of knee ROM [35,47].

The instrument most commonly used to measure range of motion in the clinical setting is the universal goniometer, which can be used at almost all joints of the body. Universal goniometers may be constructed of plastic or metal and are produced in many sizes and shapes. Whatever the size or material, the basic design of every universal goniometer includes a body and two arms - a stationary arm and a moving arm. The body resembles a protractor and may form a half circle or a full circle. A measuring scale is located around the body. While the scales on the half-circle goniometer read from 0 to 180° and from 180 to 0°, the scales on the full-circle models may read either from 0 to 180 degrees and from 180 to 0°, or from 0 to 360° and from 360 to 0°. The intervals on the scales may vary from 1 to 10° [46]. The stationary arm is structurally a part of the goniometer´s body and cannot be moved independently of the body. The moving arm is attached to the fulcrum in the centre of the body by a rivet or a screw-like device that allows the arm to move freely on the body. The length of the goniometer´s arms varies among instruments, depending on the size of the joints to be measured [46].

In gravity-dependent goniometers or inclinometers, the gravity´s effect on pointers and fluid levels is used to measure joint position and motion. Gravity-dependent goniometers are attached to or held on the distal segment of the joint to be measured. The angle between the long axis of the distal segment and the line of gravity is noted. Although inclinometers may be easier to use in certain situations than universal goniometers because they do not have to be aligned with bony landmarks or centred over the axis of motion, it is critical that the proximal segment of the joint being measured is positioned vertically or horizontally to obtain accurate measurements. Furthermore, inclinometers are difficult to use on small joints and on regions with soft tissue deformity or edema [46].

Electrogoniometers are used primarily in research to obtain dynamic joint measurements. Most devices have two arms which are attached to the proximal and distal segments of the joint being measured. The two arms are connected to a potentiometer. If the joint position changes, the resistance in the potentiometer varies. The resulting change in voltage is used to indicate the amount of joint motion. Electrogoniometers are expensive and their use is time-consuming because they have to be accurately calibrated and attached to the subject.

Other joint measurement methods used more commonly in research setting are radiographs, photographs, film, videotapes, and computer-assisted video motion analysis systems [46].

Furthermore, the range of motion is often assessed by visual estimation in clinical practice [48].

1.5 Methods to evaluate pain

There are a number of scales for evaluating pain. Among these, the Visual Analogue Scale (VAS), Numerical Rating Scale (NRS), Verbal Rating Scale (VRS), and the Faces Pain Scale-Revised (FPS-R) are the most commonly used measures of pain intensity in clinical and research settings [49-51].

The VAS is a 10 cm line anchored by verbal descriptors, usually “no pain” and “worst imaginable pain”. The patient is asked to mark the line corresponding to the intensity of present pain. This distance from the zero anchor to the patient´s mark gives the score. Using a millimetre scale to measure the patient´s score provides 101 levels of pain intensity [49,50]. One of the limitations of the VAS is that it must be administered on paper or electronically and photocopying the scale can lead to significant changes in it´s length [50]. Because graphical orientation of the VAS may be important, both horizontal (VAS-H) and vertical (VAS-V) orientations were employed [49]. The graphical orientation of the VAS should be decided according to the normal reading tradition of the population on which it is being used [50]. An advantage of the VAS is that pain is measured continuously. However, the VAS is cumbersome to administer because it requires adequate levels of visual acuity, motor function, and the cognitive ability to translate a sensation of pain into a distance measure [52].

The Verbal Rating Scale (VRS) comprises a list of adjectives used to denote increasing pain intensities. The most common words used are “no pain”, “mild pain”, and “severe or intense pain”. For ease of recording these adjectives are assigned numbers [50]. Patients are asked to pick the single word that best describes their current pain intensity, and their VRS intensity level is the number associated with the word the patient chose [51]. The VRA is ordinal [50].

A commonly used clinical measure of pain is the numerical rating scale (NRS) [50]. The Numerical Rating Scale is an 11, 21, or 101 point scale where the end points are the extremes of no pain and pain as bad as it could be, or worst pain. The NRS can be graphically or verbally delivered. When presenting graphically the numbers are often

enclosed in boxes and the scale is referred to as an 11 or 21 point box scale depending on the number of levels of discrimination offered to the patient [50]. For the verbally delivered NRS, patients are asked to indicate the intensity of pain by reporting a number that best represents it. NRS provides interval level data [50].
Williamson and Hoggart (2005) reviewed the three pain rating scales VAS, NRS, and VRS. They concluded that all three of the pain-rating scales are valid, reliable and appropriate for use in clinical practice, although the VAS was the most difficult to use in clinical practice and had the highest failure rate. Further, they stated that the NRS is probably more useful than the VRS or the VAS as a tool for pain assessment as well as for audit and research [50].

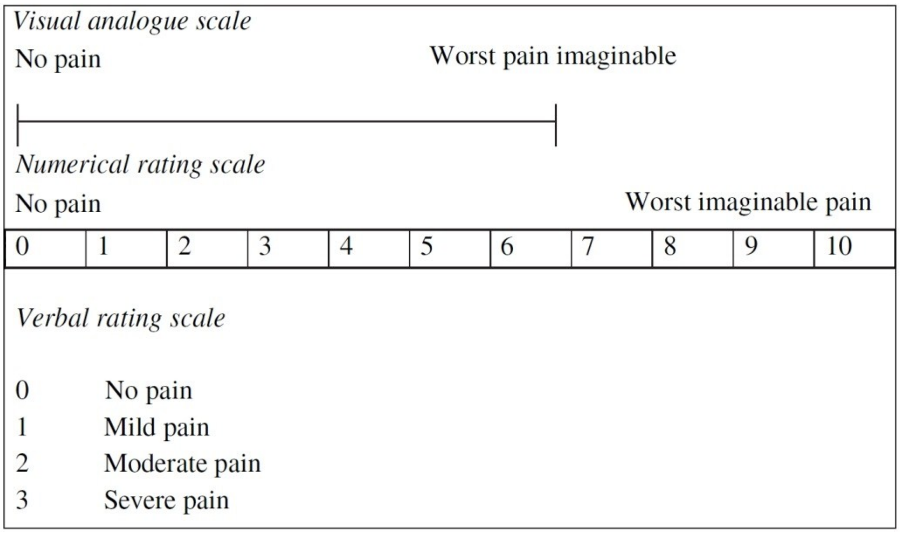


Figure 4: Scales to evaluate pain (VAS, NRS, VRS)
NRS Numeric Rating Scale; NRS, numerical rating scale; VDS Verbal Descriptor Scale
Source: Williamson A, Hoggart B. J Clin Nurs. 2005.[50]

1.6 Measurement error, Agreement and Reliability

1.6.1 Measurement error

No measurement is perfect. As all instruments and observers or measurers (raters) are fallible to some extent and all humans respond with some inconsistency, any observed score X is the sum of the true value T and an error component E (1) [53-55]:

$$X = T \pm E \quad (1)$$

Thus, the difference between the true value and the observed value is the measurement error. The total error consists of systematic error and random error [53-56].

Systematic (or bias) errors are consistent, repeatable errors and refer to a general trend for measurements to be different in a particular direction (either positive or negative) between repeated tests. For example, there might be a trend for a retest to be higher than a prior test due to a learning effect being present [53,56,57];.

Random (or precision) errors are inconsistent and unrepeatable and refer to sources of error that are due to chance factors, like luck, alertness, attentiveness by the tester and normal biological variability. Random errors increase and decrease measured scores on repeated testing in a random manner [53,56,57].

1.6.2 Terms reproducibility, agreement, reliability and responsiveness

An essential requirement of every measuring instrument is to be valid and reproducible or reliable. Reproducibility concerns the degree to which repeated measurements in stable objects, e.g. subjects or patients, provide similar results [58]. The repeated measurements may concern the same observer at different times to investigate measurement error (intra-observer variation), or different observers to investigate the variation between them (inter-observer variation) [59,60]. Repeated measurements may differ because of biologic variation in subjects, e.g. in the form of day-to-day differences or a circadian rhythm. *Other sources of variation may originate from the measurement instrument itself, or the circumstances under which the measurements take place* [59].

Reproducibility is an umbrella term for the two concepts of reliability and agreement, which are often incorrectly used interchangeably [58,59].

Measures of agreement refer to the absolute measurement error and determine the ability of observers to achieve the same value in repeated measurements. They are expressed in the units of the measurement, which is an important advantage for clinical interpretation [5,59].

Measures of reliability assess the ability to differentiate among subjects in a group, despite measurement errors, and thus refer to the relative measurement error. They provide insight into the ability of observers to differentiate between subjects in a group. The measurement error is related to the variability between persons. Consequently, reliability parameters are highly dependent on the heterogeneity of the study sample, which is not the case for agreement parameters, which are based on measurement error [5,59].

A reliability parameter has a typical basic formula (2):

$$\text{reliability} = \frac{\text{variability between study objects}}{\text{variability between study objects} + \text{measurement error}} \tag{2}$$

Reliability relates the measurement error to the variability between study objects. The reliability parameter approaches 1, if the measurement error is small compared to the variability between persons. In this case the discrimination of persons is hardly affected by measurement error, and thus the reliability is high. Conversely, if the measurement error is large compared to the variability between persons, the reliability parameter approaches 0. In such a case, the discrimination will be affected by the measurement error [59].
The responsiveness of a measure is the ability of the tool to detect a real change [44]. To assess the responsiveness of an instrument, an external criterion can be used to define whether a subject or patient has changed. *The external criterion determines the minimum change that is considered to be clinically relevant* [61]. Responsiveness is strongly related to the level of agreement [5,62].

1.6.3 Statistical Methods to assess reproducibility

A variety of statistical procedures have been proposed to assess reliability and agreement in circumferential and goniometric measurements. These include the intraclass correlation coefficient (ICC), the standard error of measurement (SEM), the Pearson correlation coefficient, the coefficient of variation, and the Bland and Altman 95% limits of agreement method (LOA) [44,45,58]. A brief review of literature for circumferential and/or goniometric reliability studies showed, that the most common methods involve the use of correlation coefficients (ICC, Pearson's, Spearman) and/or hypothesis tests (ANOVA) (Table 1). In 26 of the 30 cited articles (1976-2011), the used statistical method was a correlation coefficient, alone, or in combination with other statistical methods. Other methods cited in the reviewed literature involved standard error of measurement (SEM), coefficient of variation (CV), or the Bland and Altman 95% limits of agreement (LOA).
There is no consensus regarding the appropriate statistical test in reproducibility studies of continuous data. Several authors have discussed the inappropriateness of using tests such as Pearson´s correlation, t-tests, coefficient of variation, per cent agreement and chi-square [63]. De Vet pointed out that regardless of the preferred statistical methods, presenting one single reliability coefficient is insufficient and a visual presentation of the data is advisable [58]. Rankin and Stokes stated that the intraclass correlation coefficient and Bland and Altman tests are appropriate and recommended the combined use for the analysis of reliability studies [63].
Considering responsiveness, the smallest real difference (SRD) or the smallest detectable difference (SDD) derived from reproducibility can be used to define responsiveness [61].

The following chapters will look in more detail at the intraclass correlation coefficient (ICC) and the Bland and Altman 95% limits of agreement, which were used in the statistical analysis.

Statistical method		Measurement study		
		Circumference (n = 14)	Goniometry (n = 18)	Circumference and goniometry (n = 3)
Correlation coefficient	ICC	7	10	2
	Pearson´s	3	3	
	Spearman		1	
	CCC	-	1	-
	type not specified	-	1	1
ANOVA		5	3	1
LOA		1	3	1
SEM		3	1	1
SDD		-	1	-
SRD		-	-	1
CD		1	-	-
CV		1	-	-
CR		1	-	-
RP		1	-	-

Table 1: Statistical methods used in circumferential and goniometric reliability studies
Review of 11 articles for leg circumferential [39,64-73], 15 articles for knee goniometric studies [5,35,44,45,47,48,74-82], and 4 articles for both circumferential and goniometric measurement reliability [4,83-85]; ICC; intraclass correlation; ANOVA; analysis of variance; SEM; standard error of measurement; CV; coefficient of variation; SDD; smallest detectable difference; SRD; smallest real difference; CD; critical difference; CR; coefficients of repeatability; LOA; limits of agreement; RP; relative precision;

1.6.3.1 Intraclass correlation coefficient (ICC)

The ICC is defined as the ratio of the variance of interest, e.g. persons, to the total variance (3). These variances are derived from analysis of variance (ANOVA) [57-59]. The ICC is a measure of reliability and considers both random and systematic error, but does not allow to discriminate between random and systematic error [45,57].

$$\mathrm{ICC} = \frac{\text{Variance in persons}}{\text{Variance in persons} + \text{Variance (error)}} \tag{3}$$

The ICC is unitless and can theoretically vary between 0.0 and 1.0. An ICC of 0 indicates no reliability, whereas an ICC of 1 indicates perfect reliability [57]. An ICC of 0.95 means that an estimated 95% of the observed score variance is due to true value variance [57]. Some sources have attempted to delineate good, medium, and poor levels for the ICC, ranging from 'questionable' or "acceptable" (0.7 to 0.8) to 'high' (>0.9) [56,57,62]. Anyway, there is no consensus as to what constitutes a good ICC. This might be due to the fact that there are six commonly used versions of the ICC and the resulting ICC value varies depending on which version of the ICC is used [56,57].
The structure of the ANOVA model for the appropriate ICC depends on whether the observers are drawn at random from a large population of observers (random effects) or whether they are the only observers of interest (fixed effects), whether each observer rates each subject or not, and whether the unit of analysis is a single observer or the mean of several observers (Table 2) [58,86].
Furthermore, to compute an ICC with the statistical software IBM SPSS Statistics [87] or MedCalc for Windows [88], one has to distinguish between two further types of ICCs, $ICC_{agreement}$ and $ICC_{consistency}$. These differ whether or not systematic differences between observers are taken into account (4, 5):

$$\mathrm{ICC_{agreement}} = \frac{\sigma_p^2}{\sigma_p^2+\sigma_o^2+\sigma_{residual}^2} \tag{4}$$

$$\mathrm{ICC_{consistency}} = \frac{\sigma_p^2}{\sigma_p^2+\sigma_{residual}^2} \tag{5}$$

The variance in persons σ_p^2represents the variability between persons (i.e. subjects or patients), and σ_o^2 represents the variance due to systematic differences between observers. In case of $ICC_{agreement}$, the measurement error consists of $\sigma_o^2+\sigma_{residual}^2$, while it is only $\sigma_{residual}^2$ in case of $ICC_{consistency}$. The $ICC_{agreement}$ differs from the $ICC_{consistency}$ in the extra term σ_o^2 in the denominator, taking into account the systematic difference between the observers. Thus, the $ICC_{consistency}$ ignores systematic differences. The specifications "agreement" and "consistency" for the type of ICC are somewhat confusing, because both are reliability parameters, which are dependent on the heterogeneity of the population sample with respect to the characteristic of the study [59].

The ICC is derived from an ANOVA table, which implies that the estimation of the reliability of ratings is based on the assumptions that the data are normally distributed and

that the variances are homogeneous between ratings [54]. The ICC includes the variance term for individuals and is therefore dependent upon heterogeneity of the study population [45,56,57,59]. A large ICC can mask poor trial-to-trial consistency when between-subject variability is high. Conversely, if the between-subjects variability is low, a low ICC can be found even when trial-to-trial variability is low [57]. Thus, a given method can have different reliability determined from the ICC, depending on the characteristics of the individuals included in the analysis [57]. Due to these limitations, the ICC should not be employed as the sole statistic [56].
In addition, the ICC is a unitless ratio of variances and cannot be interpreted clinically because it gives no indication of the magnitude of disagreement between measurements. Therefore, it should be complemented by calculation of the standard error of measurement (SEM) or the Bland and Altman 95% limits of agreement tests [58,63].

<table>
<tr><td rowspan="2">Observers</td><td colspan="2">Each subject was assessed by a different set of randomly selected observers</td><td colspan="4">Each subject was assessed by each observer</td></tr>
<tr><td colspan="2">Observers were selected at random</td><td colspan="2">Observers were selected at random</td><td colspan="2">Observers are the only observers of interest</td></tr>
<tr><td>Unit of analysis</td><td>Single rating</td><td>Average of k ratings</td><td>Single rating</td><td>Average of k ratings</td><td>Single rating</td><td>Average of k ratings</td></tr>
<tr><td>ICC type</td><td>ICC(1,1)</td><td>ICC(1,k)</td><td>ICC(2,1)</td><td>ICC(2,k)</td><td>ICC(3,1)</td><td>ICC(3,k)</td></tr>
<tr><td rowspan="2">SPSS/ Medcalc Model</td><td colspan="2">One-way random</td><td colspan="2">Two-way random</td><td colspan="2">Two-way mixed</td></tr>
<tr><td>single measure</td><td>average measure</td><td>single measure</td><td>average measure</td><td>single measure</td><td>average measure</td></tr>
</table>

Table 2: ICC types and corresponding SPSS/MedCalc model
[86]

1.6.3.2 The Bland and Altman limits of agreement

The 95% limits of agreement method proposed by Bland and Altman (1986) is based on the mean and standard deviation (SD_{diff}) of the difference between two ratings of the same subject [89]. The mean difference between two observers or measurements (mD) indicates systematic error (or bias) and the SD_{diff} of the difference between two observers or measurements indicates random error [89]. The closer mD is to zero and the smaller the value of SD_{diff}, the better the agreement between measures [63]. An mD deviating substantially from 0 indicates a systematic difference between measurements [62]. The 95% limits of agreement (LOA) are determined from mD and SD_{diff} (6):

$$\mathrm{LOA} = \mathrm{md} \pm 1.96 \cdot \mathrm{SD}_{\mathrm{diff}} \tag{6}$$

Thus, the 95% limits of agreement are a measure of random (mD) error and systematic error ($1.96 \cdot SD_{diff}$) of the measurement method. If the differences are normally distributed, 95% of the differences will lie between the limits of agreement [89]. Large limits of agreement show poor agreement between the variables. The minimum acceptable level of agreement depends on the clinical use and situation. It is the task of the researcher to judge, using analytical goals, whether the limits of agreement are narrow enough for the test to be of practical use [56,60].

To explore agreement between measurements, Bland and Altman recommended the visual examination of data patterns with the Bland and Altman plot and quantification of the mD between measurements and the corresponding SD_{diff} [90]. The Bland and Altman plot is a scatter diagram that consists of the average of the paired values from each measurement on the x-axis and the difference of each pair of readings on the y-axis (Figure 5). Further, horizontal lines are drawn at the mean difference mD and at the limits of agreement (LOA). From this graph, the size of each difference, the range of differences and their distribution about zero, which corresponds to perfect agreement, can be seen clearly [63].

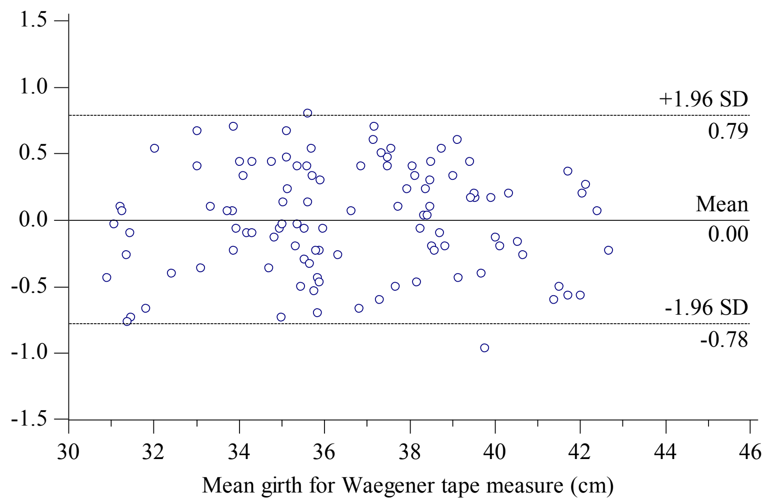


Figure 5: Example for a Bland and Altman plot
with mean difference (solid black line) and limits of agreement (broken black lines);

The presentation of the findings from the Bland and Altman limits of agreement analysis includes the mean value of the measurement, the mean difference mD, the standard deviation SD of the difference, and the limits of agreement (LOA). Further the Bland-Altman plot is displayed as a graphic [57].

From the standard deviation of the difference SD_{diff}, the smallest detectable difference (SDD) can be calculated according to (7):

$$SDD = 1.96 \cdot SD_{diff} \tag{7}$$

The smallest detectable difference (SDD) expresses the smallest difference between two independently obtained measures that can be interpreted as “real” above measurement error [62].

The limits of agreement depend on some assumptions about the data: The mean and the SD are constant throughout the range of measurements and the differences are from an approximately normal distribution. The measurements themselves do not have to follow a normal distribution [89].

Compared to ICC, the limits of agreement do not systematically depend on the heterogeneity of the study population. Further, it quantifies the amount of measurement error in units of the measurement scale used, e.g. in cm or degrees. The limits of agreement method clearly visualises systematic differences and random errors. Apart from that, it is only minimally influenced by the number of observers for inter-observer comparisons [45]. However, the Bland and Altman 95% limits of agreement indicate a range of error, but this must be interpreted with reference to the range of measurement values obtained. Therefore, Bland and Altman tests should be complemented by raw data and/or ranges [63].

1.7 Study hypothesis

The first study hypothesis was that both lower extremity circumferences in the knee region and knee flexion can be reliably assessed with the available measuring instruments.

Further, it was hypothesized that the reproducibility of circumferential measurements depends on the measurement site and the tape measure used. Considering the measuring tapes, it was suspected that the tension controlled tape measures would show higher reliability and agreement than the standard tape measure.

The second study hypothesis was that the rehabilitation of in-patients undergoing TKA can be evaluated by means of circumferential girth measurements to assess swelling, goniometric measurements to assess ROM and the numeric rating scale to assess postoperative pain. Further, it was hypothesised that postoperative swelling has an influence on passive ROM.

2 Methods and Materials

2.1 Reliability and agreement measurements

2.1.1 Subjects and observers

Initially, twenty subjects voluntarily signed a consent form to participate in the study, which had been approved by the ethics committee of Medical University of Graz. The mean age of the subjects was 40.6 (range, 23-60) and 50 % were female. The mean body mass index (BMI) was 23.20 kg/m^2 (range 19.1-30.2 kg/m^2). Subjects were eligible for participation if they gave informed consent, were between 18 and 90 years of age, had no recent history of lower extremity pathology, BMI less than 40 kg/m^2, and were able to co-operate (e.g., no unsoundness of mind). Subjects suffering from a current fracture around the knee, current infection or status post infection, tumour around the knee, rheumatoid knee arthritis, any operations done around the knee except arthroscopic knee surgery, and depression or anxiety disorder, were excluded.
The measurements were performed by five observers covering a broad spectrum of medical and non-medical professions. Three observers had practical experience in girth and ROM measurements (O1, O2, and O3), while the others had never before assessed lower leg circumference with any tape measure used in this study as well as knee flexion with a universal goniometer. Due to time constraints, only two observers (O1 and O2) measured all 20 subjects (Table 3). The observers O3, O4, and O5 measured the same 5 subjects (subjects 5, 8, 13, 14, and 18).

Observer	Sex	Profession	Experience	Subjects measured
O1	Female	Medical/engineering student	yes	20
O2	Female	Medical resident	yes	20
O3	*Male*	*Orthopaedic surgeon*	*yes**	*5*
O4	*Male*	*Engineering student*	*no*	*5*
O5	*Female*	*Secretary*	*no*	*5*

Table 3: Observers
*occasional experience with the Waegener tape measure used in the study

For girth measurements, twenty subjects were measured on the first measuring day by the observers O1 and O2, and eighteen subjects on the second measuring day, because two subjects did not return for their second testing (Figure 6, solid line). The sample size n was thus 20 for the first and 18 for the second measuring days, except for Gulick I tape measure: On the second measuring day, only 14 subjects were measured with this measuring tape. The subjects 5, 8, 13, 14, and 18 were measured by observers O1, O2, O3, O4, and O5 on both measuring days with all tape measures (n=5) (Figure 6, broken line).

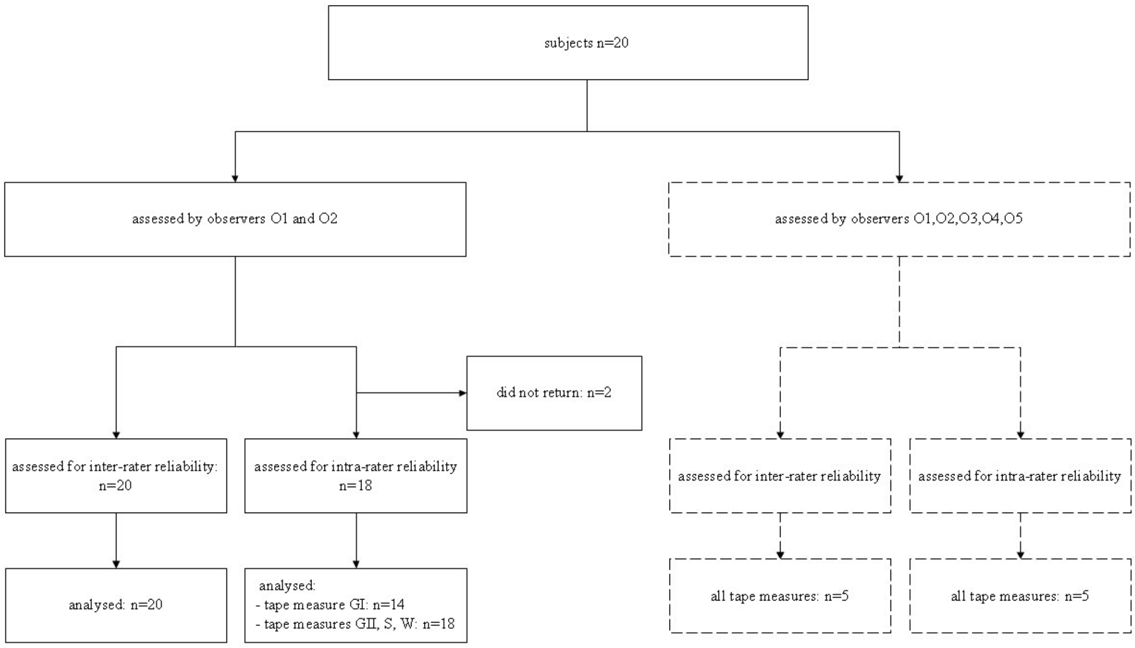


Figure 6: Number of subjects (girth measurements)
n, sample size; GI, Gulick I tape measure; GII, Gulick II plus tape measure; S, standard tape measure; W, Waegener tape measure

For goniometric measurements, twenty subjects were measured on the first measuring day and eighteen subjects on the second measuring day, because two subjects did not return (Figure 7). Accidentally, the knee flexion measurements in subject 19 were taken with the first positioning device (2.1.3.2) on the first measuring day, and with the second positioning device on the second measuring day. Thus, the data of subject 19 (male) was excluded from further statistical analysis, reducing the sample size to n=19 on the first measuring day and n=17 on the second measuring day. The subjects 5, 8, 13, 14, and 18 were measured by observers O1, O2, O3, O4, and O5 on both measuring days (n=5).

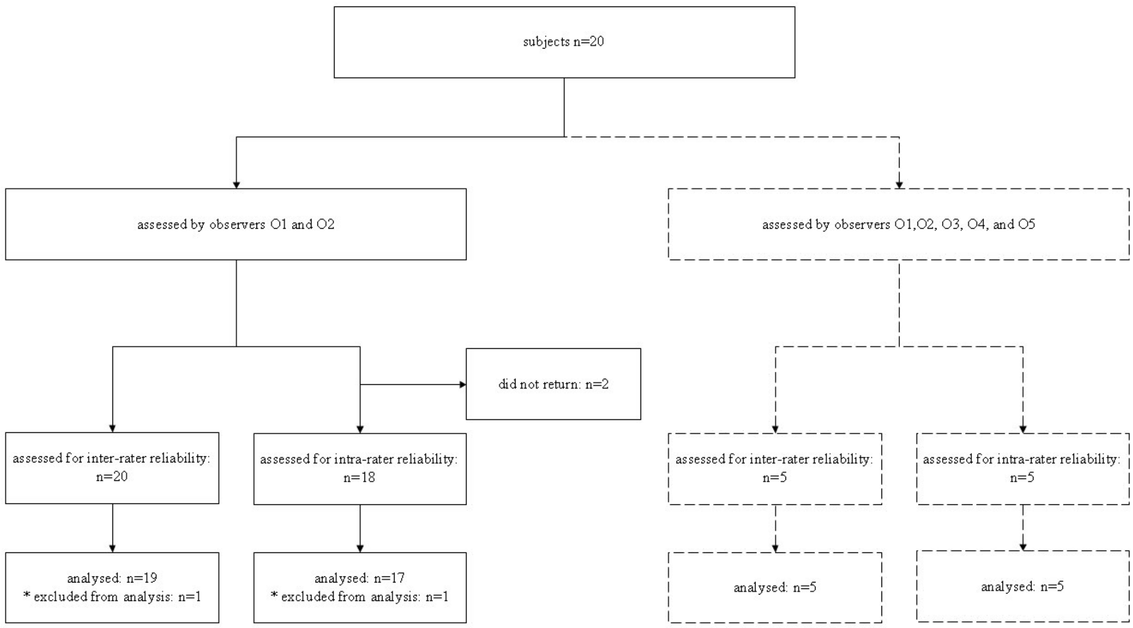


Figure 7: Number of subjects (knee flexion measurements)

2.1.2 Measuring devices

2.1.2.1 Tape measures

Four commercially available tape measuring devices were used for this study: A standard tape measure (Prym Consumer GmbH, Stollberg, Germany), a Gulick I measuring tape (Baseline Evaluation Instruments, Fabrication Enterprises, New York, USA), a Gulick II plus, Model 67019 measuring tape (Country Technology, Inc., Gays Mills, USA) and a circumference tape measure (Waegener, Beerse, Belgium). The tape portion of every device was constructed of non-stretchable material. All measurements were recorded to the closest 1 millimetre (mm). The measuring devices were used as recommended in the operator´s manual, if available.

2.1.2.1.1 Standard tape measure

The standard tape measure (S) was 1.8 cm in width and 150 cm in length. The measuring range of this tape measure was 0-150 cm (increment 0.1 cm). The scale was printed in the direction perpendicular to the axis of the tape.

In general, the measurement obtained with an ordinary tape depends on how tightly the tape is pulled. The harder the observer pulls the tape, the greater the tissue compression and the smaller the measured circumference. Further, it is likely that different observers apply a different tension on the tape.

To measure a limb girth, the standard tape measure was wrapped around the lower extremity and positioned at the measurement site as shown in Figure 8. The observers were advised that the tape should have contact with the skin, conform to the body surface being measured, and not compress the underlying soft tissue.

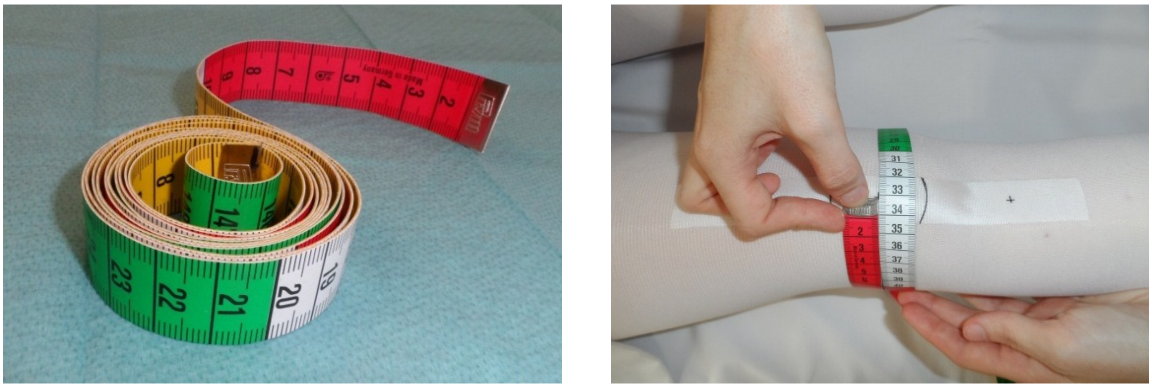


Figure 8: Standard tape measure

2.1.2.1.2 Circumference tape measure "Waegener"

The Waegener tape measure (W) is a spring tape that was designed specifically for circumferential measurements, as the portion of the device which contacts the skin has a concave surface to conform to the limb´s rounded surface (Figure 9). The tape was 1.3 cm in width and 157 cm in length with a measuring range from 0 to 150 cm (increment 0.1 cm). The scale was printed in the direction perpendicular to the axis of the tape.

The Waegener tape measure contains a spring mechanism which maintains an unknown constant tension on the tape during measurement. This avoids the tendency to apply different tensions on the tape at different measuring times.

Figure 9: Waegener circumferential tape measure

To take a measurement with the Waegener tape measure, the tape was pulled out of the box and positioned around the measurement site. The pin located at the end of the tape measure was hooked into the tape measure body forming a loop. The release button was then pressed to tighten the tape measure around the limb with a uniform amount of tension. The circumference was read from the ruler.

A characteristic of this device is that it enables taking measurements and recording of the measured values by the same person. After the release button is pressed, the measuring tape stays in place, while the observer has both hands to make adjustments and record measurements.

2.1.2.1.3 Gulick I measurement tape

The Gulick I measurement tape (GI) was designed with a six ounce spring-loaded tension indicator mounted at the end of the tape (Figure 10). Taking a measurement, the spring exerts a constant, calibrated tension on the tape. According to the manufacturer, this eliminates excessive compression of soft tissue of the limb and resulting measurement inaccuracies.

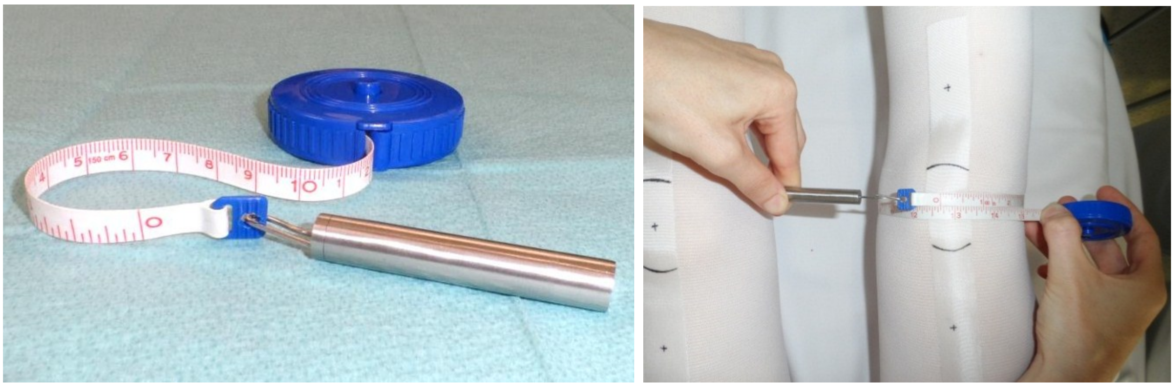


Figure 10: Gulick I tape measure

The Gulick I measurement tape was made of non-stretchable, flexible vinyl that is 0.7 cm in width and 150 cm in length (increment 0.1 cm) with a measuring range from 0 to 150 cm and 0 to 60 in. The scale (in cm on one side of the tape and in inches on the other side) was printed in direction of the tape axis.

To use the Gulick I measurement tape, the tape is pulled out of the housing and wrapped around the leg at the site to be measured. The tape's "zero line" is aligned alongside of the tape graduations. Then, the observer has to pull at the end of the tensioning mechanism until the calibration mark is just seen (Figure 11). The measurement next to the tape's "zero line" is read.

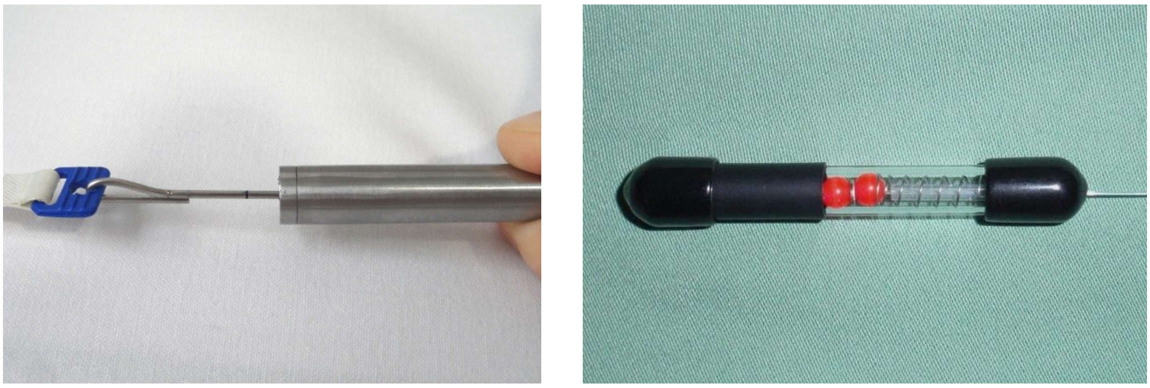


Figure 11: Tension indicators of the Gulick I (left) and Gulick II tape measures (right)

2.1.2.1.4 Gulick II plus tape measure

The Gulick II plus tape measure (GII) consists of a non-stretch, pliable, self-retracting fiberglass tape with both Metric (in centimetres) and English (in inches) gradations and a tensioning device attached to the measuring tape (Figure 12). As for the Gulick I tape measure, the tensioning device provides a known amount of tension while a measurement is being taken and is calibrated to indicate a four-ounce tension. The Gulick II plus tape measure is 1.5 cm in width and 305 cm in length. The measuring range is 0-305 cm and 0 – 120 in. The scale (in mm and inches on one side, in inches only on the other side) is printed in direction of the tape axis.

To measure a leg circumference, the measuring tape is wrapped around the lower extremity to be measured and the tape's "zero line" (end of tape) is aligned alongside of the tape graduations. One hand pulls at the end of the tensioning mechanism until the calibration mark between the two red balls is just seen (Figure 11). Then the measurement is read next to the tape's "zero line".

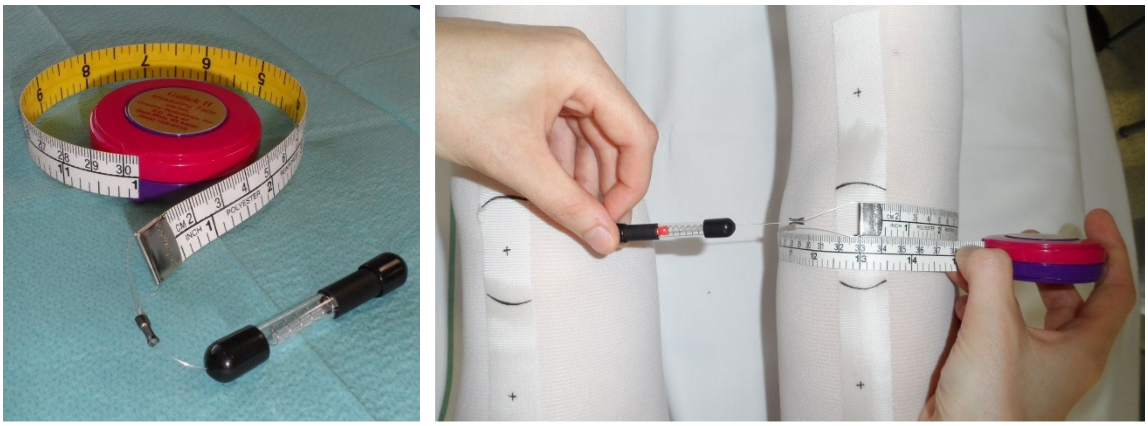


Figure 12: The Gulick II plus tape measure

2.1.2.2 Universal goniometer

For the goniometric measurements, a short arm universal goniometer was used. This goniometer type was chosen because it was thought to be the most commonly used in the clinical setting. Further, Rothstein et al. (1983) had demonstrated that the reliability of a small plastic goniometer with 6-in[a] movable arms, comparable to the one used in this study, was as reliable as a large metal goniometer with 12-in moveable arms and a large plastic goniometer with 10 in-movable arms [82]. The goniometer used in this study had a scale of 360 degrees in steps of 2-degree increments and was made out of clear plastic (Figure 13). The arms were 18 cm in length and provided a linear scale in centimetres.

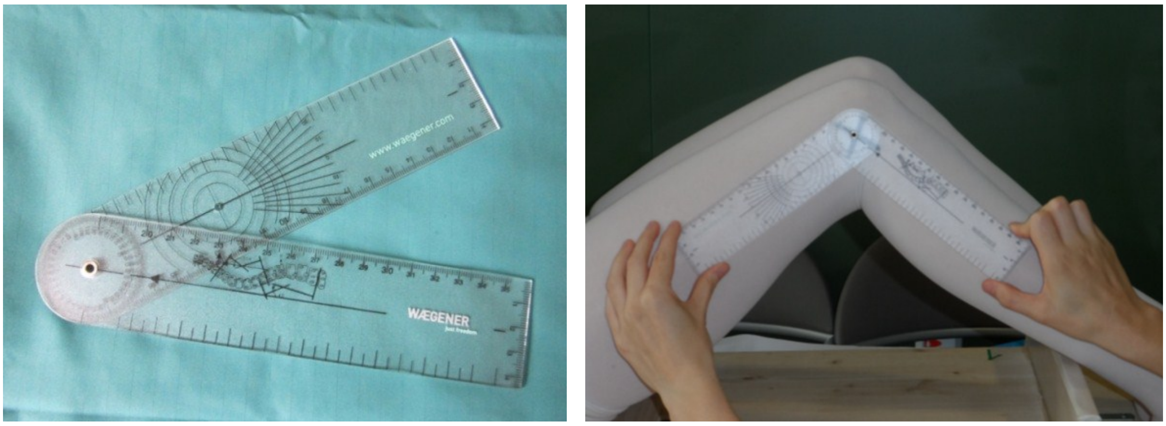


Figure 13: Universal Goniometer

2.1.3 Measurement procedures

2.1.3.1 Girth measurements

The subjects were measured when relaxed and recumbent with their left and right leg in full extension. The limbs were placed on a firm cylindrical paper roll used in hospital to cover the examination bed. The paper roll had a diameter of 16 cm and a length of 50 cm and was positioned under the patient´s heels of the foot (Figure 14).
In order to avoid marking the skin of the subjects, the marks for the three measurement sites were taken on a 2.5 cm white adhesive tape. The adhesive tape ensured that the marks would not remain on the extremity after measurement of the site. It was felt that any residual marks might have influenced subsequent measurements.
To avoid irritation of the skin when pulling off the adhesive tape after completing the measurements of one observer, the subjects were instructed to wear stockings usually used

[a] 6 in = 15.24 cm, 10 in = 25.40 cm, 12 in = 30.48 cm

in medical thrombosis prophylaxis. The adhesive tape was put on the stocking after positioning the legs on the paper roll.

The lower extremity girth was determined by measurement of the transverse plane circumference of the knee at mid-patellar height, as well as 7 cm below mid-patellar height and 7 cm above mid-patellar height. The tester stood closest to the examined leg. The patella was located by palpation, and the superior border and lower inferior pole of the patella were marked on the adhesive tape to determine the length of the patella. The mid-patella was then defined by dividing the length of the patella in half. Starting from the mid-patella site a distance of 7 cm was measured to determine the measurement sites at 7 cm proximal and distal of mid-patella.

The circumferential measurements were performed at these locations in the hope that they would reflect changes in fluid and synovial tissue (mid-patella), muscle atrophy and fluid in the suprapatellar pouch (7 cm above the patella) [70]. The sequence of measurements was repeated three times at each knee by the examiner in the same order. Circumferential measurements were recorded to the nearest 0.1 cm.

After finishing the measurements on a subject, the observer removed the adhesive tape. No semipermanent marks were left on the stocking, so that the next examiner had to identify the measurement site independently.

The observers measured each subject with each measuring tape at the three measurement sites. Thus, all subjects had 36 measurements taken on the right and on the left leg during each visit. A total of 72 measures were obtained on each leg during the first and the second visit.

The entire examination was repeated approximately one week later on the same subjects.

Figure 14: Subject positioning for girth measurements

2.1.3.2 Goniometric measurements

To measure knee flexion, the fulcrum of the goniometer was aligned with the lateral epicondyle of the femur (Figure 15). The stationary arm was placed parallel to the long axis of the femur along a line extending from the greater trochanter to the lateral condyle, and the moving arm was placed parallel to the long axis of the fibula in line with the head of the fibula and the lateral malleolus.

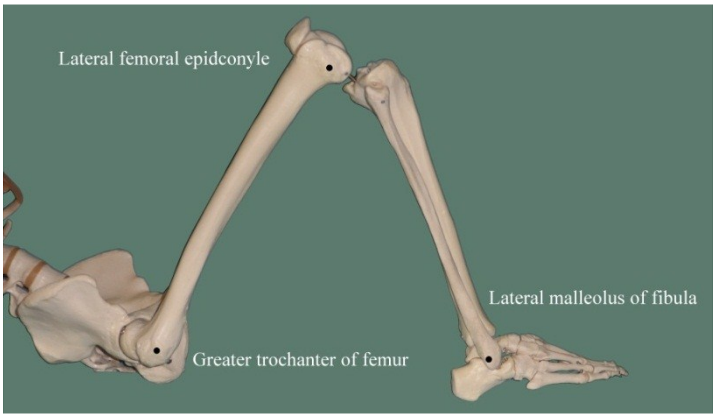


Figure 15: Anatomic landmarks used for the alignment of the universal goniometer

To fix the lower extremities of the subjects in a standardised knee position, two different custom-made devices were used (Figure 16 and Figure 17).

The first positioning device consisted of a 80 cm x 30 cm rectangular ground plate, a 30 cm x 20 cm thigh support plate, a 45 cm x 20 cm calf support plate and an 18 cm x 20 cm foot support plate, made of spruce wood (Figure 16). The ground plate, the thigh support plate and the calf support plate were connected to one another by hinges. On the ground plate, two pins each were mounted at a distance of 48 cm and 59 cm from the end of the plate. These pins served for adjusting the knee joint position, allowing the subject´s leg to be fixed in two knee positions (P1 and P2).

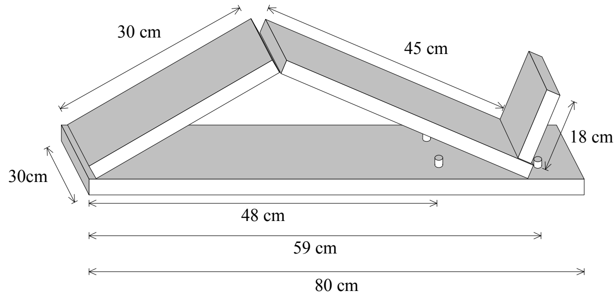


Figure 16: First positioning device and subject in position P2

For goniometric measurements, the subject´s leg was positioned on the thigh and calf support plates as shown in Figure 16 on the right. During the measurements of the first ten subjects, it turned out that this positioning device was only suitable for subjects with a sufficient thigh length.

To overcome the limitations of the first positioning device, another device was constructed. This second positioning device consisted of a 105 cm x 49 cm rectangular ground plate made of spruce wood and three rectangular boards (1, 2, 3) mounted on the ground plate (Figure 17). Two of these boards (1, 3) were mounted at both ends of the ground plate, and one (2) at a distance of 70 mm from the first board (1). For goniometric measurements, the subject was sitting in the positioning device with his/her lower back in contact with the first board and the tip toes either in contact with the second board (flexion position P1) or the third board (flexion position P2). Therefore, the second positioning device provided two standardised knee joint positions.

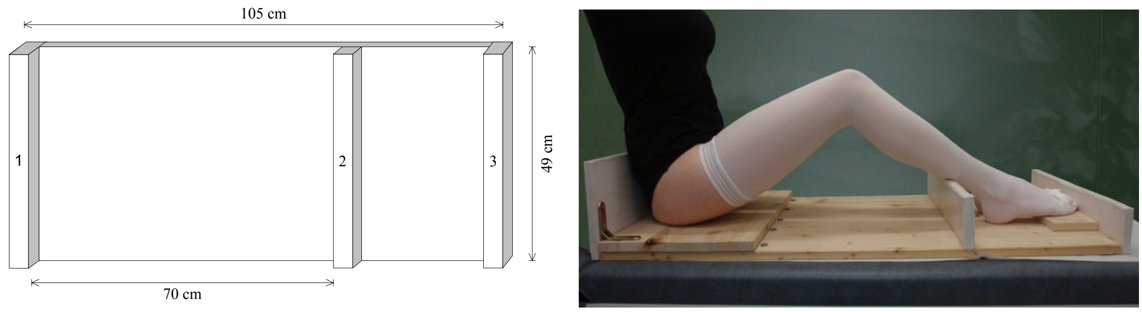


Figure 17: Second positioning device and subject in position P2

2.2 Clinical course after TKA

2.2.1 Patients

Over a period of three months (March to July 2012) patients undergoing TKA at the Department of Orthopaedics and Orthopaedic Surgery of the Medical University of Graz were recruited. Twenty-nine patients voluntarily participated in the study. The patients' clinical data such as age, gender, height, weight, body mass index (BMI) were recorded. The mean age of patients was 69.8 (range, 50-85) and 58.6% were male (Table 4). The mean body mass index (BMI) was 28.8 kg/m^2 (range, 23.0-36.9 kg/m^2). TKA of the left knee had been performed in 13 cases, while 16 patients had a right knee TKA.

Patients	n	Mean age (range) in years	Mean BMI (range) in kg/m^2
Female	12	74.3 (65 to 85)	29.7 (23.0 to 36.9)
Male	17	66.6 (50 to 79)	28.1 (23.1 to 33.9)
Total	29	69.8 (50 to 85)	28.8 (23.0 to 36.9)

Table 4: Characteristics of patient sample (n=29)
n number of subjects, BMI body mass index

The criteria which were a prerequisite for inclusion in the study (inclusion criteria) or led to exclusion from the study and from further evaluation (exclusion criteria) are listed below (Table 5). The study was part of a larger trial approved by the ethics committee of the Medical University of Graz.

Inclusion criteria	Exclusion criteria
Sex either, age >18 and < 90 years	Age < 18 or > 90 years
BMI < 40 kg/m^2	BMI > 40 kg/m^2
Degenerative or posttraumatic	Varus or valgus deformity greater than 10°, impaired extension
Scheduled operation for TKA	Any operations done around the knee except arthroscopic knee
Patient agrees with study design, therapy	Incompliance concerning patient controlled analgesia
Education form is signed by patient and	No signed education form
	Current fracture around the knee, current infection or status post infection, rheumatoid arthritis at knee, tumour around the knee
	Active systemic infection (HIV, HBV, HCV)
	Obstructive sleep apnea
	Opioid intolerance
	Circulatory disorder in the affected leg
	Fibromyalgia or other chronic pain syndromes
	Taking of immune modulating medication such as cortisone, interferon or similar
	Depression or anxiety disorder
	Addicted to drugs or alcohol
	Pregnancy or possible pregnancy without adequate contraception
	Unsoundness of mind

Table 5: Inclusion and exclusion criteria (patients)

All patients had patient-controlled analgesia (Hydromorphine) for 72 hours postoperatively and received daily cryotherapy. Drainage was removed 48 hours after the operation. Low-molecular-weight heparin subcutaneously was administered post-operatively for medical thrombosis prophylaxis. In mean, the patients were discharged 8.5 (range, 6-14) days postoperatively. The stitches were removed 2 weeks after surgery. During the in-hospital phase patients received daily physiotherapy that consisted of active and passive mobilisation of the knee and functional exercises including transfers from a supine position to sitting and from sitting to standing, walking and stair climbing.

2.2.2 Data acquisition

The clinical course after TKA was evaluated by circumferential leg measurements, measurement of knee range of motion, and recording of the NRS pain score.
The knee joint swelling after total knee replacement was evaluated by circumferential leg measurements on the day before TKA (d-1) and every day after knee surgery until the dismissal day. Girth measurements were taken of the involved and uninvolved legs at mid-patella (MP), 7 cm proximal of mid-patella (MP) and 7 cm distal of mid-patella (DP). At least two measurements were done at each measurement site, and the mean value was used for analysis. The measuring tape used was the Waegener tape measure, whose reliability and agreement were evaluated before in the reproducibility study. This tape measure was chosen, because it had shown a reproducibility level comparable to the Gulick I and standard tape measures, but was elected the most user-friendly of all tape measures evaluated.
For the lower leg circumference measurements, patients lay supine with their knees in full extension and lower extremity musculature relaxed. A firm cylindrical paper roll with a diameter of 16cm and a length of 50 cm was placed underneath the heels of the foot. The patients were told to relax their limbs. All lower extremity girth measurements were conducted according to the described protocol.
The passive range of motion (ROM) was assessed by measurements of the maximum passive flexion and extension possible. The ROM measurements were performed on the day before TKA and daily from the second day after TKA (d2) until the dismissal day. The first day after TKA (d1) was excluded from ROM measurements because of existing drainage. The maximum passive flexion and extension was measured in supine lying patients using the same universal goniometer that was used in the reproducibility measurements.

The knee pain intensity after TKA was quantified utilizing a numerical rating scale (NRS), ranging from 1-10. The NRS score was assessed by questioning the patient concerning maximum and minimum pain within the last twenty-four hours (“What number on a 1 to 10 scale would you give your pain when it is the worst that it gets and when it is the best that it gets?“). The NRS score was recorded on the day before TKA and every day after knee surgery until the dismissal day.

Measurements of circumferences and knee range of motion and recording of the pain intensity by means of the NRS were performed by one single observer (observer O1). Table 6 gives an overview of the performed data acquisition.

Measure	d-1	d1	d2	d3	d4	d5	d6	d7
Girth	x	x	x	x	x	x	x	x
ROM	x		x	x	x	x	x	x
NRS	x	x	x	x	x	x	x	x

Table 6: Data acquisition

2.3 Statistical methods

Statistical analyses were performed using MedCalc for Windows, version 12.7.0 (MedCalc Software, Ostend, Belgium) [88], IBM SPSS Statistics version 20.0 (IBM Corp., Armonk, NY, USA) [87], and Microsoft Office Excel 2010 (Microsoft, Redmond, USA) [91].

The two sources of variability examined in this study were as follows. The variability arising when a single observer made repeated measurements around the knee on the same subject, referred to subsequently as intra-observer repeatability. The variability arising from differences between observers making measurements, referred to subsequently as inter-observer repeatability.

To quantify inter-observer and intra-observer reproducibility, agreement was determined using the Bland and Altman´s method, and reliability was assessed using the intraclass correlation coefficient (ICC).

For agreement, the mean difference mD between two examiners (inter-observer agreement) and between measuring days (inter-observer agreement) and the standard deviation SD_{diff} of these differences was calculated. The magnitude of the SD_{diff} expresses the extent to which the examiners are able to achieve the same value [5]. Subsequently, the 95% limits of agreement were calculated, defined as the mean difference between examiners $\pm 1.96 \cdot SD_{diff}$ of this mean difference [62]. Only differences exceeding the limits of agreement can be interpreted as “real” differences above measurement error [62,89]. Further, the smallest detectable difference was obtained. Because mD was unequal to zero and thus systematic bias present, the SDD was corrected with the absolute value of mD, extending the formula of the SDD according to (8) [45]:

$$\mathrm{SDD} = |\mathrm{mD}| + 1.96 \cdot \mathrm{SD}_{\mathrm{diff}} \tag{8}$$

These corrected SDDs represent the 95% threshold for change that can be detected by the particular device beyond measurement error [45].

Although there are no clear criteria for the acceptable degree of inter- and intra-observer agreement, differences exceeding 1 cm in case of the girth measurements and 10° in case of the goniometric measurements were considered to be low agreement.

Further, the percentage of differences between two measurements within 1 cm in case of the circumferential measurements and 10° in case of the flexion measurements were calculated.

For reliability, the intraclass correlation coefficient (ICC) was derived from a two-way random-effects analysis of variance (ANOVA), corresponding to model 2.1 according to the guidelines specified by Shrout and Fleiss [86], for absolute agreement. The ICCs for intra-observer reliability were calculated by comparing the first and the second measurements taken by each observer, while the ICCs for inter-observer reliability were calculated by comparing the measurements of each observer.

Pearson´s correlation coefficient was computed to establish a possible relationship between circumference change and passive ROM on third (d3) and sixth (d6) postoperative day. Here, a circumference change was defined as follows (9):

$$\text{Circumference change} = \text{Circumference}_{\text{d3 or d4}} - \text{Circumference}_{\text{d}-1} \tag{9}$$

with d-1 representing the circumference measured preoperatively. A P level of 0.05 or less was considered statistically significant.

3 Results

In this chapter, the results of the inter-observer reliability and agreement are presented for the observers O1 and O2 on the first measuring day. The results for the second measuring day, and for the observers O1, O2, O3, O4, and O5 who measured five subjects are available in the appendix (A.1 and 0, respectively). Further, in order to keep the tables for inter-observer and intra-observer agreement as simple and clear as possible, the confidence intervals of the mean difference and of the limits of agreement are not shown in the tables in this chapter. These may be found in the appendix as well (A.3).

3.1 Reliability and agreement of girth measurements

3.1.1 Descriptive statistics

Table 7 shows the descriptive statistics for the circumferential girth measurements of the observers O1 and O2.

		Mean girth ± SD	(Range) (cm)
Observer	O1	37.1 ± 2.3	(29.8 to 45.5)
	O2	37.0 ± 2.3	(30.1 to 45.2)
Leg side	left	37.0 ± 3.3	(29.8 to 45.4)
	right	37.1 ± 3.2	(29.8 to 45.5)
Measuring day	t1	37.1 ± 3.2	(29.8 to 45.5)
	t2	37.0 ± 3.2	(29.8 to 45.4)
Measurement site	PP	39.7 ± 2.6	(34.3 to 45.5)
	MP	37.3 ± 2.1	(37.3 to 41.9)
	DP	34.1 ± 2.2	(29.8 to 38.8)
Tape measure	GI	36.3 ± 3.1	(29.8 to 43.1)
	GII	37.6 ± 3.4	(30.7 to 45.5)
	S	37.5 ± 3.3	(30.9 to 45.2)
	W	36.7 ± 2.9	(30.7 to 43.4)

Table 7: Descriptive statistics for the circumference measurements
SD, standard deviation; PP, measurement site at 7 cm proximal of mid-patella; MP, measurement site at mid-patella; DP, measurement site at 7 cm distal of mid-patella; GI, Gulick I tape measure; GII, Gulick II plus tape measure; S, standard tape measure; W, Waegener tape measure; t1, first measuring day; t2, second measuring day;

Measured circumference differed between observers (P=0.0107), between measuring days (P<0.0001), between leg sides (P=0.0029) and between measurement sites (P<0.0001). There was an increase in circumference as measurements proceeded proximally. The circumferences measured with the different tape measures showed significant differences (P≤0.0112) between all pairs of tape measure (GI-GII, GI-S, GI-W, GII-S, GII-W, S-W). Girth measured with the Gulick II plus and standard tape measures were in mean 0.8 to 1.3 cm larger than with the Gulick I and Waegener tape measures.

3.1.2 Inter-observer reproducibility

The results of the inter-observer agreement and reliability with regard to different measuring positions and tape measures are presented in Table 8.
Across measuring positions and tape measures, the SDD ranged from 0.5 to 2.1 cm, and the ICC ranged from 0.93 to 0.98.

3.1.2.1 Measurement sites

Considering the measurement sites, highest agreement was observed at the site 7 cm distal of mid-patella (SDD range, 0.7 to 1.0 cm) followed by the mid-patella site (SDD range, 0.9 to 1.2 cm) and at the site at 7 cm proximal of mid-patella (SDD range, 0.9 to 2.1 cm). Accordingly, reliability was slightly higher at 7 cm distal of mid-patella (ICC, 0.98) than at mid-patella (ICC range, 0.97 to 0.98) and 7 cm proximal of mid-patella (ICC range, 0.93 to 0.98). Thus, reliability and agreement were influenced by the measurement site in the way that both increased from proximal to distal.

3.1.2.2 Measuring tapes

Having a closer look at the different measuring tapes, the Waegener tape measure showed highest agreement (SDD range, 0.9 to 1.0 cm), followed by the Gulick I tape measure (range, 0.9 to 1.2 cm), and the standard tape measure (SDD range, 0.7 to 1.7 cm). The Gulick II plus tape measure had the lowest agreement (SDD range, 1.0 to 2.1 cm). Correspondingly, reliability was slightly higher for the Waegener tape measure (ICC 0.98) than for the Gulick I tape measure (ICC range 0.97 to 0.98) and the standard tape measure (range, 0.95 to 0.98). The Gulick II tape measure showed lowest reliability (ICC range, 0.93 to 0.98). Thus, the Waegener tape measure showed highest reproducibility, while the Gulick II plus tape measure showed lowest reproducibility.

Site	Tape	O1 (cm)	O2 (cm)	Agreement: O1-O2 (cm)			Reliability
		Mean ± SD	Mean ± SD	mD ± SD_{diff}	LOA	SDD	ICC [95% CI]
PP	GI	39.1 ± 2.4	38.9 ± 2.5	0.2 ± 0.5	-0.9 to 1.2	1.2	0.97 [0.95 to 0.99]
	GII	40.9 ± 2.5	40.5 ± 2.5	0.4 ± 0.8	-1.2 to 2.1	2.1	0.93 [0.84 to 0.97]
	S	40.7 ± 2.5	40.4 ± 2.6	0.3 ± 0.7	-1.0 to 1.7	1.7	0.95 [0.89 to 0.98]
	W	39.1 ± 2.4	39.1 ± 2.4	0.0 ± 0.4	-0.8 to 0.9	0.9	0.98 [0.97 to 0.99]
MP	GI	36.7 ± 2.1	36.7 ± 2.0	0.0 ± 0.5	-0.9 to 1.0	1.0	0.97 [0.95 to 0.99]
	GII	37.8 ± 2.2	37.8 ± 2.1	-0.1 ± 0.6	-1.2 to 1.0	1.2	0.97 [0.94 to 0.98]
	S	37.8 ± 2.2	37.7 ± 2.1	0.2 ± 0.5	-0.7 to 1.1	1.1	0.97 [0.94 to 0.99]
	W	37.1 ± 2.1	37.1 ± 2.0	0.0 ± 0.4	-0.8 to 0.9	0.9	0.98 [0.96 to 0.99]
DP	GI	33.4 ± 2.1	33.6 ± 2.1	-0.2 ± 0.4	-0.9 to 0.5	0.9	0.98 [0.95 to 0.99]
	GII	34.3 ± 2.1	34.5 ± 2.0	-0.2 ± 0.4	-1.0 to 0.6	1.0	0.98 [0.93 to 0.99]
	S	34.5 ± 2.1	34.5 ± 2.0	0.0 ± 0.4	-0.7 to 0.7	0.7	0.98 [0.97 to 0.99]
	W	34.1 ± 2.0	34.3 ± 2.0	-0.2 ± 0.4	-1.0 -to 0.6	1.0	0.98 [0.95 to 0.99]

Table 8: Girth - Inter-observer reproducibility for observers O1 and O2 (n=40 legs)
LOA, limits of agreement; SDD, smallest detectable difference; D ≤1cm, difference between observers O1 and O2 within 1 cm; ICC, intraclass correlation coefficient; CI, confidence interval; n, sample size

These results are displayed graphically in the form of Bland and Altman (B-A) plots for the measurement site 7 cm proximal of mid-patella, mid-patella, and 7 cm distal of mid-patella in Figure 18, Figure 19, and Figure 20, respectively. In these plots, each point represents the difference between observers for each subject´s left and right lower extremities (n=40).

- For the measurement site at 7 cm proximal of mid-patella, the Bland and Altman plots show the highest agreement for the Waegener tape measure, closely followed by the Gulick I tape measure, while the Gulick II plus tape measure had the largest limits of agreement (Figure 18). Further, the Gulick I and Gulick II plus tape measures show one outlier each clearly exceeding the limits of agreement.
- In case of the mid-patella measurement site, the Waegener standard tape measure showed highest agreement, followed by the standard and GI tape measures (Figure 19). The Gulick II plus tape measure, again, had the widest limits of agreement, even though the limits of agreement were substantially smaller at mid-patella than at 7 cm proximal of mid-patella. Each plot shows outliers clearly exceeding the limits of agreement. With the Waegener tape measure, it was possible to fulfil the a priori criterion and detect differences in girth within 1 cm (maximum SDD, 1.0 cm).

- In contrast to the results at the measurement sites at 7 cm proximal of mid-patella and mid-patella, the B-A plot of the Waegener tape measure shows larger limits of agreement at 7 cm distal of mid-patella than the B-A plots of the other tape measures. The B-A plots of the Gulick I and standard tape measures have the smallest limits of agreement. Again, the plots show outliers clearly exceeding the limits of agreement.

Interpreting these results, however, one should bear in mind that the limits of agreement at 7 cm distal of mid-patella are generally smaller than at the other measurement sites. Thus, the limits of agreement of the Waegener tape measure at 7 cm distal of mid-patella (-1.0 to 0.7 cm) are in the range of the limits of agreement measured at 7 cm proximal of mid-patella (-0.8 to 0.9 cm) and at mid-patella (-0.8 to 0.9 cm) with this tape measure.

In summary, the level of inter-observer reproducibility depended on the measurement site and the tape measure used. At 7 cm distal of mid-patella, all tape measures showed good agreement. The results of the Waegener tape measure were reproducible independent of measurement site. The reliability across measurement sites and tape measures was high (mean 0.98).

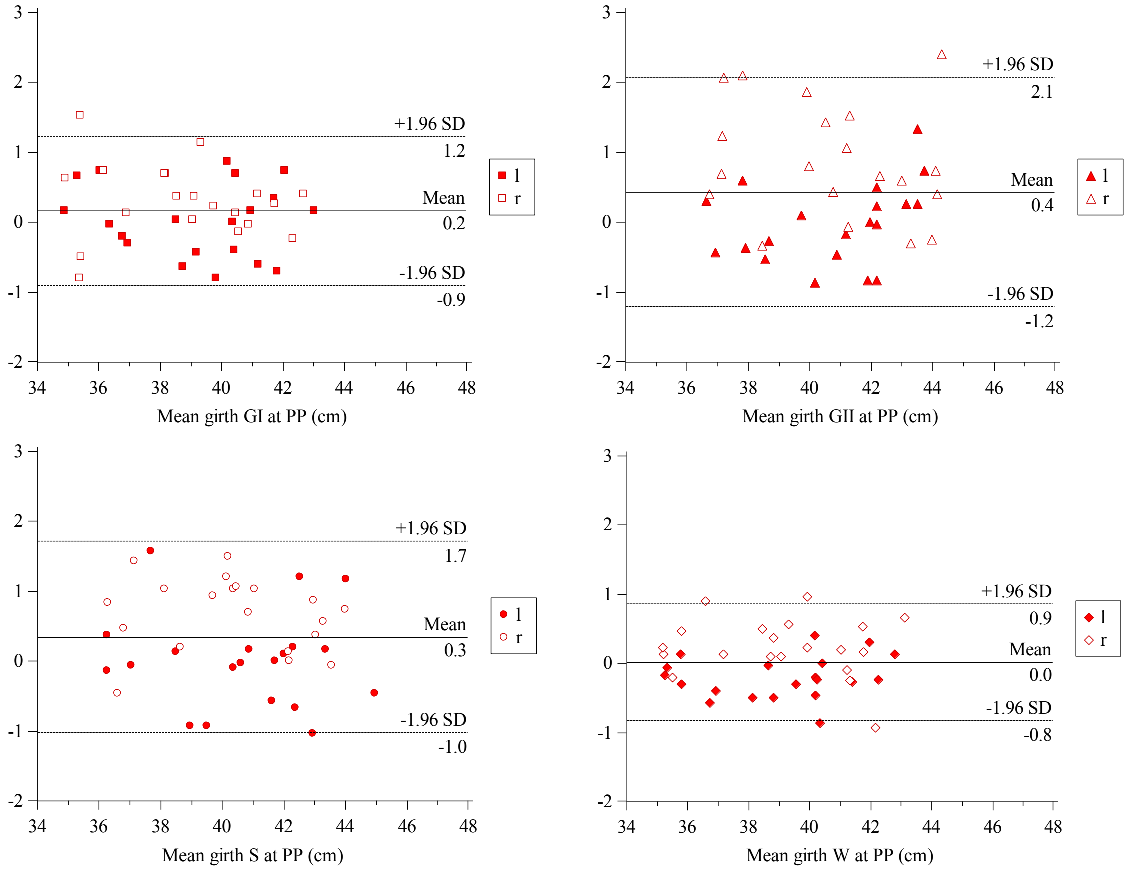


Figure 18: Girth - Inter-observer B-A plots at PP for the observers O1 and O2
with mean difference between observers (solid black line) and limits of agreement (broken black lines); l, left leg; r, right leg; The closer the limits of agreement, the higher agreement between observers for each tape measure. A mean difference differing from zero indicates a systematic bias present.

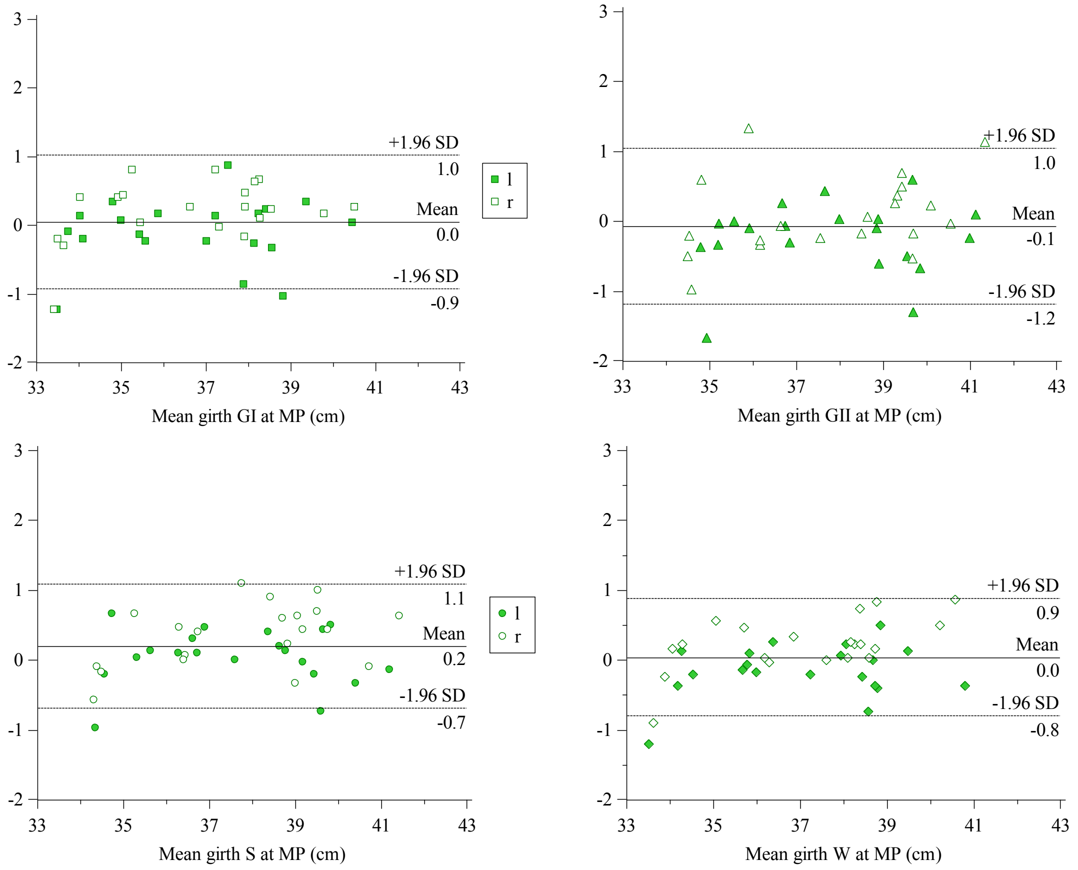


Figure 19: Girth - Inter-observer B-A plots at MP for the observers O1 and O2

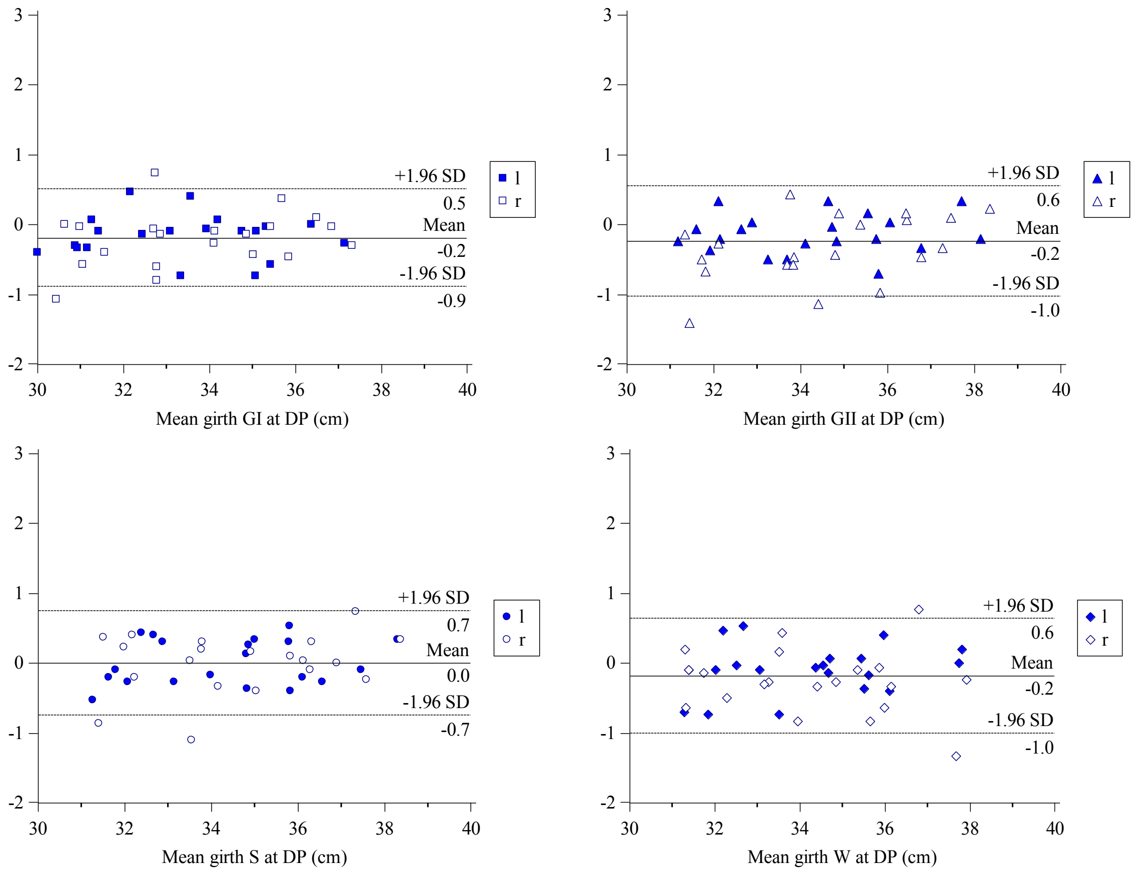


Figure 20: Girth - Inter-observer B-A plots DP for the observers O1 and O2

3.1.3 Intra-observer reproducibility

The results of the intra-observer agreement and reliability of the circumferential girth measurements for the observers O1 and O2 are presented in Table 9. Of the corresponding Bland and Altman plots, only those for the measurement sites at 7 cm proximal of mid-patella are shown in this chapter (Figure 22). The Bland and Altman plots for the measurement sites at mid-patella and 7 cm distal of mid-patella are to be found in the appendix A.1.1.2 (Figure 56, and Figure 57, respectively).

Site	Tape	O1				O2			
		mD ± SD_{diff} (cm)	LOA (cm)	SDD (cm)	ICC [95% CI]	mD ± SD_{diff} (cm)	LOA (cm)	SDD (cm)	ICC [95% CI]
PP	GI	0.1 ± 0.6	-1.1 to 1.3	1.3	0.97 [0.93 to 0.98]	0.1 ± 0.6	-1.2 to 1.3	1.3	0.97 [0.93 to 0.99]
	GII	0.1 ± 0.8	-1.4 to 1.7	1.7	0.95 [0.90 to 0.97]	-0.1 ± 0.8	-1.6 to 1.4	1.6	0.96 [0.91 to 0.98]
	S	0.2 ± 0.5	-0.9 to 1.2	1.2	0.98 [0.95 to 0.99]	0.0 ± 0.6	-1.1 to 1.1	1.1	0.98 [0.95 to 0.99]
	W	0.1 ± 0.5	-0.9 to 1.1	1.1	0.98 [0.96 to 0.99]	0.1 ± 0.6	-0.9 to 1.2	1.2	0.98 [0.95 to 0.99]
MP	GI	0.1 ± 0.4	-0.7 to 0.9	0.9	0.98 [0.96 to 0.99]	0.0 ± 0.4	-0.8 to 0.9	0.9	0.98 [0.96 to 0.99]
	GII	0.0 ± 0.5	-0.9 to 1.0	1.0	0.98 [0.95 to 0.99]	0.0 ± 0.5	-1.1 to 1.0	1.1	0.97 [0.94 to 0.98]
	S	0.0 ± 0.5	-1.1 to 1.0	1.1	0.97 [0.94 to 0.98]	0.0 ± 0.4	-0.9 to 0.9	0.9	0.98 [0.96 to 0.99]
	W	0.0 ± 0.5	-1.0 to 1.0	1.0	0.97 [0.94 to 0.99]	0.1 ± 0.4	-0.8 to 0.9	0.9	0.98 [0.96 to 0.99]
DP	GI	0.1 ± 0.5	-0.8 to 1.1	1.1	0.98 [0.95 to 0.99]	0.2 ± 0.4	-0.6 to 1.0	1.0	0.98 [0.94 to 0.99]
	GII	0.1 ± 0.5	-0.9 to 1.2	1.2	0.97 [0.94 to 0.98]	0.2 ± 0.5	-0.8 to 1.3	1.3	0.97 [0.92 to 0.98]
	S	0.0 ± 0.5	-0.9 to 1.0	1.0	0.98 [0.95 to 0.99]	0.2 ± 0.4	-0.7 to 1.0	1.0	0.98 [0.95 to 0.99]
	W	0.1 ± 0.5	-0.8 to 1.1	1.1	0.97 [0.95 to 0.99]	0.1 ± 0.5	-0.8 to 1.1	1.1	0.97 [0.95 to 0.99]

Table 9: Girth - Intra-observer reproducibility for observers O1 and O2

3.1.3.1 Observer O1

For observer O1, the SDD ranged from 0.9 to 1.7 cm across the three measurement sites and four tape measures (Table 9). The ICC ranged from 0.95 to 0.98.

- Considering the different measurement sites, agreement was higher at mid-patella (SDD range, 0.9 to 1.0 cm) than at 7 cm distal of mid-patella (SDD range, 1.0 to 1.2 cm). As for inter-observer agreement, the intra-observer agreement was lowest at 7 cm proximal of mid-patella (SDD range, 1.1 to 1.7 cm). Accordingly, reliability was higher at mid-patella and 7 cm distal of mid-patella (ICC range, 0.97 to 0.98) than at 7 cm proximal of mid-patella (ICC range, 0.95 to 0.98. Thus, the measurement site at 7 cm proximal of mid-patella showed the lowest level of intra-observer reproducibility.

- Taking a closer look at the different measuring tapes, highest agreement was found for the Waegener tape measure (SDD range, 1.0 to 1.1 cm), followed by the standard tape measure (SDD range, 1.0 to 1.2 cm) and the Gulick I tape measure (SDD range, 0.9 to 1.3 cm). Again, the Gulick II plus tape measure showed lowest agreement (SDD range, 1.0 to 1.7 cm). Accordingly, reliability was higher for the Gulick I, and standard tape and Waegener tape measures (ICC range, 0.97 to 0.98) than for the Gulick II plus tape measure (ICC range, 0.95 to 0.97). Thus, intra-observer reproducibility was lowest for the Gulick II plus tape.

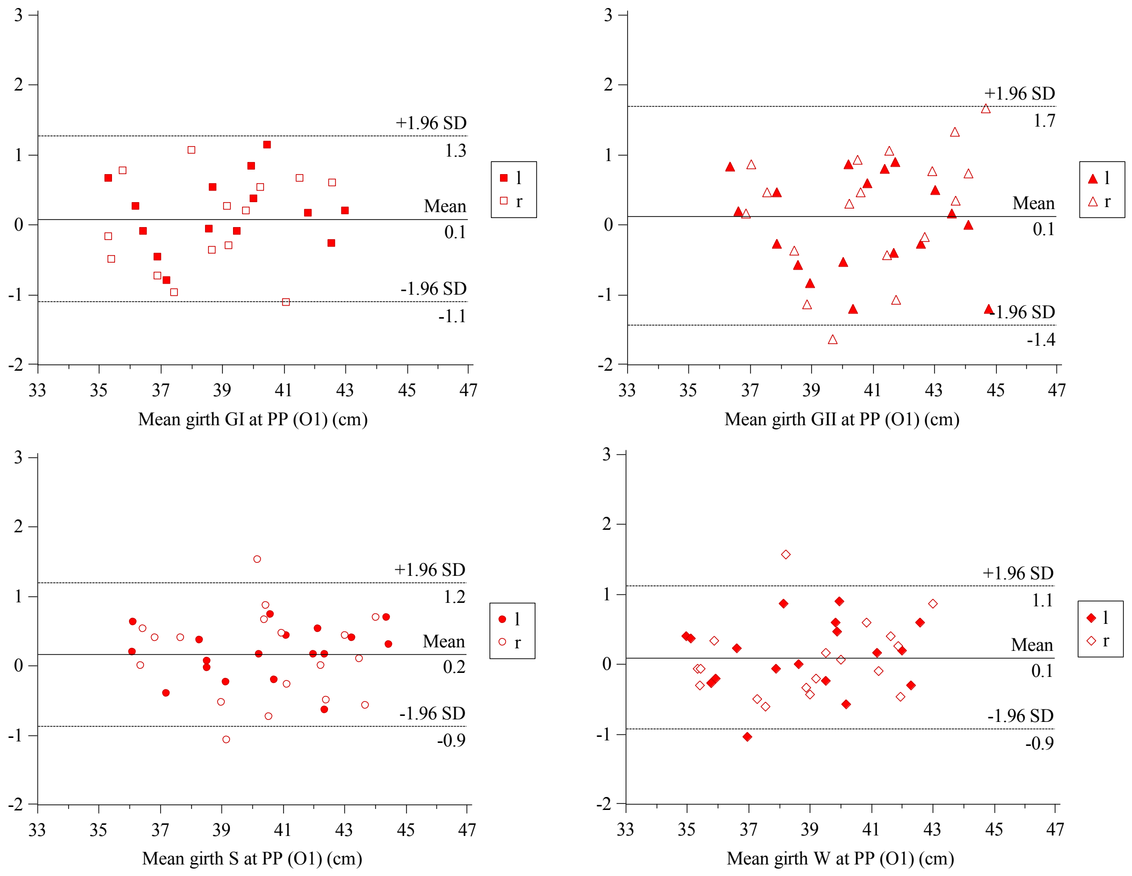


Figure 21: Girth - Intra-observer B-A plots at PP for the observers O2

3.1.3.2 Observer O2

Observer O2 showed slightly higher level of reproducibility than observer O1 (SDD range, 0.9 to 1.6 cm and ICC range, 0.96 to 0.98).

- Considering measurement sites, reproducibility was higher at the mid-patella site (SDD range, 0.9 to1.1 cm and ICC range, 0.97 to 0.98) than at 7 cm distal of mid-patella

(SDD range, 1.0 to 1.3 cm and ICC range, 0.97 to 0.98) and 7 cm proximal of mid-patella (SDD range, 1.1 to 1.6 cm and ICC range, 0.96 to 0.98).

- Observer O2 performed most reproducible with the standard tape measure (SDD range, 0.9 to 1.1 cm and ICC 0.98), followed by the Waegener tape measure (SDD range, 0.9 to 1.2 cm and ICC range, 0.97 to 0.98) and the Gulick I tape measure (SDD range, 0.9 to 1.3 cm and ICC range, 0.97 to 0.98). Again, the Gulick II plus provided least reproducible results (SDD range, 1.1 to 1.6 cm and ICC range, 0.96 to 0.97).

Summarising these results, the level of intra-observer reproducibility depended on the measurement site and the tape measure used. Both observers performed the 'least best' at 7 cm proximal of mid-patella and with the Gulick II plus tape measure. Observers were not able to fulfil the a priori criterion (difference between measuring days ≤1cm) with neither tape measure.

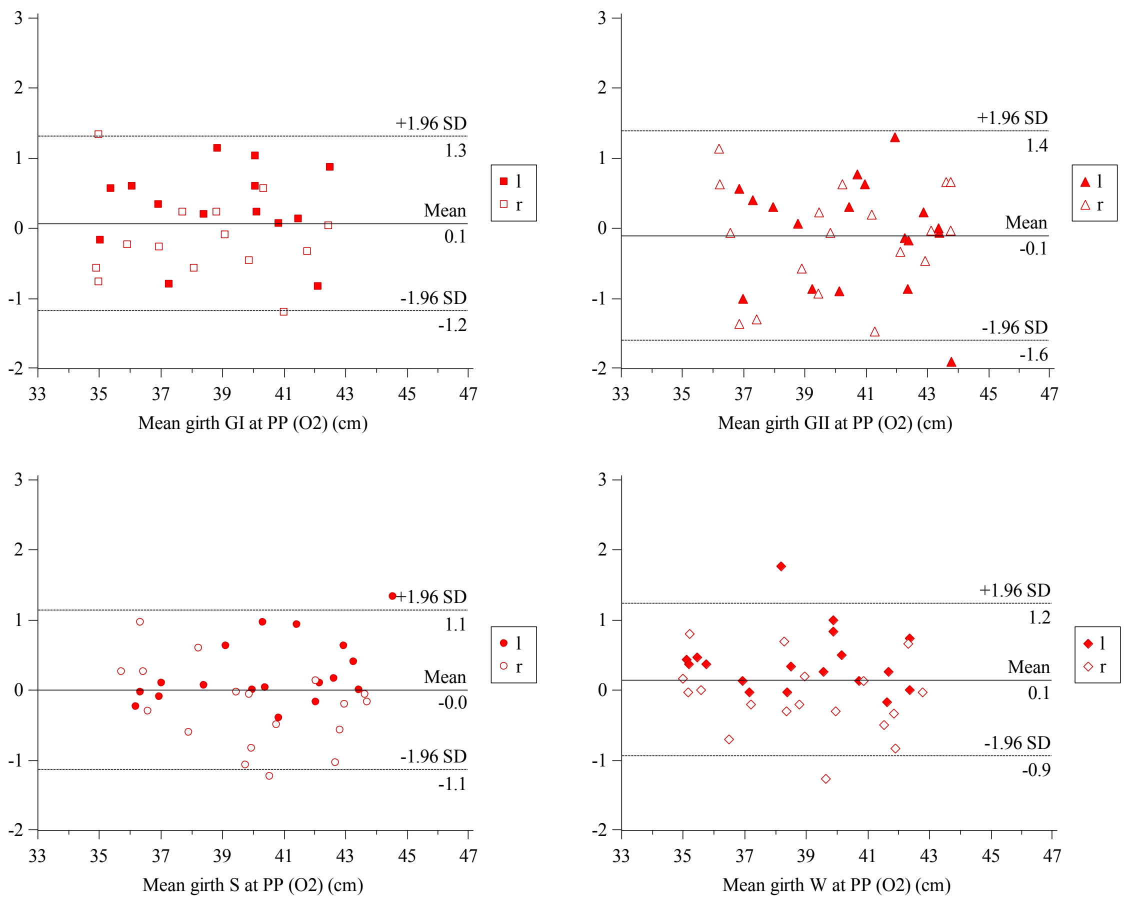


Figure 22: Girth - Intra-observer B-A plots at PP for the observers O2

3.2 Reliability and agreement of knee flexion measurements

3.2.1 Descriptive statistics

Table 7 shows the descriptive statistics of the goniometric measurements for the observers O1 and O2 and both legs. Flexion values were comparable between observers (P=0.1344) and leg sites (P=0.1020). Mean flexion was higher on second measuring day than on first measuring day (P<0.0001). Further, mean flexion values in test position P1 were higher than in test position P2 (P<0.0001). For further analysis, both legs were considered as independent entities.

		Mean flexion ± SD	(Range) (°)
Measuring day	t1	97.0 ± 12.7	(64.0 to 134.5)
	t2	98.9 ± 19.7	(68.0 to 138.0)
Leg side	left	97.9 ± 19.9	(64.0 to 138.0)
	right	98.2 ± 19.5	(64.0 to 138.1)
Observer	O1	98.3 ± 20.3	(64.0 to 138.0)
	O2	97.9 ± 19.1	(70.0 to 132.0)
Test position	P1	112.4 ± 17.5	(90.0 to 138.0)
	P2	83.7 ± 7.6	(64.0 to 102.0)

Table 10: Descriptive statistics for the goniometric measurements

3.2.2 Inter-observer reproducibility

Table 11 summarizes the results of the inter-observer reliability and agreement analysis for the observers O1 and O2 and 19 subjects. Figure 23 shows the corresponding Bland and Altman plots.

Position	O1	O2	Agreement: O1-O2			Reliability
	Mean ± SD (°)	Mean ± SD (°)	mD ± SD_{diff} (°)	LOA (°)	SDD (°)	ICC [95% CI]
P1	112.0 ± 18.1	111.1 ± 17.1	1.0 ± 2.5	-4.0 to 5.9	5.9	0.99 [0.98 to 0.99]
P2	82.4 ± 8.3	85.1 ± 8.8	-1.1 ± 3.6	-8.2 to 6.0	8.2	0.88 [0.77 to 0.93]

Table 11: Flexion – Inter-observer reproducibility for observers O1 and O2 (n=38 legs)
P1, first position; P2, second position

The SDD ranged from 5.9 to 8.2°, and the ICCs ranged from 0.88 to 0.99. Considering the different measuring positions, inter-observer agreement was higher for knee position P1 than for P2. This was also reflected by the ICC, which was 0.99 for knee position P1, and 0.88 for knee position P2. Accordingly, the Bland and Altman plots in Figure 23, show smaller limits of agreement for knee position P1. Further, there is an outlier exceeding a mean difference of 10° for test position P2.

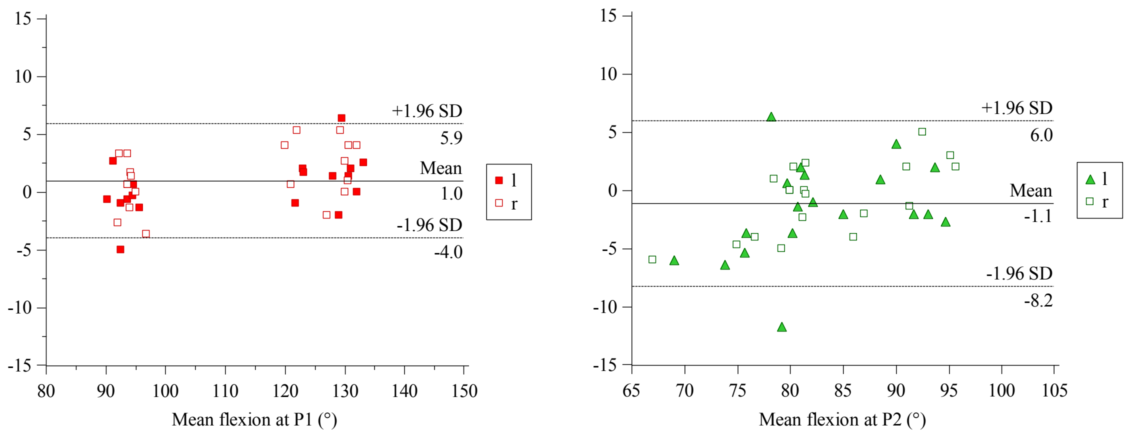


Figure 23: Flexion - Inter-observer B-A plots for the observers O1 and O2
with mean difference between observers (solid black line) and limits of agreement (broken black lines);

3.2.3 Intra-observer reproducibility

Table 12 summarizes the intra-observer agreement and reliability of knee flexion measurements for the observer O1 and O2. Figure 24 shows the corresponding Bland and Altman plots.

The SDD ranged from 7.1° to 8.1°, and the ICC ranged from 0.87 to 0.99. Considering the different positions, there was hardly any difference in agreement between knee position P1 and P2 (SDD range, 7.1 to 8.1° and 7.2 to 7.7°. However, reliability was higher for position P1 than P2 (0.98, and range, 0.87 to 0.90, respectively).

Having a closer look at differences in agreement between the observers O1 and O2 shows that observer O1 had slightly smaller SDDs than observer O2 (range, 7.1 to 7.7°, and 7.2 to 8.1°, respectively). This was also reflected by the ICC values, which were slightly higher for observer O1 than O2 (range, 0.90 to 0.98, and 0.87 to 0.98, respectively).

Comparing the SDD values with the ICC values in Table 12 shows that there are certain inconsistencies in the statistical results. The lowest agreement was found for observer O2 at right knee position P1 (SDD 8.1°). However, the corresponding ICC value was 0.98, suggesting high reliability. In contrary, the second lowest agreement was found at left knee position P2 for observer O1 (SDD 7.7 °). In this case, the ICC was 0.90, which appears to be a reasonable value. These unexpected results will be considered in more detail in the discussion chapter (4.3).

Position	O1				O2			
	mD ± SD_{diff} (°)	LOA (°)	SDD (°)	ICC [95% CI]	mD ± SD_{diff} (°)	LOA (°)	SDD (°)	ICC [95% CI]
P1	-1.8 ± 2.7	-7.1 to 3.6	7.1	0.98 [0.95 to 0.99]	-1.9 ± 3.2	-8.1 to 4.3	8.1	0.98 [0.93 to 0.99]
P2	-3.2 ± 2.3	-7.7 to 1.2	7.7	0.90 [0.15 to 0.97]	-0.8 ± 3.3	-7.1 to 5.6	7.1	0.87 [0.76 to 0.93]

Table 12: Flexion – Intra-observer reproducibility for observers O1 and O2 (n=34 legs)

Another interesting fact is that the mean difference between measuring days was negative for both observers. This negative mean difference suggests that both observers measured higher flexion values on the second measuring day.

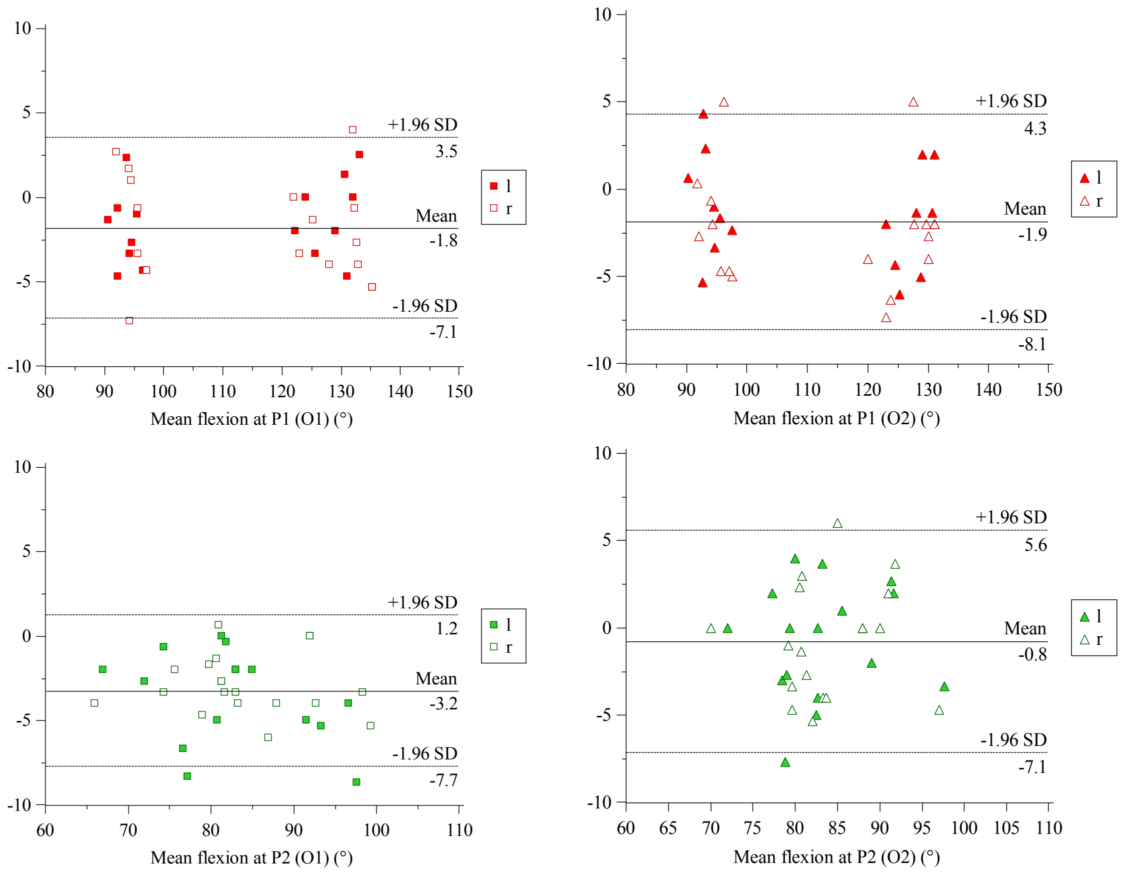


Figure 24: Flexion - Intra-observer B-A plots for the observers O1 and O2 at knee positions P1 and P2 with mean difference (solid black line) and limits of agreement (broken black lines)

In summary, the intra-observer reproducibility was low or, at best, acceptable. However, the differences between measurements did not exceed the a priori criterion of 10°. The statistical measures ICC and SDD showed inconsistencies, which have to be further discussed (4.3).

3.3 Clinical Course after TKA

In this chapter, the presented results refer to mean values obtained from the data of 29 patients evaluated for changes in girth, passive ROM and NRS.

3.3.1 Changes in lower limb girth

Figure 25 shows the chronological course of the lower extremity girth after TKA surgery at the three measurement sites 7 cm proximal of mid-patella (PP), mid-patella (MP) and 7 cm distal of mid-patella (DP) of the operated lower extremity.

After operation, there was an increase of girth at all measurement sites. The site 7 cm proximal of mid-patella showed the highest increase in girth, with in mean 5.1 cm (range, 2.3 to 7.6 cm). The mean increase at the mid-patella site and at 7 cm distal of midpatella was smaller (3.8 cm, range, 1.9 to 9.8 cm, and 2.8 cm, range, 1.7 to 7.2 cm, respectively).

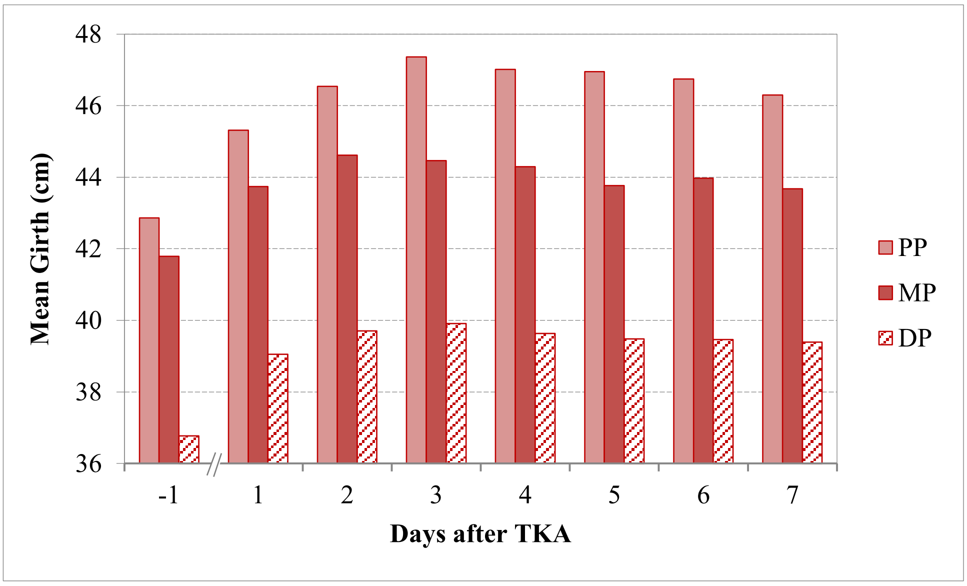


Figure 25: Changes in mean girth of the operated lower leg at the three measurement sites; PP, 7 cm proximal of mid-patella; MP, mid-patella; DP, 7 cm distal of mid-patella

For the measurement sites 7 cm proximal and distal of mid-patella, the maximum girth was reached 4.0 days after surgery (range, 2 to 11 days), and slightly earlier at mid-patella (mean 3.6 days after TKA, range, 1 to 11 days). In one patient, the maximum girth was measured on the dismissal day, thus one can not be sure if this was the maximum, actually.

Excluding this patient from analysis, the maximum girth was reached earlier (mean 3.7, 3.4 and 3.8 days after TKA for PP, MP, and DP, respectively).

Figure 26 shows the chronological lower extremtity circumference change of the operated leg (OP) and the contralateral, uninvolved leg (CL) for the three different mesasuring sites. It can be seen that not only the girth of the operated leg, but also the girth of the uninvolved leg showed changes in girth over time. The circumference of the uninvolved leg showed a decrease during hospitalisation. Taking a closer look at the mid-patella site, the course of the curves of the operated and the contralateral leg seem to be similar, having peaks at the same timepoints of the x-axis. In the girth curves at 7 cm proximal and distal of mid-patella, this course was only indicated by a small peak on the second postoperative day.

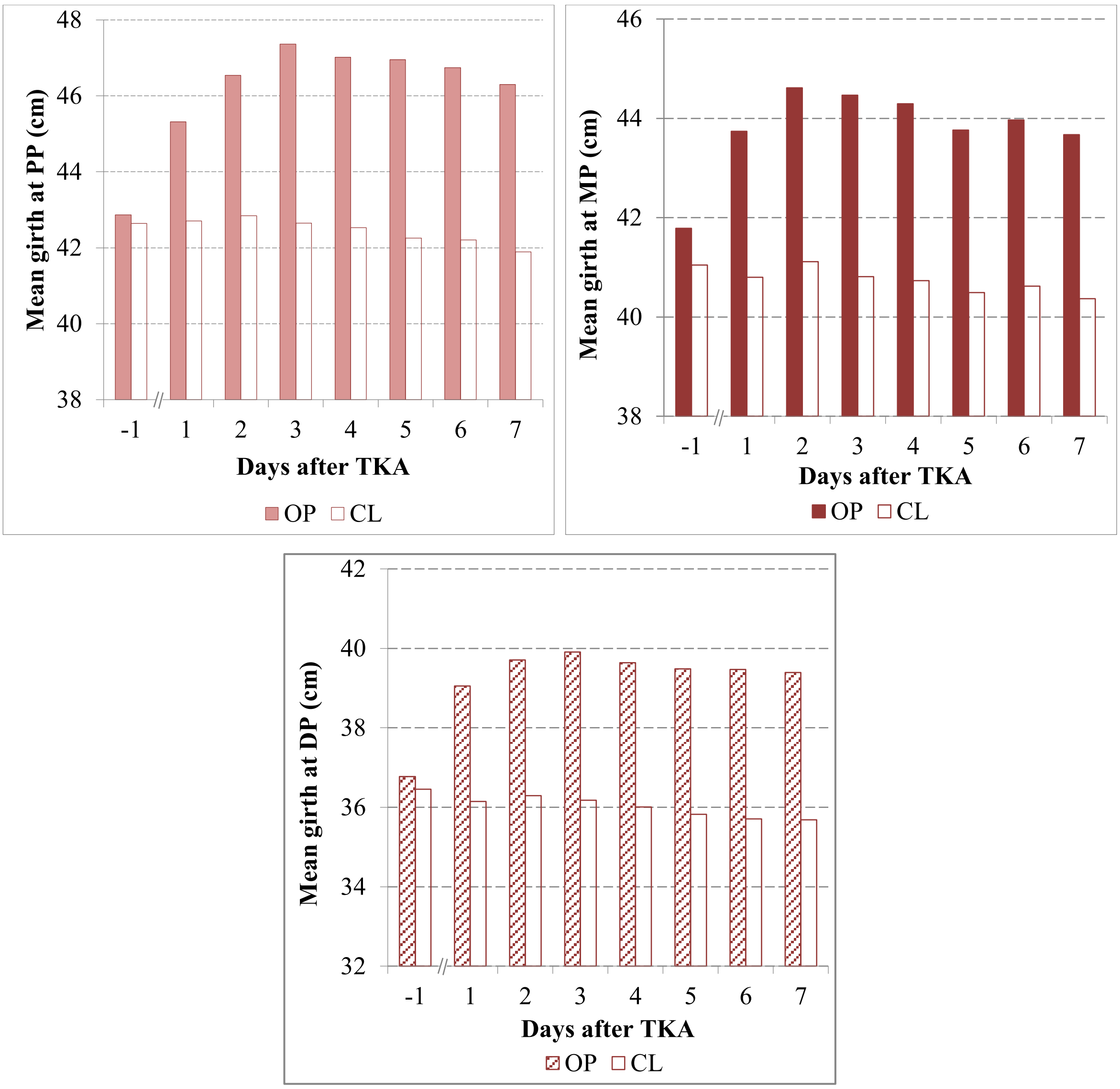


Figure 26: Changes in mean girth of the operated and the contralateral leg
OP, operated leg; CL, contralateral leg

3.3.2 Changes in passive range of motion

The postoperative change of passive knee ROM after TKA surgery is shown in Figure 27. After surgery, the mean passive knee ROM decreases from 117.1° (range, 80 to 135°) preoperatively to 54.5° (range, 30 to 75°) on the second postoperative day. However, during hospital stay, the mean passive ROM increases again from day to day, reaching 79.0° (range, 55 to 100°) on the sixth day after TKA operation.

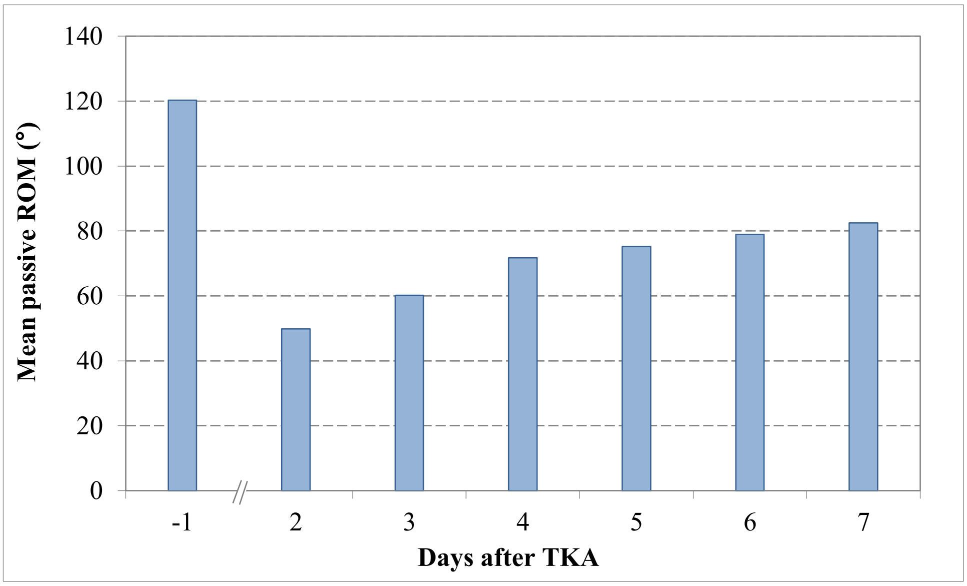


Figure 27: Changes in mean passive ROM
ROM, range of motion

3.3.3 Changes in pain intensity (NRS)

Figure 28 shows the chronological changes in minimum and maximum pain in the form of the NRS reported by the patients.

Interestingly, the maximum pain was stated preoperatively (mean 7.0, range 4 to 9), and not, as was expected, postoperatively. The reason for this might be the fact that all patients had patient-controlled analgesia for 72 hours postoperatively. Further, pain is a primary indication for TKA, while pain release is a primary goal.

After surgery, the maximum pain decreased from the first postoperative day (mean NRS 6.6, range, 3 to 9) to the dismissal day (mean NRS 3.0, range 1 to 9). The minimum NRS curve showed a similar course as the maximum NRS curve, except for a very small increase on the first postoperative day, indicating that the minimum pain increased for a short period after surgery.

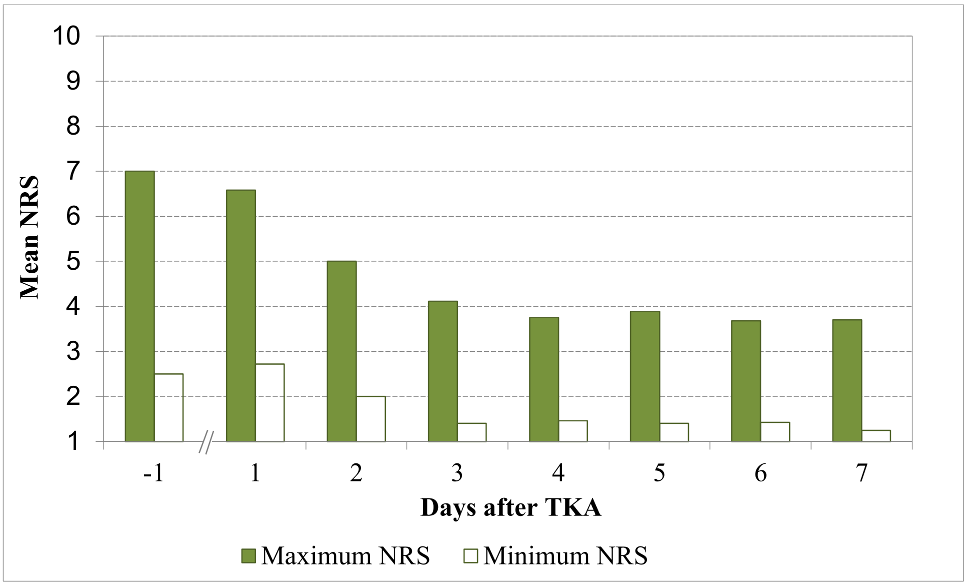


Figure 28: Changes in mean minimum and maximum NRS
NRS, numerical rating scale;

3.3.4 Relationship between girth, ROM and pain changes

In the following figures, the changes in girth and passive ROM (Figure 29), girth and NRS pain scale (Figure 30), and passive ROM and reported NRS pain scale (Figure 31) are compared. Because the lower leg circumferences at the different measurement sites changed in a similar way, only the results of the measurement site at 7 cm proximal of mid-patella will be presented in this chapter. Figures comparing the circumferential changes at mid-patella and at 7 cm distal of mid-patella with passive ROM and the NRS are to be found in the appendix (A.4).

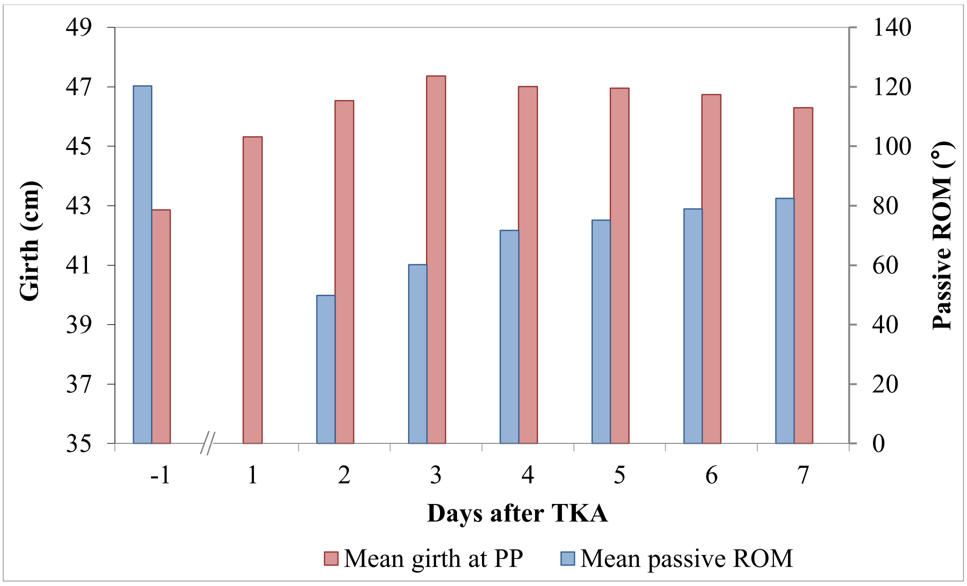


Figure 29: Clinical course - Girth at PP and passive ROM

A comparison of girth at 7 cm proximal of mid-patella and the passive ROM showed an opposite course of the curves (Figure 29). Girth increased up to a maximum on the third postoperative day, from which it decreased slowly until the dismissal day. The passive ROM first decreased after surgery, but kept increasing in the course of inpatient stay. Figure 30 shows the course of girth and mean maximum reported NRS. On the fifth postoperative day, there was a small peak in the NRS curve. The girth curve showed a corresponding slower decrease on the same day.

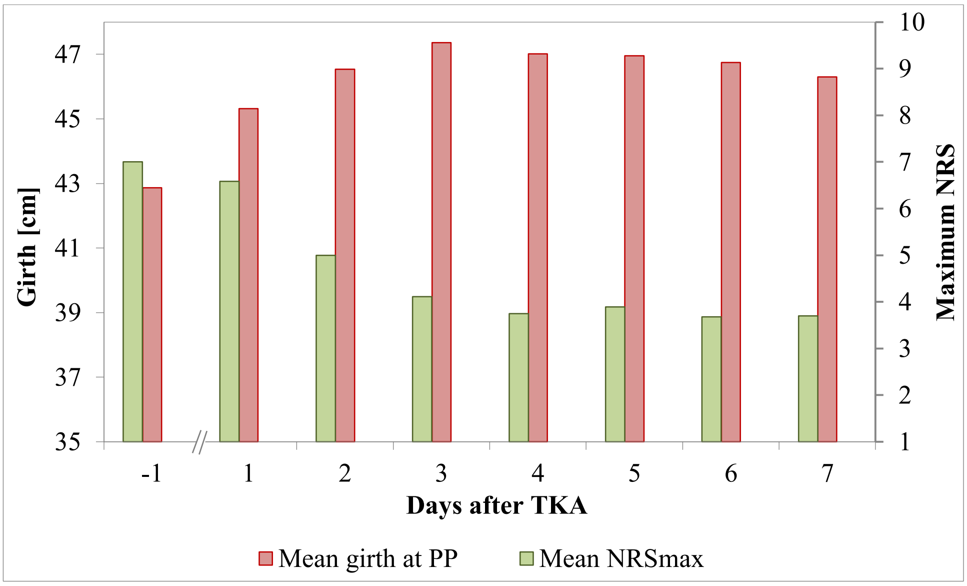


Figure 30: Clinical course - Girth at PP and maximum reported NRS

The curves of the passive ROM and reported NRS are compared in Figure 31. While the reported NRS kept decreasing from the beginning, the passive ROM kept increasing.

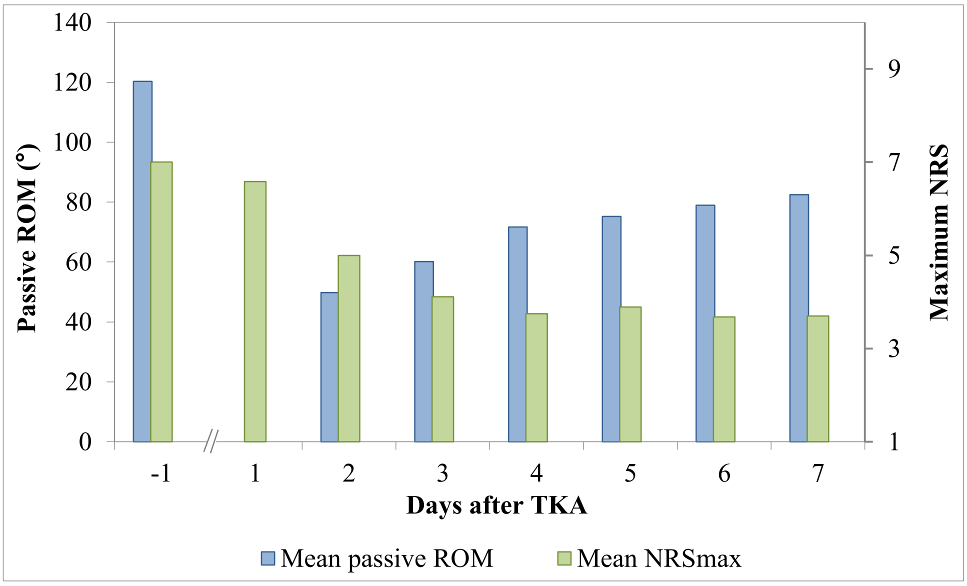


Figure 31: Clinical course – Passive ROM and maximum NRS

The Pearson´s correlation coefficient between the circumference change on the third and sixth postoperative and the corresponding passive ROM showed no significant correlation between swelling and passive ROM (P≥0.1375).

3.3.5 Postoperative follow up examination six weeks after TKA

Girth measurements of the operated leg and passive ROM measurements were performed in 10 of the 29 patients during six weeks follow up check. Table 13 shows the differences in girth and passive ROM between this postoperative 6-weeks check and the day of dismissal ("6 weeks-dismissal"). Further, the girth and ROM values were compared to those obtained preoperatively ("preoperative - 6 weeks check").

Patient No.	Difference (6 weeks check - dismissal)				Difference (preoperative – 6 weeks check)			
	Girth (cm)			PROM (°)	Girth (cm)			PROM (°)
	PP	MP	DP		PP	MP	DP	
2	0.1	-0.5	0.2	25	0.4	1.9	1.2	-14
5	-2.5	-1.3	-1.3	-5	2.0	0.8	2.2	-42
8	-0.5	0.5	-0.9	25	1.7	2.1	2.3	-15
11	-7.1	-4.6	-5.1	30	2.1	0.9	0.6	-6
14	-3.3	-1.3	-1.9	15	1.5	-0.4	1.0	-19
17	-2.8	-2.4	-1.7	20	2.1	0.8	0.8	-21
19	-1.3	0.1	-0.7	35	1.9	1.4	0.9	3
22	-2.4	-1.5	-0.1	30	3.3	4.2	3.6	-13
24	-1.2	-3.7	0.2	55	-0.2	-2.8	0.6	14
29	0.4	1.0	1.0	25	3.4	2.6	1.4	-13

Table 13: 6-weeks check: Changes in girth and PROM
compared to day of dismissal and compared to values obtained preoperatively; PROM;, passive range of motion;

There was a mean decrease in girth of -2.1 cm (range, -7.1 to 0.4 cm) at 7 cm proximal of mid-patella, -1.4 cm (range, -4.6 to 1.0 cm) at mid-patella, and -1.0 cm (range, -5.1 to 1.0 cm) at 7 cm distal of mid-patella between the day of dismissal and the six weeks check. The passive ROM showed a mean increase of 25.5° (range, -5 to 35°). Considering the values presented above, the negative values mean a decrease, while the positive values mean an increase in girth or passive ROM. Thus, neither did swelling decrease in all patients examined, nor passive ROM increase. Table 13 further shows that the lower extremity circumferences were higher at the six weeks follow-up check than preoperatively in most of the cases, suggesting that there was still swelling present six weeks after surgery. Comparing the passive ROM before surgery and six weeks after, it is seen that patients did not reach preoperative passive ROM within six weeks postoperatively (mean -2.1, range -7.1 to 0.4 °), except for two patients (patient No. 2 and patient No. 29).

3.3.6 Influence of gender and BMI on postoperative swelling

To consider a possible influence of the BMI on limb swelling after TKA, the patients were divided into BMI <30 kg/m^2 and a group with BMI ≥30 kg/m^2 groups. Ten of the 29 patients (34%) had a BMI ≥30 kg/m^2. The results of the comparisons in Table 14 show that patients with a BMI less than 30 kg/m^2 reached the maximum girth earlier than adipose patients. At 7 cm proximal of mid-patella, there was no difference in maximum swelling, which was defined as the difference of maximum girth and minimum girth, i.e. preoperatively. However, at mid-patella and 7 cm distal of mid-patella, the maximum swelling was higher for the adipose group.
Having a closer look at possibly existing differences due to gender, the results of female patients were compared with those of male patients. Male patients reached the maximum swelling earlier than the females (Table 14). Further, at mid-patella and 7 cm distal of mid-patella, female patients showed higher maximum swelling.

Factor		Day of maximum swelling			Mean maximum swelling (cm)		
		PP (range)	MP (range)	DP (range)	PP (range)	MP (range)	DP (range)
BMI (kg/m^2)	<30	3.8 (2-6)	3.5 (1-8)	3.8 (2-8)	5.1 (2.6-7.6)	3.4 (1.9-5.3)	3.6 (1.7-5.4)
	≥30	4.4 (2-11)	3.8 (2-11)	4.4 (2-11)	5.1 (2.3-7.2)	4.4 (2.1-9.8)	4.0 (1.8-7.2)
Gender	f	4.4 (2-11)	3.7 (1-11)	4.3 (3-11)	5.0 (2.3-7.0)	4.3 (2.4-9.8)	3.9 (1.8-7.2)
	m	3.7 (2-6)	3.6 (2-8)	3.8 (2-8)	5.2 (2.6-7.6)	3.4 (1.9-5.6)	3.6 (1.7-5.4)

Table 14: Effects of BMI and gender on lower extremity swelling after TKA
Maximum swelling = Difference of maximum girth and minimum girth; f, female; m, male;

Summarizing these results, one could speculate that the BMI and the sex have an influence on postoperative swelling. However, taking a closer look at the sex distribution in the BMI groups, it shows that in the adipose group there are six females and four males, while in the group with BMI <30 kg/m^2 there are thirteen males and six females. For this reason, the sex related differences sex might be due to the fact that more female than male patients in the study population were adipose. Further, a statistical analysis of these data with appropriate methods would be necessary in order to make a statement on the impact of gender and BMI on postoperative swelling.

3.3.7 Subjective judgment of knee swelling vs. girth measurements

Questions concerning an existing swelling are included in many outcome questionnaires, e.g., the Knee Injury and Osteoarthritis Outcome Score (KOOS), Lysholm Knee Scoring Scale, and the Knee Outcome Survey - Activities of Daily Living (KOS-ADL). In the course of data acquisition in this study the question arose whether patients are able to assess changes in postoperative swelling. Therefore, patients were asked if the swelling in the knee region had increased or decreased within the last 24 hours prior to the knee girth measurements. These answers were compared with the results of the girth measurements at the mid-patella measurement site according to (10):

$$\text{Change in girth} = \text{Girth}_{\text{yesterday}} - \text{Girth}_{\text{today}} \tag{10}$$

Only differences in girth of 0.5 cm or greater were considered real changes in swelling. “I don´t know” answers were excluded. A total of 70 patient answers were analysed. The results showed that the subjective judgment of the patients regarding knee swelling and the measurement results did only agree in 49%. In other words, in more than half of the cases, patients were not able to properly assess changes in knee swelling. One patient even reported no swelling although there was an increase in girth of 2.8 cm at 7 cm proximal of mid-patella, 1.3 cm at mid-patella, and 1.0 cm at 7 cm distal of mid-patella, compared with the preoperative circumference values.

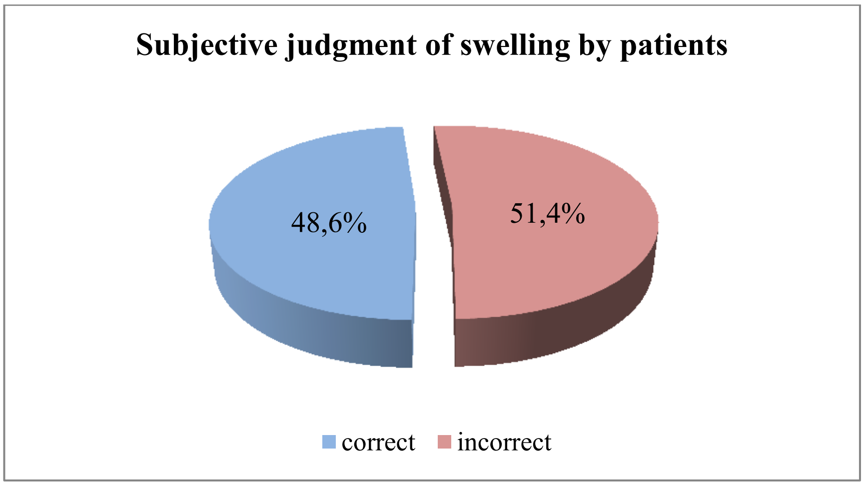


Figure 32: Subjective judgment of swelling in the knee region by patients

4 Discussion

The primary purpose of this study was to evaluate the reliability and reproducibility of circumferential tape measurements with four different types of tape measures using intra- and inter-observer coefficients. The secondary purpose was to describe the clinical course of the swelling in the knee joint region/area after/following total knee arthroplasty. The main findings of this study were

1) The level of reproducibility of circumferential measurements differed substantially dependent on the measuring position and tape measure used. With the Waegener tape measure, differences in girth exceeding 1.2 cm can be considered a real change above measurement error;
2) Knee flexion changes above 9° seems to detect a real change;
3) After TKA, swelling in the knee region was observed in all patients, being highest at 7 cm proximal of mid-patella. Mean passive ROM on the day of dismissal was 81.7°. Reported pain intensity was highest preoperatively.

4.1 Discussion of the girth reproducibility measurements

4.1.1 Discussion of the results

Based on the results of inter- and intra-observer agreement, the smallest detectable differences would lie between 0.4 cm and 2.1 cm, clearly exceeding the a priori criterion of 1 cm. Inter-observer agreement was slightly higher for the second measuring day. Further, observers O1 and O2 showed comparable results in intra-observer comparisons. Reliability was generally high, with ICC values ranging from 0.93 to 0.99.

Measurements	SDD (cm)	ICC
total	0.7 to 2.1	0.93 to 0.99
t1: O1-O2	0.7 to 2.1	0.93 to 0.98
t2: O1-O2	0.7 to 1.8	0.94 to 0.99
O1: t1-t2	0.9 to 1.7	0.95 to 0.98
O2: t1-t2	0.9 to 1.6	0.96 to 0.98

Table 15: Girth - Summary of the results (observers O1 and O2)

Intra-tester reproducibility for observers O1 and O2 was higher than inter-tester reproducibility. This is in agreement to previous reports on circumferential measurements of the lower extremity [39,67,72,92], which found higher intra-observer reliability than inter-observer reliability.

In literature, there is no agreement on reproducibility of lower extremity circumference measurements with tape measures. Whitney et al. (1995) stated that girth measurements in the clinic can be highly repeatable (ICC 0.91 to 1.00) in experienced clinicians by using a simple standardized procedure [73]. In their study, girth of thirty subjects was assessed at five different lower extremity sites by two experienced physical therapists with a standard tape measure. Harrelson et al. (1998) performed lower extremity circumference measurements in twenty-one subjects at three measurement sites with a standard tape measure and a Lufkin tape measure with a Gulick spring-loaded handle attached and reported high reliability (ICC 0.98 to 0.99) [67]. Others evaluated circumferential measurements of the involved and uninvolved legs in nine patients recovering from anterior cruciate ligament reconstructive surgery [72]. They stated that “the measurements established sufficiently high reliability to justify their use both within and between examiners for subjects recovering from surgery of the anterior cruciate ligament” (ICC 0.82 to 1.0 and 0.72 to 0.97, respectively). This is in agreement with te Slaa et al. (2011), who concluded that “tape measurements have been proved to be a reliable and reproducible method to assess the lower limb circumference” [39]. Jakobsen et al (2010) examined 19 outpatients having received a TKA reported that circumference measurements were generally reliable (ICC 0.98 to 0.99) and that changes in “knee joint circumference of more than 1.0 cm and 1.63 cm represent a real clinical improvement (SRD) or deterioration for a single individual within and between physiotherapists, respectively” [4].

Maylia et al. (1999), in contrast, stated that the degree of inaccuracy of tape measurements of the thigh is sufficient to indicate that it is of little value in the assessment of the lower limb. They concluded that the technique is not a reliable method of monitoring the rehabilitation of a patient [93].

The results vary considerably across studies and estimates are difficult to compare due to differences in study design, subject groups, observers, and measurement methods.

4.1.1.1 A priori criterion

Prior to the study, the maximum clinically acceptable SDD was set 1 cm for girth measurements, because it was felt that a change in girth of 1 cm would be clinically

relevant and should thus be detectable. The results show that the SDD in part exceeded this a priori criterion for all measurement sites and all measuring tapes. Table 16 shows the percentage of differences between observers and between measuring days exceeding 1 cm, 1.5 cm and 2 cm, respectively, for all measurement sites and tape measures.
Nicholas et al. (1976) stated that a change in circumference noted by different observers on two different days was significant if it exceeds 1.5 cm at mid-patella, 2.7 cm at 7 cm above, and 3.5 cm at 15 cm above the patella [70]. Further they reported that the change needed to exceed only 1.0, 2.0, and 2.7 cm, respectively, to be significant if a single observer performed both measurements [70]. Thus, the a priori criterion might have been set at an unrealistic low level.

	Tape	D >1.0 cm [%]			D >1.5 cm [%]			D >2.0 cm [%]		
		PP	MP	DP	PP	MP	DP	PP	MP	DP
Inter-observer O1-O2	GI	7.4	4.4	1.5	0.0	0.0	0.0	0.0	0.0	0.0
	GII	19.7	5.3	2.6	6.6	1.3	0.0	3.9	0.0	0.0
	S	13.2	1.3	1.3	1.3	0.0	0.0	0.0	0.0	0.0
	W	0.0	1.3	1.3	0.0	0.0	0.0	0.0	0.0	0.0
Intra-observer t1-t2	GI	10.7	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	GII	19.4	4.2	2.8	4.2	0.0	0.0	0.0	0.0	0.0
	S	6.9	5.6	2.8	0.0	0.0	0.0	0.0	0.0	0.0
	W	4.2	5.6	4.2	2.8	0.0	1.4	0.0	0.0	0.0

Table 16: Girth - Inter-observer reproducibility for observers O1 and O2 (n=40 legs)
D >1.0 cm/>1.5 cm/>2.0 cm, difference between measurements exceeding 1.0 cm /1.5 cm/2.0 cm, difference between measurements exceeding;

4.1.1.2 Location of measurement

In this study, the different measurement sites influenced reproducibility. The lowest level of agreement was found for the measurement site at 7 cm proximal of mid-patella (maximum SDD 2.1 cm). The measurement sites at mid-patella and 7 cm distal of mid-patella showed higher agreement both within and between observers with maximum SDD values of 1.1 cm and 1.3 cm, respectively. This was confirmed by greater ICCs for both inter-tester and intra-tester variation at the sites mid-patella and 7 cm distal of mid-patella compared to the site at 7 cm proximal of mid-patella.
These results are in agreement with the findings of Nicholas et al. (1976) who performed measurements of the circumference of the knee with an ordinary tape measure at mid-

patella, 7 cm above the superior border of the patella, and 15 cm above the superior border of the patella. They reported that intra-observer and inter-observer variations were smallest at the mid-patella and increased as the location of measurement moved from the mid-patella to 7 and 15 cm above the superior border of the patella. They stated that this observation was probably due to the cylindrical shape of the thigh, since the circumference of a cone varies significantly as one moves along the long axis [70]. Kirwan et al. (1979) and Soderberg et al. (1996) noted variations by site of measurement. Kirwan et al. (1979) made girth measurements with "*tape measures available in hospital*" at the mid-patella level and at 1 cm and 15 cm above the palpated upper boarder of the patella [68]. They stated that the circumference 1 cm above the patella could be measured "most precisely" of the three circumferences tested. In the study of Soderberg et al. (1996), circumferential measurements of the uninvolved and involved legs in patients recovering from anterior cruciate ligament reconstructive surgery were taken at 15 cm inferior to the joint line, 5 cm, 10 cm, and 15 cm superior to the joint line, and at mid-thigh with a specially designed device. They reported that correlation coefficients (ICC) were lowest for the joint line [72]. Inter-observer reproducibility was higher than intra-observer reproducibility for the measurement site at 7 cm distal of mid-patella and for the Gulick I and Waegener tape measures. This is in contradiction to previous reports on circumferential measurements of the knee joint region, which observed higher intra-observer reliability and/or agreement [4,39,66,68,70,72,84].

4.1.1.3 Tape measures

Comparing the circumferences measured with the different tape measures, the measured girth values differed significantly and were smaller for the Gulick I and Waegener tape measures than for the Gulick II and standard tape measure (Table 17). This can be explained by the amount of tension applied to the tape and thereby on the soft tissue. When handled correctly, a six ounce tension was applied with the Gulick I tape measure, which was higher than the Gulick II tape measure tension (four ounces). The tension applied with the Waegener tape measure was unknown, but due to the results, one might speculate that the tension value was between four and six ounces. Observers were instructed not to apply tension on the standard tape measure, and results show that they followed these instructions on the whole. The limit of reproducibility was lowest for the Gulick II tape measure and highest for the Waegener tape measure, although the differences between the Gulick I, standard and Waegener tape measures were rather small. This indicates that reproducibility of circumferential measurements obtained with expensive devices, i.e.

Gulick I and Gulick II tape measures, were not higher than those obtained with inexpensive devices, i.e. Waegener and standard tape measures.

Tape	Mean ± SD	t1: O1-O2		t2: O1-O2		O1: t1-t2		O2: t1-t2	
		SDD	ICC	SDD	ICC	SDD	ICC	SDD	ICC
GI	36.3 ± 3.1	0.5-1.2	0.97-0.98	0.4-1.3	0.97-0.99	0.9-1.3	0.97-0.98	0.9-1.3	0.97-0.98
GII	37.6 ± 3.4	0.6-2.1	0.93-0.98	0.5-1.8	0.94-0.99	1.0-1.7	0.95-0.98	1.0-1.4	0.96-0.97
S	37.5 ± 3.3	0.7-1.7	0.95-0.98	0.7-1.2	0.98-0.99	1.0-1.2	0.97-0.98	0.9-1.1	0.98
W	36.7 ± 2.9	0.6-0.9	0.98	0.6-1.0	0.98-0.99	1.0-1.1	0.97-0.98	0.9-1.2	0.97-0.98

Table 17: Comparison of measuring tapes
t1: O1-O2, inter-observer reproducibility on first measuring day; t2: O1-O2, inter-observer reproducibility on second measuring day; O1: t1-t2, intra-observer reproducibility of observer O1; O2: t1-t2, intra-observer reproducibility of observer O2

Studies investigating the reproducibility of different tape measures are hard to find in literature. Harrelson et al. (1998) performed lower extremity circumference measurements in twenty-one subjects at three measurement sites with a standard tape measure and a Lufkin tape measure with a Gulick spring-loaded handle attached, comparable to the Gulick I tape measure used in this study [67]. They reported high reliability but a significant difference between the two tape measures and stated that the Lufkin tape measure with the Gulick handle attachment was associated with less measurement error and was therefore preferable to a standard tape measure. Geil (2005) investigated the accuracy and reliability of a standard tape measure, a spring tape measure and a circumferential tape measure, comparable to the Gulick I and Waegener tape measures, respectively, used in the present study on three foam positive models of transtibial amputee residual limbs and concluded that the type of tape measure used did not affect the results [94]. However, the measurements were performed on foam models with fixed marking of the measurement sites, limiting the comparability to other studies.
Baker et al. (2010) evaluated four different devices for obtaining circumferential measurements at four locations on the canine hindlimb and forelimb (the Gulick II tape measure, a retractable tape, an ergonomic tape measure and a circumference measuring tape comparable to the Waegener tape measure used in the present study). They stated that devices were equally precise for repeated measurements although the absolute measurement varied by device. The measurements with the ergonomic and circumference measuring tapes were significantly larger than with the retractable and Gulick II tape measures [92].

4.1.1.4 Influence of experience and profession of observers on reproducibility

In order to assess a possible influence of tester experience and profession on girth measurements, five observers with different degree of experience in girth measurements covering a broad spectrum of medical and non-medical professions were selected (Table 3). The observers O4 and O5 had no medical background and had never before assessed lower leg circumference with any tape measure used in this study. Observer O1 was a student of both medicine and electrical engineering, while observer O2 was a medical resident and observer O3 an orthopaedic surgeon. Further, the observers O1, O2, and O3 had experience in the usage of one of the used tape measures (Waegener tape measure) in girth measurements. There were differences in the degree of experience between the observers O1, O2, and O3. While the observers O1 and O2 had used the Waegener tape measure over a couple of months prior to the study, observer O3 had only occasionally performed circumferential measurements with this tape measure. None of the observers had clinical experience with the other measuring tapes (Gulick I, Gulick II plus, and standard tape measures).

A comparison of the observers across measurement sites and tape measures showed that the observers O1 and O2 had the highest intra-observer agreement (SDD range, 0.7 to 1.5 cm, and 0.5 to 1.4 cm, respectively) (Table 18). The observers O3 and O5 showed lowest agreement (SDD range, 1.1 to 3.3 cm, and 0.6 to 2.5 cm, respectively), while the observer O4 ranked in between (SDD range, 0.9 to1.8 cm). This was also reflected by the range of ICC values, showing higher reliability for the observers O1, O2 and O4 (range, 0.97 to 0.99, 0.97 to 1.00 and 0.97 to 0.99), than for the observers O3 and O5 (range, 0.88 to 0.99 and 0.92 to 1.00, respectively).

Considering the different measuring tapes, observer O1 had very similar SDDs for the Gulick I, standard and Waegener tape measures (range, 0.7 to 1.0 cm, 0.9 to 1.0 cm, and 0.9 cm, respectively), which was also reflected by the corresponding ICC values (range, 0.98 to 1.00 and 0.99), but lower SDD and ICC values for the Gulick II plus tape measure (SDD range, 0.8 to 1.5 cm and ICC range, 0.97 to 0.99). Observer O2 showed higher reproducibility for the Waegener and standard tape measures (SDD range, 0.7 to 1.1 cm and 0.5 to 1.3 cm, and ICC range, 0.98 to 0.99 and 0.98 to 1.00, respectively), than for the Gulick II and Gulick I tape measures (SDD range, 0.7 to 1.4 cm and 0.8 to 1.4, and ICC range, 0.97 to 0.99 and 0.98 to 0.99). For observer O3, agreement and reliability were highest for the Gulick I tape measure (SDD range, 1.1 to 1.2 cm and ICC 0.98) and lowest for the Gulick II plus tape measure (SDD range, 1.3 to 3.3 cm and ICC range, 0.88 to

0.99). For observers O4 and O5, reproducibility was lowest for the Gulick II plus tape (SDD range, 1.1 to 1.8 cm and 1.2 to 2.5 cm, and ICC range, 0.97 to 0.99 and 0.92 to 0.98, respectively). While observer O4 reached highest reproducibility with the Waegener tape measure (SDD range, 0.9 to 1.1 cm and ICC range, 0.98 to 0.99), observer O5 performed similar with the Gulick I, standard tape and Waegener measures (SDD range, 1.1 to 1.9 cm, 0.6 to 2.0 cm and 0.7 to 2.0 cm, and ICC range, 0.95 to 0.98, 0.95 to 1.00, and 0.95 to 0.99, respectively).

Site	Tape	O1		O2		O3		O4		O5	
		SDD	ICC	SDD	ICC	SDD	ICC	SDD	ICC	SDD	ICC
PP	GI	1.0	0.99	1.4	0.98	1.2	0.98	1.5	0.98	1.9	0.96
	GII	1.5	0.97	1.4	0.97	3.3	0.88	1.8	0.97	2.5	0.92
	S	1.0	0.99	1.3	0.98	1.8	0.96	1.6	0.98	1.6	0.97
	W	0.9	0.99	1.0	0.99	1.7	0.96	1.1	0.99	2.0	0.95
MP	GI	0.7	0.99	0.8	0.99	1.1	0.98	1.1	0.98	1.1	0.98
	GII	0.8	0.99	0.7	0.99	1.6	0.96	1.2	0.98	1.2	0.98
	S	0.9	0.99	0.6	0.99	1.3	0.97	1.3	0.98	0.6	1.00
	W	0.9	0.99	0.7	0.99	1.5	0.96	1.0	0.98	0.7	0.99
DP	GI	1.0	0.98	0.8	0.99	1.2	0.98	0.9	0.99	1.7	0.95
	GII	1.1	0.98	0.8	0.98	1.3	0.99	1.1	0.98	2.0	0.94
	S	0.9	0.99	0.5	1.00	1.5	0.97	1.1	0.98	2.0	0.95
	W	0.9	0.99	1.1	0.98	1.1	0.98	0.9	0.99	1.8	0.95

Table 18: Girth - Intra-observer reproducibility for the observers O1-O5 (n= 10 legs)

A comparison of the percentage of intra-observer differences exceeding 1.0 cm, 1.5 cm, and 2.0 cm shows that the difference between measuring days measured by the observers O1 and O2 never exceeded 1.5 cm, in contrary to the other observers (Table 19). For observer O1, the difference between measuring days exceeded 1 cm in only two cases (at 7 cm proximal of mid-patella measured with the Gulick II plus tape measure, and at mid-patella, measured with the standard tape measure).

In summary, the results of the girth measurements by the observers O1, O2, O3, O4 and O5 showed remarkable variations. As was to be expected, the most experienced observers O1 and O2 performed best. However, observer O3 with occasional experience in girth measurements showed lower performance than the inexperienced observer O4. One could speculate that reproducibility can only be increased if the observer regularly performs girth measurements. Another explanation for these findings may be that it is not experience, but talent, which has a greater influence on repeatability in girth measurements.

Site	Tape	D >1.0 cm [%]					D >1.5 cm [%]					D >2.0 cm [%]				
		O1	O2	O3	O4	O5	O1	O2	O3	O4	O5	O1	O2	O3	O4	O5
PP	GI	0	20	0	10	30	0	0	0	0	10	0	0	0	0	0
	GII	10	20	50	10	40	0	0	20	10	30	0	0	20	10	0
	S	0	0	30	20	10	0	0	0	0	10	0	0	0	0	0
	W	0	0	20	10	30	0	0	0	0	10	0	0	0	0	0
MP	GI	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0
	GII	0	0	20	10	10	0	0	0	10	0	0	0	0	0	0
	S	10	0	10	10	0	0	0	0	0	0	0	0	0	0	0
	W	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
DP	GI	0	0	10	0	20	0	0	0	0	10	0	0	0	0	0
	GII	0	0	0	0	20	0	0	0	0	10	0	0	0	0	0
	S	0	0	10	10	30	0	0	0	0	10	0	0	0	0	0
	W	0	10	0	0	20	0	0	0	0	10	0	0	0	0	0

Table 19: Percentage of differences exceeding 1 cm, 1.5 cm, and 2 cm (O1-O5)

Harrelson et al. (1998) stated that reliability of lower extremity circumference measurements was not influenced by tester experience [67]. In their study, an athletic trainer and a graduate assistant athletic trainer measured lower extremity girth at three sites in twenty-one subjects using a standard flexible tape measure and a Lufkin tape measure with a Gulick spring-loaded handle attachment, comparable with the Gulick I tape measure used in this study. However, no information could be found in their publication considering the experience of the testers in girth measurements. Further, Jakobsen et al. (2010) reported that tester experience appeared not to influence the degree of reliability [4]. In their study, an experienced physiotherapist with 10 years of experience in orthopaedic physiotherapy and an inexperienced physiotherapy student performed circumference measurements in nineteen outpatients having received a TKA, using a non-elastic tape measure at 1 cm proximal of the base of the patella.

In general, the previously reported studies considering tester experience compared only one experienced observer with one inexperienced observer. In order to make a statement about the influence of tester experience on reproducibility of a measuring device, measurements taken by a larger number of experienced and inexperienced testers should be compared.

The corresponding Bland and Altman plots for the measurement sites 7 cm proximal of mid-patella, mid-patella, and 7 cm distal of mid-patella are to be found in Figure 45, Figure 46, Figure 47, Figure 48 and Figure 49, respectively, in the appendix. Further, there are detailed tables of the results of the inter-observer and intra-observer analysis for the observers O1-O5, including mean difference with confidence intervals, standard deviation of the mean difference, limits of agreement with confidence intervals, and the ICCs with confidence intervals in the appendix (A.2.1.1 and A.2.1.2, respectively).

4.1.2 Sources of measurement error

In general, measurement error can originate from the measurement device itself, i.e., tape measure used, the user, and the subject [80]. In particular, sources of measurement error of girth measurements in this study include the determination of the measurement sites, the alignment of the measuring tape perpendicular to the leg axis, the use of the tape measures as recommended in the user´s manual, the applied tension on the tape, and reading errors. Further, biological changes might have an influence on the subject´s lower extremity circumference. Apart from that, a learning effect might be present which may cause higher variation mimicking less reproducibility. Additionally, the examined lower extremities of the subjects were considered independent entities because it was assumed that there was no influence of leg side on the assessment of reproducibility. Finally, the sample size was limited.

4.1.2.1 Determination of the measurement sites

The measurement sites for this study were defined at 7 cm proximal of mid-patella, at mid-patella, and at 7 cm distal of mid-patella because of two reasons. First, it was thought that these measurement sites are appropriate to describe circumferential changes due to swelling after TKA. Second, it was assumed that the upper and lower boarder of the patella were simply identifiable bony landmarks.

To determine the measuring positions, the first step was to identify the upper and lower patella borders in order to determine the mid-patella. During measurements, it turned out that this was a difficult task in some subjects due to a varying proportion of surrounding soft tissue in the knee region.

If an observer failed to identify the patella borders correctly, the marks for the different measurement sites subsequently differed from other observers and maybe even from one

measuring point of time to the other of the same observer. As a consequence, the girth was measured at a different site, which, in turn, led to different the measured circumferences, increasing the inter-observer and intra-observer variation. To investigate this possible source of error, the observers were asked to document the measured length of patella, which was used to identify the mid-patella site by division by 2. Table 20 shows the mean longitudinal length of the left and right patella on the first and second measuring day and the differences in measured patella length between measuring days. The patella length ranged from 4.0 to 7.0 cm. In literature, the mean longitudinal length ± SD of the patella has been reported to range from 40 ± 2.6 mm for females to 45.6 ± 3.0 for males in MRI measurements [95] and 3.70 ± 0.29 cm for females to 4.12 ± 0.29 cm for males in skeletons [96]. The values in this study are higher (females 5.1 cm, range 4.0 to 6.1 cm and males 5.6 cm, range 4.6 to 7.0 cm), because patella length was estimated by palpation on the body surface and thus the soft tissue inevitably enlarged the measured length.
The difference in patella length between measuring days ranged from 0.0 to 1.5 cm, which was up to 37.5% for a patella of 4 cm in length. Interestingly the inter-observer variation was smallest for the observers O3 and O5, who had shown the least reliability and agreement in girth measurements. This may indicate that other sources of variation make a greater contribution to the total measurement error.

Observer	Mean length t1 (cm)		Mean length t2 (cm)		Mean difference t1-t2 (cm)	
	Left patella	Right patella	Left patella	Right patella	Left patella	Right patella
O1	5.6 (5.1-6.4)	5.3 (4.0-6.0)	5.4 (4.6-6.5)	5.4 (4.3-6.5)	0.4 (0.0-0.4)	0.6 (0.0-1.5)
O2	5.2 (4.6-6.2)	5.1 (4.5-6.0)	5.2 4.0-6.5)	5.2 (4.0-6.6)	0.7 (0.2-1.3)	0.7 (0.2-1.5)
O3	5.5 (4.8-6.1)	5.5 (4.8-6.2)	5.4 (5.0-6.0)	5.4 (5.0-6.0)	0.2 (0.0-0.5)	0.2 (0.0-0.5)
O4	5.5 (5.0-6.2)	5.6 (5.0-7.0)	5.8 (4.6-6.5)	5.8 (4.5-7.0)	0.6 (0.0-1.0)	0.7 (0.0-1.5)
O5	5.3 (4.8-6.0)	5.3 (4.7-6.0)	5.2 (4.9-6.1)	5.2 (4.7-6.2)	0.3 (0.0-1.0)	0.4 (0.0-1.0)

Table 20: Mean difference in measured patella length for the observers O1-O5
With range of length in semicolons; t1, first measuring day; t2, second measuring day

4.1.2.2 Alignment and use of the tape measures

Prior to the first measurement day, the use of the tape measures as recommended in the user´s manual was shown to the observers. In case of the standard tape measure, the observers were instructed to avoid applying tension to the tape.

Apart from that observers were instructed to measure circumferentially, not elliptically. However, there was no training of the observers prior to the measurements. One could speculate that reproducibility was higher if the observers had pre-study training in the correct use of the tape measures.

The lower reproducibility of girth measurements at the measurement site 7 cm proximal of mid-patella might be explained by the cylindrical shape of the thigh and less firm soft tissue in this region, in combination with elliptically misalignment of the tape measure.

4.1.2.3 Biological changes

Lower extremity circumference may change from day to day, and during the day, due to time of day, outside temperature, and physical load between measurements. The measurements were performed at different daytimes and outdoor temperatures.

Further, there was a gap of five days between the measuring days. Although the subjects were asked to avoid excessive physical loading prior to and between the measurement days, it cannot be ruled out that there was a change in lower extremity girth due to biological changes.

Te Slaa at al. reported that reliability of repeated tape measurements decreased when the time interval between measurements increased and explained this drop in reliability by biological fluctuations [39].

4.1.2.4 Reading errors

Reasons for reading the wrong numbers from the tape measure might be an ambiguously defined zero line, an inadequate font size of the scaling, or a disadvantageous scaling. Of the measurement tapes used in this study, only the Gulick tape measure had a clear zero line. In case of the Gulick II plus and standard tape measure, the end of the tape was the zero line. The zero line of the Waegener tape measure was determined by the entry opening of the tape into the housing. Depending on the viewing angle, the zero line thus changed because it was not part of the tape, but of the housing.

The orientation of the numbers printed on the scaling in direction of the tape axis on the Gulick I and Gulick II plus tape measures, and perpendicular to the axis of the tape on the Waegener and standard tape measures (Figure 33, from top in this order).

The scale of the Gulick I tape measure turned out to be disadvantageous. Only the tens on the tape consisted of first and second digit, while the numbers in between consisted of one digit instead of two, e.g., 10-1-2-3-4-5-6-7-8-9-20 (Figure 33, tape on the top). This made the reading more difficult and prone to error. Further, the observer had to ensure that the side with the metric scale faced upwards to report measurements in cm. In case of the Gulick II plus tape measure, both a metric and an inch scale were found on one side of the tape, and only an inch scale on the other tape side. Thus the observer had to ensure to read from the right scale (cm).

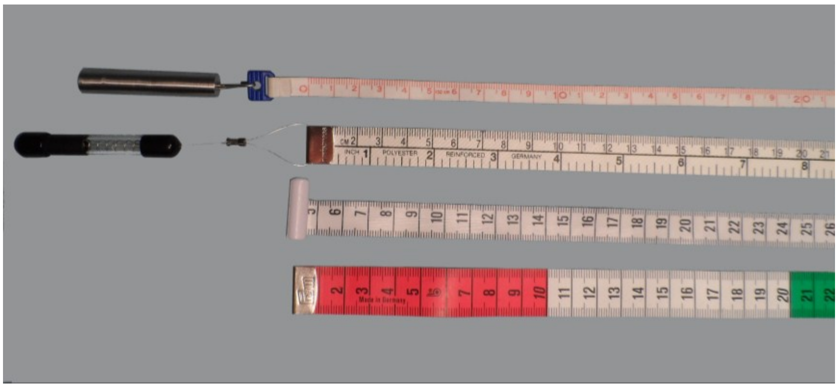


Figure 33: Scales of the used tape measure

4.1.2.5 Learning effect

Eighty percent of the observers stated in the questionnaire that their measurement reliability and agreement would improve with the duration of the measurement procedure because they got used to handling the measuring tapes. Thus, to consider a possible learning effect, the results of inter-observer reproducibility of the first and second measuring day are compared in Table 21. These day to day comparisons were made for the observers O1 and O2 only, who measured all subjects, because a learning effect was considered unlikely for those observers who measured only five subjects.

Inter-observer reproducibility was substantially higher for the second measuring day than for the first (SDD range, 0.7 to 1.8, and 0.7 to 2.1 cm, respectively, and ICC range, 0.94 to 0.99 and 0.93 to 0.98, respectively). In particular, agreement improved for all measurement sites, with 7 cm proximal of mid-patella and at mid-patella showing higher differences in maximum SDD (-0.3 cm) than 7 cm distal of mid-patella (-0.1 cm). Considering the different tape measures, agreement improved for the Gulick II and standard tape measures (-0.3 cm and -0.5 cm, respectively).

Summarising these results, improvement in reproducibility between first and second measuring day was observed for all measurement sites as well as the Gulick II and the standard tape measures. It is likely that the observers improved in tape positioning and usage of the tape measures.

Site	Tape	t1: O1-O2				t2: O1-O2			
		mD ± SD_{diff}	LOA	SDD	ICC	mD ± SD_{diff}	LOA	SDD	ICC
PP	GI	0.2 ± 0.5	-0.9 to 1.2	1.2	0.97	0.2 ± 0.6	-1.0 to 1.3	1.3	0.97
	GII	0.4 ± 0.8	-1.2 to 2.1	2.1	0.93	0.2 ± 0.8	-1.4 to 1.8	1.8	0.94
	S	0.3 ± 0.7	-1.0 to 1.7	1.7	0.95	0.1 ± 0.6	-0.9 to 1.2	1.2	0.98
	W	0.0 ± 0.4	-0.8 to 0.9	0.9	0.98	0.1 ± 0.5	-0.8 to 1.0	1.0	0.98
MP	GI	0.0 ± 0.5	-0.9 to 1.0	1.0	0.97	0.0 ± 0.4	-0.8 to 0.8	0.8	0.98
	GII	-0.1 ± 0.6	-1.2 to 1.0	1.2	0.97	-0.1 ± 0.3	-0.8 to 0.5	0.8	0.99
	S	0.2 ± 0.5	-0.7 to 1.1	1.1	0.97	0.2 ± 0.4	-0.5 to 0.9	0.9	0.98
	W	0.0 ± 0.4	-0.8 to 0.9	0.9	0.98	0.1 ± 0.3	-0.5 to 0.7	0.7	0.99
DP	GI	-0.2 ± 0.4	-0.9 to 0.5	0.9	0.98	-0.1 ± 0.3	-0.7 to 0.4	0.7	0.99
	GII	-0.2 ± 0.4	-1.0 to 0.6	1.0	0.98	-0.1 ± 0.4	-0.9 to 0.7	0.9	0.98
	S	0.0 ± 0.4	-0.7 to 0.7	0.7	0.98	0.1 ± 0.3	-0.5 to 0.7	0.7	0.99
	W	-0.2 ± 0.4	-1.0 to 0.6	1.0	0.98	-0.2 ± 0.4	-0.9 to 0.6	0.9	0.98

Table 21: Girth - Inter-observer reproducibility of first and second measuring day

4.1.2.6 Left vs. right leg

In clinical routine, a measuring tape has to show high reproducibility in assessing the circumference of the involved leg. Thus, differences between left and right leg were not considered clinically relevant and, therefore, both legs were treated as independent entities for statistical analysis. Further, it was hypothesised that differences in reproducibility of girth measurements between left and right leg do not exist. To test this assumption, the SDD and ICC values of the left and right lower extremity were compared. It was found that agreement (SDD) of left and right leg differed considerably, in inter-observer comparison for the first (SDD range, 0.7 to 1.4 cm and 0.9 to 2.5 cm, respectively) and second measuring day (SDD range, 0.6 to 1.4 and 0.6 to 2.2 cm, respectively), and in intra-observer comparison for observer O1 (SDD range, 0.8 to 1.4 cm and 1.1 to 2.0 cm, respectively). Observer O2 showed no side differences in SDD (range, 0.7 to 1.6 cm). Having a closer look at the results of observer O1 in Table 22, it shows that reproducibility was higher for the left leg (SDD range, 0.8 to 1.4 cm and ICC range, 0.96 to 0.99) than for the right leg (SDD range, 1.1 to 2.0 cm and ICC range, 0.94 to 0.98). Further, the measurement site at 7 cm proximal of mid-patella and the Gulick II plus and standard tape measures showed highest leg differences. Differences in SDD between left and right leg were negligible at the measurement site 7 cm distal of mid-patella and for the Waegener tape measure.

These results indicate that handedness of the observers may have an influence on reproducibility of circumferential leg measurements.

	O1				O2			
	Left leg		Right leg		Left leg		Right leg	
	SDD	ICC	SDD	ICC	SDD	ICC	SDD	ICC
total	0.8 to 1.4	0.96 to 0.99	1.1 to 2.0	0.94 to 0.98	0.7 to 1.6	0.95 to 0.99	0.7 to 1.6	0.95 to 0.99
PP	0.9 to 1.4	0.96 to 0.99	1.1 to 2.0	0.94 to 0.98	1.2 to 1.6	0.95 to 0.98	1.2 to 1.6	0.96 to 0.98
MP	0.8 to 1.0	0.97 to 0.99	1.1 to 1.3	0.96 to 0.98	1.0 to 1.1	0.97 to 0.98	0.7 to 1.2	0.97 to 0.99
DP	0.9 to 1.3	0.96 to 0.98	1.1 to 1.3	0.97 to 0.98	0.7 to 1.1	0.98 to 0.99	1.0 to 1.5	0.95 to 0.98
GI	0.8 to 1.2	0.98 to 0.99	1.1 to 1.3	0.96 to 0.98	0.9 to 1.5	0.97 to 0.99	0.7 to 1.4	0.97 to 0.99
GII	0.8 to 1.4	0.96 to 0.98	1.1 to 2.0	0.94 to 0.98	1.0 to 1.6	0.95 to 0.98	1.2 to 1.6	0.95 to 0.97
S	1.0 to 0.9	0.98 to 0.99	1.1 to 1.4	0.96 to 0.97	1.0 to 1.2	0.98 to 0.98	0.9 to 1.4	0.98 to 0.98
W	1.0 to 1.1	0.97 to 0.98	1.1 to 1.1	0.97 to 0.98	0.7 to 1.3	0.98 to 0.99	0.9 to 1.4	0.96 to 0.98

Table 22: Left and right leg intra-observer comparisons (SDD and ICC) (n=18)

4.1.2.7 Number of cases

The sample size for this study was limited. 20 subjects voluntarily participated in the study. On the first measuring day, all 40 legs of the 20 subjects were measured by observers O1 and O2. Two of the 20 subjects did not return for measurements on measuring day two (10%). Observers O3, O4, and O5 measured only 10 legs of five subjects (25%). The small sample size may reduce the statistical power of the results. It should be noted that all of our subjects were healthy individuals and, thus, the generalizability of our results to a patient population may be limited. This applies equally to the goniometric measurements.

4.1.3 Usability of the different tape measures

Although reliability and agreement are essential features, the usability of a measuring device used in the clinical setting plays a non-negligible role. To assess the user-friendliness of the four different measuring tapes used in this study, the observers were asked to fill in a questionnaire after the first measuring day (*Questionnaire on the usability of the measuring tapes*). This questionnaire contained questions on the pre-study clinical experience of the observers with the used tape measures, the advantages and disadvantages of each tape measure, the precision of each measuring tape and the required measurement time.

Three of the five observers (60% - O1, O2, and O3), had experience in handling one of the used tape measures (Waegener tape measure). While the observers O1 and O2 had used the Waegener tape measure over a couple of months prior to the study, observer O3 had only occasional experience. None of the observers had clinical experience with the other measuring tapes (Gulick I, Gulick II plus, and standard tape measures).

In the questionnaire, the observers were asked to rank each tape measure according to its level of usability. All five observers ranked the Waegener tape measure the most user friendly tape measure. Four of five observers (83%) ranked the Gulick II plus tape measure the least user friendly tape measure, while one observer selected the standard tape measure as the least easy-to-use measuring tape.

Another question addressed the subjective feeling of the observers concerning the precision of the tape measures. Four of five observers ranked the Waegener tape measure the most precise tape measure, while one observer selected the standard tape measure as the most precise measuring tape. However, the observers were less agreed on the least precise tape measure. Three observers selected the Gulick II plus tape measure, one observer the standard tape measure, and another observer the Gulick I tape measure to be the least precise one. When asked for the required measuring time, there was consensus that the Waegener tape measure enabled the quickest measurement. Four of five observers (80%) stated that the measurement took longest with the GII tape measure, while one observer selected the Gulick II plus tape measure to be the most time consuming.

Interestingly, none of the observers felt that the accuracy of measurements decreased with the duration of the measurement procedure due to increasing fatigue. On the other hand, four of five observers thought that their measuring accuracy increased with the duration of the measurement procedure because they got used to handling the measuring tapes.

When asked about the most important features of the Waegener tape measure which was considered the most user-friendly, all observers answered that easy handling and quick measuring due to the suspension mechanism were reasons to choose this tape measure. In this context, an easy to read scaling, a clearly visible zero line, an easy alignment of the tape, and a tape that doesn't slip were important properties (80%). Further, four of five observers stated that a precise measurement was a reason for choosing the Waegener tape measure.

Conversely, reasons for choosing the Gulick II plus tape measure as the least user-friendly were complicated handling (100%), long measuring time (100%), and difficult alignment due to tape thickness (80%). Four of five observers criticised that the tape easily slipped because it was too thick and therefore the contact surface between skin and tape too small, due to the conical leg shape.

Considering the tension control provided by the Gulick I, Gulick II plus, and Waegener tape measures, the tension applied by the Gulick I tape measure was considered too tight, while the tension of the Gulick II plus was considered too weak to tighten the tape. In general, the observers appreciated the tension control insofar as it increases the precision of the measurement. However, in case of the Gulick I and Gulick II plus tape measures, correct application of the tensioning device made them more difficult to handle. This was partly due to the fact that both hands were necessary for the measurement - one hand for the tensioning device and the other hand for stabilising the tape. If the observer had to reposition the tape, e.g. because it was not aligned perpendicular to the longitudinal axis of the leg, it was not possible to do the correction without releasing the tape and starting the measuring procedure from the beginning. As a consequence, the measurements with the Gulick I and Gulick II plus tape measures took longer compared with the standard and Waegener tape measures. Furthermore, the marking for the correctly applied tension was hard to identify in case of the Gulick I tape measure (Figure 11, left picture). Another reported disadvantage of the Gulick I and Gulick II plus tape measures was the scale on the tapes (see chapter 4.1.2.4).

While the numbers of the scaling on the Gulick I tape measure were of adequate type size, this did not apply for the Gulick II plus tape measure. One side of the tape provided a scale in inches, and the other side one scale in inches and one in cm. Thus, the font size of the metric scale was very small.

Although the Waegener tape measure showed the highest usability, one big disadvantage was the short life cycle of this product. The tape was torn at the pin after a few measurements. Further, the amount of tension applied to the tape was unknown and was felt to decrease with time.

In this study, reproducibility strongly depended on the tape measure used. The Waegener tape measure showed highest reliability for four of five observers. This tape measure was also selected the most user-friendly tape measure by all five observers. Further, four of five observers stated that they had the feeling to perform most precise with the Waegener tape measure, which was true for four of five observers. These results show that the most user-friendly tape measure also provided the most reproducible results. Thus it is recommended that the tape measure preferred by the clinician should be used for lower extremity girth measurements, if possible.

4.2 Discussion of the flexion reproducibility measurements

4.2.1 Discussion of the results

Based on the results of inter-and intra-observer agreement, the smallest detectable differences for goniometric measurements would lie between 5.9 and 9.0°, thus within the a priori criterion of 10°. Inter-observer reproducibility was slightly higher for the second measuring day. Further, observer O1 showed slightly higher agreement than observer O2 in intra-observer comparisons. Reliability was acceptable to high, with ICC values ranging from 0.85 to 0.99.

Intra-observer and inter-observer reproducibility were comparable. This is in contradiction to most previous reports on goniometric measurements of the knee joint reporting higher intra-tester reliability and/or agreement [4,45,48,74,75,77].

Measurements	SDD (°)	ICC
total	5.9 to 9.0	0.87 to 0.99
t1: O1-O2	5.9 to 8.2	0.88 to 0.99
t2: O1-O2	7.6 to 9.0	0.85 to 0.98
O1: t1-t2	7.1 to 7.7	0.90 to 0.98
O2: t1-t2	7.1 to 8.1	0.87 to 0.98

Table 23: Flexion: Summary of the results (observers O1 and O2)
t1: O1-O2, inter-observer reproducibility on first measuring day; t2: O1-O2, inter-observer reproducibility on second measuring day; O1: t1-t2, intra-observer reproducibility of observer O1; O2: t1-t2, intra-observer reproducibility of observer O2

Several published studies have addressed the reliability of goniometric measurements [47,48,74,75,78,81], but only a few reliability and agreement [4,5] (Table 24). The findings of our study agree with those of previous investigators who also demonstrated acceptable to high intra-tester reliability [4,48,74-77,79,82]. Compared with previous reports, the inter-tester reliability in this study was within the range of observed values or slightly higher.

Generally, higher reliability was reported for knee flexion than for extension. Rothstein et al (1983) found high intra-tester reliability for knee flexion and extension (ICC, 0.97-0.99).

Inter-tester reliability was high for knee flexion (0.89 to 0.92), but poor for knee extension (ICC 0.61 to 0.70). This is in agreement with Clapper (1988) et al., who observed high intra-observer reliability for knee flexion (ICC, 0.95) and acceptable intra-observer reliability for extension (ICC, 0.85). Watkins and colleague (1991) reported higher intra-tester and inter-tester reliability for knee flexion (ICC 0.99 and 0.90) than for knee extension (0.98 and 0.86) [48]. Käfer et al (2005) examined the reliability of visual estimation and goniometric measurements of knee range of motion. Intra-observer and inter-observer agreement was consistent regarding the goniometric assessment of flexion (r_s>0.6), whereas reliability was uniformly low for both measurements regarding the assessment of extension (r_s<0.6) [78].
Brosseau et al. (2001) reported very high intra-tester reliability in flexion (ICC, 0.997) and in extension (ICC, 0.972 to 0.985) for measurements with a universal goniometer. The inter-tester reliability was also high for flexion (ICC 0.977 to 0.982) and for extension (ICC 0.893 to 0.926) [75].
While most studies evaluated only reliability of goniometric measurements, only two addressed agreement of universal goniometers [4,5]. Lenssen et al. (2007) assessed inter-observer reproducibility of active and passive measurements of knee ROM using a long arm goniometer in TKA patients within the first four days after surgery. For passive flexion with the patient in supine position, they reported a mean difference of 1.4° with limits of agreement from 16.2° to 19.0° for the difference between the two observers. The corresponding ICC value was 0.88 [5]. In the study of Jakobsen et al. (2010), two physiotherapists with different clinical experience performed passive knee joint ROM measurements with a universal goniometer in patients having received a TKA [4]. They observed high ICC values for intra-observer and inter-observer comparisons ($ICC_{2,1}$ 0.97 to 0.98, smallest real difference SRD 5.1-6.2° and $ICC_{2,1}$ 0.96, SRD 6.4 to7.1°, respectively).
Joint mobility can be determined by visual estimation, universal goniometers, or measurement of joint angles after X-ray visualisation in maximum flexion or extension [45]. Opinions vary on the method that should be used to measure knee ROM. Watkins and colleagues examined intra-observer and inter-observer reliability of therapists who performed hand goniometry and visual estimation, and concluded that hand goniometry was superior to visual estimation for consistency of measurement [48]. This is in agreement with Lavernia et al. (2008) who stated that assessment of ROM through direct observation without a goniometer provides inaccurate findings [35]. Käfer et al (2005) reported comparable reliability for visual and goniometric assessment of knee range of

motion [78]. Conversely, Peters et al. found higher intra-observer and inter-observer reliability for visual estimation than for hand goniometry [79]. Watkins and colleagues (1991) examined the intra-observer and inter-observer reliability of hand goniometry and visual estimation of knee range of motion. They found interobserver reliability for hand goniometry to be 0.90 for flexion and 0.86 for extension, compared with 0.83 and 0.82 for flexion and extension, respectively, for visual estimation [48].

Study	Goniometer type	Intra-observer reliability/ agreement	Inter-observer reliability/ agreement
Rothstein et al. (1983)	Plastic goniometer, L=15.24 cm Plastic goniometer, L=25.4 cm Metal goniometer, L=30.5 cm	ICC: 0.99 ICC: 0.99 ICC: 0.97	ICC: 0.61-0.70 ICC: 0.61-0.63 ICC: 0.59-0.80
Gogia et al. (1987)	Standard plastic goniometer, L=30 cm		ICC: 0.99
Clapper et al. (1988)	Standard goniometer	ICC: 0.95	
Rheault et al. (1988)	Plastic universal goniometer, L=25.4 cm		Pearson´s r: 0.87
Watkins et al. (1991)	Plastic universal goniometer, L=12.7 cm	ICC: 0.99	ICC: 0.90
Brosseau et al. (1997)	Universal goniometer	ICC: 0.86-0.94[a] ICC: 0.95-0.97[b]	ICC: 0.62-0.70[a] ICC: 0.91-1.00[b]
Brosseau et al. (2001)	Universal plastic goniometer, L= 25 cm	ICC: 0.997	ICC: 0.977-0.982
Edwards et al. (2004)	Standard goniometer, L=30.5 cm	Correlation coefficient (no type specified): 0.92	Correlation coefficient (no type specified): 0.79
Käfer et al. (2005)	Universal goniometer	Spearman r_S: 0.65-0.76	r_S: 0.62-0.69
Lenssen et al. (2007)	Long-arm goniometer		ICC: 0.88[c], SDD: 17.6° LOA: -16.2-19.0°
Jakobsen et al. (2010)	Plastic goniometer, L=30 cm	$ICC_{2,1}$: 0.97-0.98 SRD: 5.1-6.2°	$ICC_{2,1}$: 0.96 SRD: 6.4-7.1°
Peters et al. (2011)	Standard plastic goniometer, L=18 cm	ICC: 0.96-0.98[d]	ICC: 0.70-0.95[d]

Table 24: Knee flexion in literature
L, length of goniometer arm; [a] small angles, [b] large angles, [c] passive flexion supine, [d] mean of two observers; length of goniometer arms reported in inches were converted to cm;

However, one has to be careful in comparing the results of different studies because of different study designs, different subjects evaluated (healthy subjects vs. patients), and different methods for statistical analysis. For example, Brosseau et al. (2001) reported very high reliability for goniometric measurements [75]. This might be explained by their measurement procedure, in which one independent observer had placed markers over the bony landmarks and the same markers were used by all testers. This approach seems questionable because an important source of error - the identification of the bony landmark by the observers – was eliminated. Thus, the reported ICC values did not reflect total reliability of the method. Rothstein et al. (1983), Watkins (1991), and Jakobsen (2009) used goniometers with covered scales [4,48,82]. The observer aligned the arms of the blinded goniometer and gave the goniometer to a data recorder who uncovered the scale and recorded the measurement obtained. This procedure was justified by the authors to avoid influence of the observers by the results of the measurement. Nevertheless, the reported reliability did not account for reading errors by the observer.
Considering statistical analysis, most of the cited studies used intraclass correlation coefficient as a measure for reliability. In this study, the $ICC_{2,1}$ as described by Shrout and Fleiss was used to describe the degree of reliability of the measurements. This type of ICC was used by some of the cited studies [4,47,75,76]. However, the type of ICC was not or not sufficiently specified in other cases [5,79,82], or different to the $ICC_{2,1}$ [48,75]. This makes comparisons difficult since the resulting ICC value varies depending on which version of the ICC is used [56,57]. Further, Rankin and Stokes (1998) pointed out that "*regardless of which reliability tests are selected, it appears that comparison of reliability results between studies is not possible unless the size and attributes of the sample tested in each case are virtually identical*" [63].

4.2.1.1 A priori criterion

Prior to the study, the maximum clinically acceptable SDD was set at 10° for goniometric measurements, because it was felt that a change in flexion of 10° would be clinically relevant and should thus be detectable. This criterion was fulfilled in all flexion measurements (SDD range, 5.9 to 9.0°).

4.2.1.2 Test positions

Reliability of goniometric measurements was reported to depend on the knee angle. Brosseau et al (1997) have shown higher reliability measurements for larger but not smaller knee angles [74].Considering the results of this study, agreement for observers O1 and O2 measuring 20 subjects was higher for test position P1 (range, 5.9 to 8.1°) than for P2 (range, 7.1 to 9.0°). Reliability differed substantially between knee positions P1 and P2, with ICCs ranging from 0.98 to 0.99 for P1 and from 0.85 to 0.90 for P2. However, for observers O1-O5 measuring five subjects, intra-observer agreement was comparable between position P1 (range, 4.9 to 9.1°) and P2 (range, 5.2 to 9.2°). Intra-observer reliability for observers O1-O5 ranged from 0.40 to 0.91 for position P1 and 0.92 to 0.97 for position P2. These inconsistent results do not allow drawing any conclusions about reproducibility of test positions.

4.2.1.3 Influence of experience and profession of observers on reproducibility

In order to assess a possible influence of tester experience and profession on reproducibility of goniometric measurements, five observers with different degrees of experience in girth and ROM measurements covering a broad spectrum of medical and non-medical professions were selected (Table 3). The observers O4 and O5 had no medical background and had never before assessed knee flexion with a universal goniometer, while the observers O1 and O2 had practised goniometric measurements daily for half of a year prior to the study. Observer O3 was an orthopaedic surgeon with three years of experience. The results of the intra-observer reproducibility for the observers O1, O2, O3, O4 and O5 and five subjects are summarized in Table 25. For the interested reader, the corresponding Bland and Altman plots are to be found in Figure 55 in the appendix (A.2.2.2).
The SDD ranged from 4.9 to 9.2°, and the ICC ranged from 0.40 to 0.97. There was no significant difference in agreement between the two testing positions P1 and P2 (SDD range, 4.9 to 9.1 ° and 5.2 to 9.2 °, respectively). However, the corresponding ICC values were lower for position P1 than for position P2 (range, 0.40 to 0.91, and 0.92 to 0.97, respectively).
Considering differences between the different observers, observer O4 showed the highest agreement (SDD range, 6.7 to 7.8°), followed by the observers O1 and O2 (SDD range, 6.9 to 8.2° and 7.2 to 8.3°, respectively). The observers O3 and O5 had lowest SDD values (range, 4.9 to 9.2° and 5.2 to 9.1°). However, the corresponding ICC values paint a completely different picture: Observer O3 showed highest reliability (ICC range, 0.91 to 0.92), followed by the observers O1 and O4 (range 0.80 to 0.93 and 0.76 to 0.97) and observer O5 (range, 0.69 to 0.97). Observer O2 had surprisingly low ICC coefficients (range 0.40 to 0.92). A closer look at Table 25 reveals that the SDD and ICC values do not correspond with respect to the level of repeatability. Observer O3, for example, had low agreement for testing position P2 (SDD, 9.2°), but reliability was high (ICC, 0.94). The

results of observers O1 and O2 measuring all subjects (3.2.3) showed similar contradictory findings, which will be discussed in chapter 4.3.

In summary, difference in agreement between observers was comparable. This is in agreement with Käfer et al (2005) and Jakobsen et al. (2010), who reported that tester experience had no influence on reliability and/or agreement [4,78].

Observer	P1: t1-t2			P2: t1-t2		
	mD ± SD_{diff}	SDD (°)	ICC [95% CI]	mD ± SD_{diff}	SDD (°)	ICC [95% CI]
O1	-2.9 ± 2.1	6.9	0.80 [0.00 , 0.96]	-3.9 ± 2.2	8.2	0.93 [0.06 to 0.99]
O2	-3.8 ± 2.3	8.3	0.40 [-0.11 , 0.81]	-0.9 ± 3.2	7.2	0.92 [0.73 to 0.98]
O3	-0.7 ± 2.2	4.9	0.91 [0.70 , 0.98]	-0.9 ± 4.2	9.2	0.94 [0.79 to 0.98]
O4	-1.6 ± 3.2	7.8	0.76 [0.31 , 0.93]	-1.5 ± 2.7	6.7	0.97 [0.87 to 0.99]
O5	-3.2 ± 3.0	9.1	0.69 [-0.01 , 0.92]	-1.1 ± 2.1	5.2	0.97 [0.90 to 1.00]

Table 25: Flexion - Intra-observer repeatability (observers O1-O5)

4.2.2 Sources of measurement error

The level of reproducibility of ROM measurements may be influenced by many factors such as the instruments and procedures applied, the joint examined, or the type of movement tested [62]. Some investigators have suggested that small joint angles might be more difficult to measure than large ones [74]. Others have reported that the professional background had an influence on reproducibility [35]. Further, reproducibility for knee goniometry relies largely on consistency in the identification of the bony landmarks on the proximal femur (greater trochanter) and the distal tibia (lateral malleolus) and visualizing the sagittal axis of movement for the knee joint [44].

Limitations of this study include a possible learning effect, possible differences in reproducibility of goniometric measurements between left and right leg, and the use of two different positioning devices.

4.2.2.1 Learning effect

In order to assess whether there is learning effect in knee flexion measurements, the results of inter-observer reproducibility of the first and second measuring day are compared in Table 26. Again, only the results of the observers O1 and O2 were used, because a learning effect was considered unlikely for those observers who measured only five subjects.

Comparison of the first and second measuring day in Table 26 did not show any improvement in reproducibility. On the contrary, there was a slight decrease in agreement and reliability between the first and second measuring day, although the SDD did not exceed 9° for both measuring days.

Position	Reproducibility t1			Reproducibility t2		
	mD ± SD (°)	LOA (°)	ICC [95% CI]	mD ± SD (°)	LOA (°)	ICC [95% CI]
P1	1.0 ± 2.5	-4.0 to 5.9	0.99 [0.98 to 0.99]	1.0 ± 3.4	-5.6 to 7.6	0.98 [0.96 to 0.99]
P2	-1.1 ± 3.6	-8.2 to 6.0	0.88 [0.77 to 0.93]	1.1 ± 4.1	-6.9 to 9.0	0.85 [0.84 to 0.96]

Table 26: Flexion - Inter-observer reproducibility of first and second measuring day

4.2.2.2 Positioning devices

Two different positioning devices were used to fix the lower extremities of the subjects in a standardised knee position, which may contribute to measurement error. The Bland and Altman plots in Figure 34 show that the differences between observers for both positioning devices are strictly separated for test position P1, while they are distributed over the range of mean flexion for test position P2. Further, the differences between observers lie in a narrower range for positioning device PD1 than for PD2.

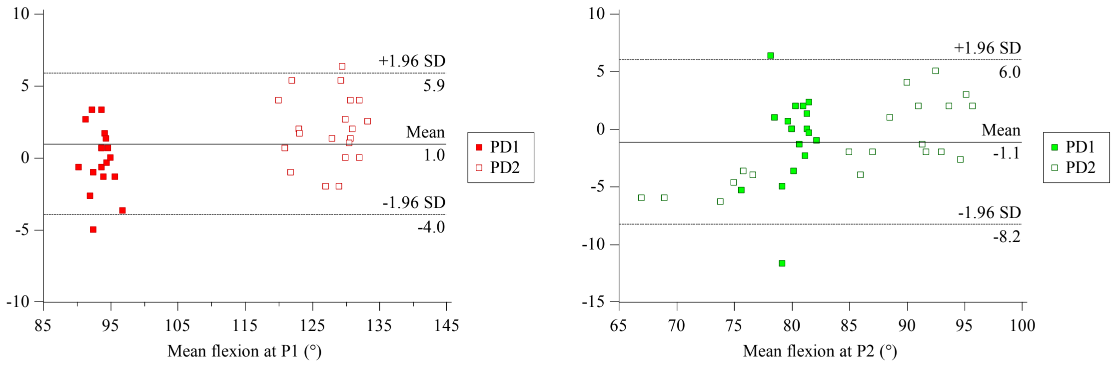


Figure 34: Influence of different positioning devices

This might be explained by the different construction of the positioning devices. In positioning device PD1, the leg was laid on two support plates positioned at a certain angle to each other, which defined the knee joint angle. These angles were approximately 70° for test position P1 and approximately 90° for test position P2. Conversely, the subjects were sitting in the positioning device PD2 and the distance between subject´s buttocks and foot plate defined the angle between thigh and limb. Thus, the knee joint angle far more depended on the height of the subject, resulting in a broader range of flexion angles than for positioning device PD1.

Considering differences in agreement between the different positioning devices, Table 27 shows there was hardly any difference in intra-observer agreement between positioning devices. Intra-observer agreement (SDD) ranged from 7.5 to 8.5° for PD1 and 5.9 to 8.5° for PD2. However, inter-observer agreement was higher for positioning device PD1 than PD2 (SDD range, 4.5 to 8.4° and 6.6 to 11.5°, respectively).

Agreement		PD1				PD2			
		P1		P2		P1		P2	
		mD ± SD	SDD	mD ± SD	SDD	mD ± SD	SDD	mD ± SD	SDD
Inter-observer	t1	-0.1 ± 2.2	4.5	-0.8 ± 3.9	8.4	2.0 ± 2.4	6.6	-1.4 ± 3.5	8.2
	t2	0.4 ± 3.4	7.0	0.6 ± 2.5	5.6	1.5 ± 3.4	8.1	1.5 ± 5.1	11.5
Intra-observer	O1	-1.9 ± 2.9	7.5	-3.0 ± 2.4	7.7	-1.7 ± 2.7	7.0	-3.5 ± 2.2	7.8
	O2	-1.3 ± 3.2	7.5	-1.4 ± 3.6	8.5	-2.4 ± 3.1	8.5	-0.2 ± 2.9	5.9

Table 27: Agreement for the positioning devices PD1 and PD2

4.2.2.3 Left vs. right leg

For statistical analysis, both legs were treated as independent entities because it was hypothesised that there was no difference in reproducibility of the goniometric assessment of left and right knee flexion. To test this assumption, the SDD and ICC values of left and right lower extremity were compared. It was found that agreement (SDD) and reliability (ICC) of left and right leg measurements differed substantially for inter-observer and intra-observer comparisons, with the right leg showing higher agreement and reliability (Table 28). These results indicate that reproducibility of goniometric measurements using a short-arm goniometer may be influenced by the handedness of the observers.

	Left leg		Right leg	
	SDD range (°)	ICC range	SDD range (°)	ICC range
t1	5.1 to 9.7	0.84 to 0.99	6.7	0.92 to 0.99
t2	6.8 to 9.3	0.86 to 0.99	8.4 to 9.0	0.86 to 0.98
O1	6.5 to 8.8	0.89 to 0.99	6.5 to 7.9	0.92 to 0.98
O2	7.1 to 7.2	0.88 to 0.98	7.4 to 9.0	0.87 to 0.97

Table 28: Flexion – Left and right leg side differences

4.3 Discussion of the statistical methods

In intra-observer comparison of the goniometric measurements, the SDD (maximum LOA) and the ICC did not behave the way they were expected, i.e. the higher the SDD, the lower the ICC and vice versa (Table 12). Even though the SDD was 7.1° for the measurements of observer O1 at P1 and observer O2 at P2, the ICC values differed substantially (0.98 and 0.89, respectively) (Figure 35). How can these unexpected results be explained?

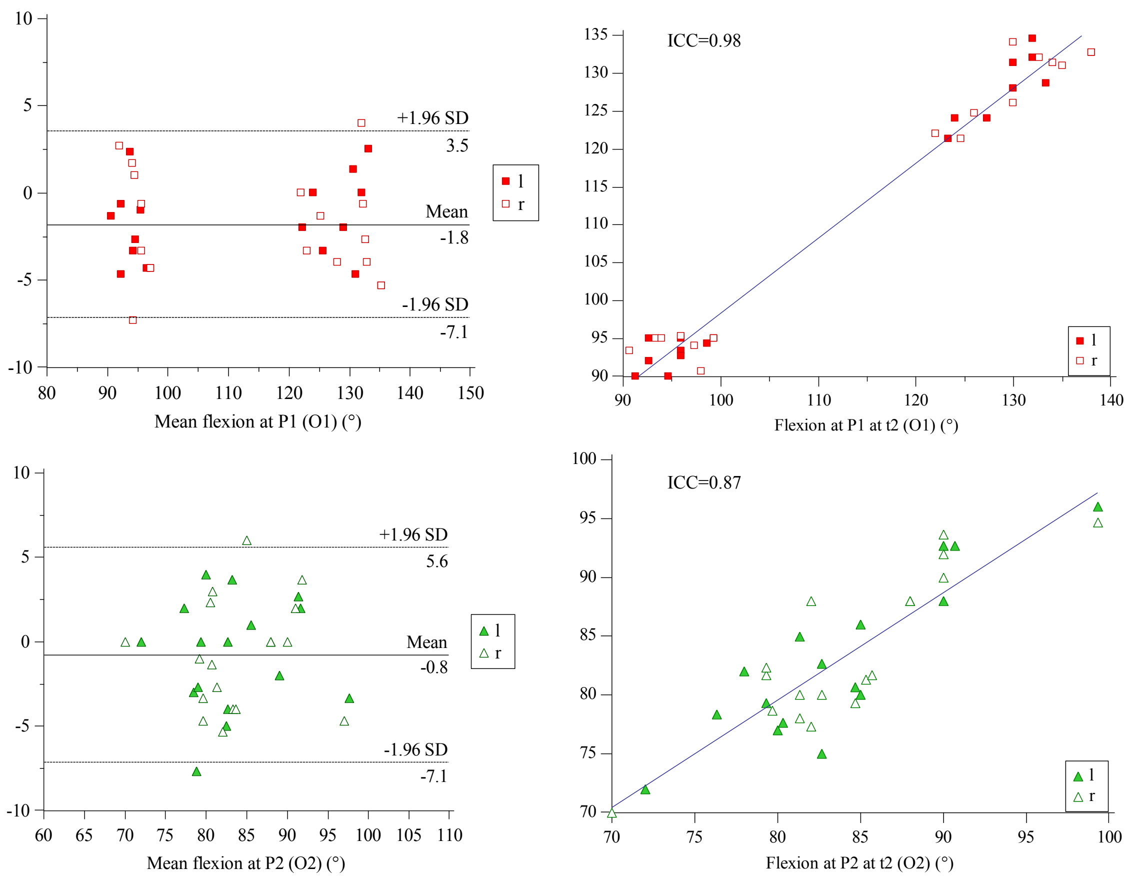


Figure 35: B-A plots vs. scatter plots

The ICC depends on the range of variables measured, which is not the case for the Bland and Altman LOA [58]. Therefore, for a group of subjects with a wide range of measurements, the ICC is likely to be greater than for a more homogenous sample group with similar flexion measurements. This is stated as a major criticism of the ICC on the one hand. On the other hand, it has been suggested that reliability should reflect true variability. In this context, a measurement error of 10° may or may not be important depending on the range of measured flexion values [63].

In case of the measurements obtained by observer O1 at test position P1, the measurement ranged from 90 to 138° (max-min=48°), which was a much wider range than in the case of observer O2 measuring test position P2 (68 to 99.3°, max-min=31.3°). Thus, an SDD of 7.1° seems to be more relevant clinically in the second case. This is reflected by the corresponding ICC values, being 0.87 for the second and 0.98 for the first case. Figure 35 compares Bland and Altman plots and scatter diagrams for both cases.

One of the advantages of the B-A limits of agreement, compared to the ICC, is that the amount of measurement error is quantified in units of the measurement scale, which simplifies the clinical interpretation of results. However, the Bland and Altman 95% limits of agreement indicate a range of error, but this must be interpreted with reference to the range of measurement values obtained. Thus, for a measuring device to be of clinical value, it has to show small limits of agreements and a high ICC. A statistical measure taking into account the advantages of ICC and the LOA would be appreciable.

4.4 Clinical course

Lower limb swelling occurred in all patients after TKA surgery. The circumference change was significantly higher above the knee (PP, mean 5.1 cm, range, 2.3 to 7.6 cm) than at mid-patella (MP, mean 3.8 cm, range, 1.9 to 9.8 cm) and below the knee (DP) (mean 2.8 cm, range, 1.7 to 7.2 cm) ($P=0.0002$ and $P<0.0001$, respetively). The mean maximum swelling was reached on the third or fourth postoperative day. The findings of this study are in agreement with the results of Gao et al. (2011), who retrospectively analysed the mean changes in limb circumferences of 286 patients who underwent primary unilateral total knee arthroplasty [2]. They reported that swelling was most pronounced from the third to the fifth post-operative day and usually occurred in both lower limbs. Further swelling was significantly more pronounced in the operated limb than in the non-operated limb. The swelling above the knee was also significantly greater than that below the knee.

Passive ROM increased continuously after TKA. On sixth postoperative day, the mean passive ROM was 79.0° (range, 55 to 100°). Swelling and passive ROM did not show any significant correlation ($P\geq0.1375$).

Postsurgical pain intensity reported by the patients was highest preoperatively (mean NRS_{max} 7.0, range 4 to 9), which might be explained by the patient-controlled analgesia (72 hours postoperatively). After surgery, the pain intensity decreased from the first postoperative day (mean NRS_{max} 6.6, range, 3 to 9) to the dismissal day (mean NRS_{max} 3.0, range 1 to 9). The minimum NRS curve showed a course similar to the maximum NRS curve, except for a very small increase at first postoperative day, indicating that the minimum pain increased for a short period after surgery.

5 Conclusion

In this study it was demonstrated that the reproducibility of lower extremity girth measurements depends on the sites where the measurements are taken, and the types of tape measures used. Based on the results of inter-and intra-observer agreement, the smallest detectable differences (SDD) would lie between 0.4 cm and 2.1 cm. This means that only changes in girth larger than these values can be detected beyond measurement error when one or different clinicians perform girth measurements in the knee region in a comparable environment. Considering the measurement sites, the inter-observer and intra-observer reproducibility increased from proximal to distal. Regarding the different tape measures, the Gulick II plus tape measure showed lowest reproducibility. Agreement was highest for the Waegener tape measure, even though differences to the Gulick I and standard tape measure were small. If subjects are assessed with the Gulick II plus tape measure at 7 cm proximal of mid-patella, differences in girth of less than 2.1 cm cannot be distinguished from measurement error. Conversely, a change in girth of more than 1.2 cm, 1.3 and 1.7 cm measured with the Waegener, Gulick I and standard tape measure, respectively, can be considered a real change beyond measurement error. However, these values clearly exceeded the a priori criterion of 1 cm and thus seem to be too large for the measurement of individual patients in clinical practice, or to assess intra-individual changes in girth over time.

The results of the knee flexion measurements with a short-arm universal goniometer showed that changes in flexion exceeding 9° can be considered a real change above measurement error.

Since the measurements were already standardized and the mean values of multiple measurements at each timepoint were used for analysis, the best way to reduce variation in measurements would seem to be training of the observers. In general, intra-observer reproducibility was reported higher than inter-observer reproducibility, so it is recommended that the same observer should be responsible for the measurement of treatment outcome for each patient.

To be useful for outcome assessment in clinical practice or research, an instrument should have high responsiveness, which is strongly determined by the level of agreement. The smallest detectable difference should be smaller than the minimal clinically important difference that one wants to detect. Given the large SDDs, the value of circumferential girth and goniometric measurements as an outcome measure can be questioned.

Clinical evaluation of in-patients undergoing TKA showed that postoperative swelling does not seem to influence passive ROM after surgery.

6 References

[1] Kurtz SM, Ong KL, Lau E, Widmer M, Maravic M, Gomez-Barrena E, et al. International survey of primary and revision total knee replacement. Int Orthop 2011 Dec;35(12):1783-1789.

[2] Gao FQ, Li ZJ, Zhang K, Huang D, Liu ZJ. Risk factors for lower limb swelling after primary total knee arthroplasty. Chin Med J (Engl) 2011 Dec;124(23):3896-3899.

[3] Munk S, Jensen NJ, Andersen I, Kehlet H, Hansen TB. Effect of compression therapy on knee swelling and pain after total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 2013 Feb;21(2):388-392.

[4] Jakobsen TL, Christensen M, Christensen SS, Olsen M, Bandholm T. Reliability of knee joint range of motion and circumference measurements after total knee arthroplasty: does tester experience matter? Physiother Res Int 2010 Sep;15(3):126-134.

[5] Lenssen AF, van Dam EM, Crijns YH, Verhey M, Geesink RJ, van den Brandt PA, et al. Reproducibility of goniometric measurement of the knee in the in-hospital phase following total knee arthroplasty. BMC Musculoskelet Disord 2007 Aug 17;8:83.

[6] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2011. Wien: Verlag Österreich; 2012.

[7] Van Manen MD, Nace J, Mont MA. Management of primary knee osteoarthritis and indications for total knee arthroplasty for general practitioners. J Am Osteopath Assoc 2012 Nov;112(11):709-715.

[8] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 1999. Wien: Verlag Österreich; 2001.

[9] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2000. Wien: Verlag Österreich; 2002.

[10] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2001. Wien: Verlag Österreich; 2003.

[11] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2003. Wien: Verlag Österreich; 2005.

[12] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2005. Wien: Verlag Österreich; 2006.

[13] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2004. Wien: Verlag Österreich; 2006.

[14] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2006. Wien: Verlag Österreich; 2007.

[15] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2007. Wien: Verlag Österreich; 2008.

[16] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2008. Wien: Verlag Österreich; 2009.

[17] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2009. Wien: Verlag Österreich; 2010.

[18] Statistik Austria editor. Jahrbuch der Gesundheitsstatistik 2010. Wien: Verlag Österreich; 2011.

[19] Thomsen MG, Husted H, Otte KS, Orsnes T, Troelsen A. Indications for knee arthroplasty have remained consistent over time. Dan Med J 2012 Aug;59(8):A4492.

[20] Guyton JL. Arthroplasty of ankle and knee. In: Calale STH, Campbell WCB, Crenshaw AHH, editors. Campbell's operative orthopaedics. 9. ed. ed. St. Louis, Mo. u.a.]: Mosby; 1998. p. 232-295.

[21] Leopold SS. Minimally invasive total knee arthroplasty for osteoarthritis. N Engl J Med 2009 Apr 23;360(17):1749-1758.

[22] Long W, Scuderi G. Total knee replacement. In: Bulstrode C, Wilson-MacDonald J, Eastwood D, MacMaster J, Fairbank J, Singh P, et al, editors. Oxford Textbook of trauma and orthopaedics. 2nd ed. New York: Oxford University Press; 2011. p. 660-664.

[23] Gopinath P, Arun B. Surgical exposures for primary total knee arthroplasty. Journal of Orthopaedics 1998;1(1):e6.

[24] Kohn D, Pohlemann T. Bikonyläre Prothese und Totalendoprothese. In: Kohn D, Pohlemann T, editors. Operationsatlas für die orthopädisch-unfallchirurgische Weiterbildung Berlin Heidelberg: Springer; 2010. p. 109-114.

[25] Jones B. Complications of total knee replacement. In: Bulstrode CJK, editor. Oxford textbook of trauma and orthopaedics. 2. ed. ed. Oxford: Oxford Univ. Press; 2011. p. 665-671.

[26] Leone JM, Hanssen AD. Management of Infection at the Site of a Total Knee Arthroplasty. The Journal of Bone & Joint Surgery 2005 October 1;87(10):2335-2348.

[27] Bauer TW, Parvizi J, Kobayashi N, Krebs V. Diagnosis of Periprosthetic Infection. The Journal of Bone & Joint Surgery 2006 April 1;88(4):869-882.

[28] Jamsen E, Huhtala H, Puolakka T, Moilanen T. Risk factors for infection after knee arthroplasty. A register-based analysis of 43,149 cases. J Bone Joint Surg Am 2009 Jan;91(1):38-47.

[29] Kim YH, Kim JS. Incidence and natural history of deep-vein thrombosis after total knee arthroplasty. A prospective, randomised study. J Bone Joint Surg Br 2002 May;84(4):566-570.

[30] Blanchard J, Meuwly JY, Leyvraz PF, Miron MJ, Bounameaux H, Hoffmeyer P, et al. Prevention of deep-vein thrombosis after total knee replacement. Randomised comparison between a low-molecular-weight heparin (nadroparin) and mechanical prophylaxis with a foot-pump system. J Bone Joint Surg Br 1999 Jul;81(4):654-659.

[31] Martin GM, Thornhill TS, Katz JN. Complications in total knee arthroplasty. In: Basow DS, editor. UpToDate Waltham, MA: UpToDate; 2013.

[32] Long WJ, Scuderi GR. Total knee replacement. In: Bulstrode CJK, editor. Oxford textbook of trauma and orthopaedics. 2. ed. ed. Oxford: Oxford Univ. Press; 2011. p. 660-664.

[33] Scott RD, 1943-. Totale Kniearthroplastik. 1. Aufl. ed. München ; Jena: Elsevier, Urban & Fischer; 2007.

[34] Dennis DA, Komistek RD, Stiehl JB, Walker SA, Dennis KN. Range of motion after total knee arthroplasty: the effect of implant design and weight-bearing conditions. J Arthroplasty 1998 Oct;13(7):748-752.

[35] Lavernia C, D'Apuzzo M, Rossi MD, Lee D. Accuracy of knee range of motion assessment after total knee arthroplasty. J Arthroplasty 2008 Sep;23(6 Suppl 1):85-91.

[36] Scott RD. Stiffness associated with total knee arthroplasty. Orthopedics 2009 Sep;32(9):10.3928/01477447-20090728-30.

[37] Sehat KR, Evans R, Newman JH. How much blood is really lost in total knee arthroplasty?. Correct blood loss management should take hidden loss into account. Knee 2000 Jul 1;7(3):151-155.

[38] Li B, Wen Y, Wu H, Qian Q, Lin X, Zhao H. The effect of tourniquet use on hidden blood loss in total knee arthroplasty. Int Orthop 2009 Oct;33(5):1263-1268.

[39] te Slaa A, Mulder P, Dolmans D, Castenmiller P, Ho G, van der Laan L. Reliability and reproducibility of a clinical application of a simple technique for repeated circumferential leg measurements. Phlebology 2011 Feb;26(1):14-19.

[40] Deltombe T, Jamart J, Recloux S, Legrand C, Vandenbroeck N, Theys S, et al. Reliability and limits of agreement of circumferential, water displacement, and optoelectronic volumetry in the measurement of upper limb lymphedema. Lymphology 2007 Mar;40(1):26-34.

[41] Rabe E, Stucker M, Ottillinger B. Water displacement leg volumetry in clinical studies-a discussion of error sources. BMC Med Res Methodol 2010 Jan 13;10:5-2288-10-5.

[42] "Archimedes' Principle." *Merriam-Webster.com*. Merriam-Webster, n.d. Web. Available at: <http://www.merriam-webster.com/dictionary/Archimedes' principle>. Accessed August, 22, 2013.

[43] Tewari N, Gill PG, Bochner MA, Kollias J. Comparison of volume displacement versus circumferential arm measurements for lymphoedema: implications for the SNAC trial. ANZ J Surg 2008 Oct;78(10):889-893.

[44] Naylor JM, Ko V, Adie S, Gaskin C, Walker R, Harris IA, et al. Validity and reliability of using photography for measuring knee range of motion: a methodological study. BMC Musculoskelet Disord 2011 Apr 18;12:77-2474-12-77.

[45] Cleffken B, van Breukelen G, Brink P, van Mameren H, Olde Damink S. Digital goniometric measurement of knee joint motion. Evaluation of usefulness for research settings and clinical practice. Knee 2007 Oct;14(5):385-389.

[46] Norkin CC, White DJ. Introduction to goniometry. Measurement of joint motion; a guide to goniometry. 3. ed. ed. Philadelphia: Davis; 2003. p. 1-53.

[47] Gogia PP, Braatz JH, Rose SJ, Norton BJ. Reliability and validity of goniometric measurements at the knee. Phys Ther 1987 Feb;67(2):192-195.

[48] Watkins MA, Riddle DL, Lamb RL, Personius WJ. Reliability of goniometric measurements and visual estimates of knee range of motion obtained in a clinical setting. Phys Ther 1991 Feb;71(2):90-6; discussion 96-7.

[49] Ferreira-Valente MA, Pais-Ribeiro JL, Jensen MP. Validity of four pain intensity rating scales. Pain 2011 Oct;152(10):2399-2404.

[50] Williamson A, Hoggart B. Pain: a review of three commonly used pain rating scales. J Clin Nurs 2005 Aug;14(7):798-804.

[51] Jensen MP, Chen C, Brugger AM. Postsurgical pain outcome assessment. Pain 2002 Sep;99(1-2):101-109.

[52] Bijur PE, Latimer CT, Gallagher EJ. Validation of a verbally administered numerical rating scale of acute pain for use in the emergency department. Acad Emerg Med 2003 Apr;10(4):390-392.

[53] Bruton A, Conway JH, Holgate ST. Reliability: What is it, and how is it measured? Physiotherapy 2000 02/01;86(2):94-99.

[54] Bravo G, Potvin L. Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. J Clin Epidemiol 1991;44(4-5):381-390.

[55] Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable? Injury 2011 Mar;42(3):236-240.

[56] Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998 Oct;26(4):217-238.

[57] Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005 Feb;19(1):231-240.

[58] De Vet H. Observer Reliability and Agreement. In: Armitage P, Colton T, editors. Encyclopedia of Biostatistics. 2nd ed.: John Wiley & Sons, Ltd; 2005. p. 3801-3805.

[59] de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006 Oct;59(10):1033-1039.

[60] Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003 Jul;22(1):85-93.

[61] de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001 Fall;17(4):479-487.

[62] Terwee CB, de Winter AF, Scholten RJ, Jans MP, Deville W, van Schaardenburg D, et al. Interobserver reproducibility of the visual estimation of range of motion of the shoulder. Arch Phys Med Rehabil 2005 Jul;86(7):1356-1361.

[63] Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998 Jun;12(3):187-199.

[64] Allison SC, Westphal KA, Finstuen K. Knee extension and flexion torque as a function of thigh asymmetry. J Orthop Sports Phys Ther 1993 Dec;18(6):661-666.

[65] Berard A, Zuccarelli F. Test-retest reliability study of a new improved Leg-O-meter, the Leg-O-meter II, in patients suffering from venous insufficiency of the lower limbs. Angiology 2000 Sep;51(9):711-717.

[66] Gross MT, McGrain P, Demilio N, Plyler L. Relationship between multiple predictor variables and normal knee torque production. Phys Ther 1989 Jan;69(1):54-62.

[67] Harrelson GL, Leaver-Dunn D, Fincher AL, Leeper JD. Inter- and intratester reliability of lower extremity circumference measurements. Journal of Sport Rehabilitation, J Sport Rehabil 1998;7(4):300-306.

[68] Kirwan JR, Byron MA, Winfield J, Altman DG, Gumpel JM. Circumferential measurements in the assessment of synovitis of the knee. Rheumatol Rehabil 1979 May;18(2):78-84.

[69] Labs KH, Tschoepl M, Gamba G, Aschwanden M, Jaeger KA. The reliability of leg circumference assessment: a comparison of spring tape measurements and optoelectronic volumetry. Vasc Med 2000;5(2):69-74.

[70] Nicholas JJ, Taylor FH, Buckingham RB, Ottonello D. Measurement of circumference of the knee with ordinary tape measure. Ann Rheum Dis 1976 Jun;35(3):282-284.

[71] Ross M, Worrell TW. Thigh and calf girth following knee injury and surgery. J Orthop Sports Phys Ther 1998 Jan;27(1):9-15.

[72] Soderberg GL, Ballantyne BT, Kestel LL. Reliability of lower extremity girth measurements after anterior cruciate ligament reconstruction. Physiother Res Int 1996;1(1):7-16.

[73] Whitney S, Mattlocks L, Irrgang J. Reliability of lower girth measurements and right- and left-side differences. Journal of Sport Rehabilitation, J Sport Rehabil 1995;4(2):108-115.

[74] Brosseau L, Tousignant M, Budd J, Chartier N, Duciaume L, Plamondon S, et al. Intratester and intertester reliability and criterion validity of the parallelogram and

universal goniometers for active knee flexion in healthy subjects. Physiother Res Int 1997;2(3):150-166.

[75] Brosseau L, Balmer S, Tousignant M, O'Sullivan JP, Goudreault C, Goudreault M, et al. Intra- and intertester reliability and criterion validity of the parallelogram and universal goniometers for measuring maximum active knee flexion and extension of patients with knee restrictions. Arch Phys Med Rehabil 2001 Mar;82(3):396-402.

[76] Clapper MP, Wolf SL. Comparison of the reliability of the Orthoranger and the standard goniometer for assessing active lower extremity range of motion. Phys Ther 1988 Feb;68(2):214-218.

[77] Edwards JZ, Greene KA, Davis RS, Kovacik MW, Noe DA, Askew MJ. Measuring flexion in knee arthroplasty patients. J Arthroplasty 2004 Apr;19(3):369-372.

[78] Kafer W, Fraitzl CR, Kinkel S, Clessienne CB, Puhl W, Kessler S. Outcome assessment in total knee arthroplasty: is the clinical measurement of range of motion a reliable measurable outcome variable? Z Orthop Ihre Grenzgeb 2005 Jan-Feb;143(1):25-29.

[79] Peters PG, Herbenick MA, Anloague PA, Markert RJ, Rubino LJ,3rd. Knee range of motion: reliability and agreement of 3 measurement methods. Am J Orthop (Belle Mead NJ) 2011 Dec;40(12):E249-52.

[80] Piriyaprasarth P, Morris ME, Winter A, Bialocerkowski AE. The reliability of knee joint position testing using electrogoniometry. BMC Musculoskelet Disord 2008 Jan 22;9:6-2474-9-6.

[81] Rheault W, Miller M, Nothnagel P, Straessle J, Urban D. Intertester reliability and concurrent validity of fluid-based and universal goniometers for active knee flexion. Phys Ther 1988 Nov;68(11):1676-1678.

[82] Rothstein JM, Miller PJ, Roettger RF. Goniometric reliability in a clinical setting. Elbow and knee measurements. Phys Ther 1983 Oct;63(10):1611-1615.

[83] Alonso A, Hekeik P, Adams R. Predicting a recovery time from the initial assessment of a quadriceps contusion injury. Aust J Physiother 2000;46(3):167-177.

[84] Marks JS, Palmer MK, Burke MJ, Smith P. Observer variation in examination of knee joints. Ann Rheum Dis 1978 Aug;37(4):376-377.

[85] Theiler R, Stucki G, Schutz R, Hofer H, Seifert B, Tyndall A, et al. Parametric and non-parametric measures in the assessment of knee and hip osteoarthritis: interobserver reliability and correlation with radiology. Osteoarthritis Cartilage 1996 Mar;4(1):35-42.

[86] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428.

[87] IBM Corp. IBM SPSS Statistics for Windows, Version 20.0. Released 2011. IBM Corp., Armonk, NY

[88] MedCalc Software. MedCalc for Windows version 12.7.0. MedCalc Software, Ostend, Belgium

[89] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 8;1(8476):307-310.

[90] Hanneman SK. Design, analysis, and interpretation of method-comparison studies. AACN Adv Crit Care 2008 Apr-Jun;19(2):223-234.

[91] Microsoft. Microsoft Office Excel. 2010. Microsoft, Redmond, USA

[92] Baker SG, Roush JK, Unis MD, Wodiske T. Comparison of four commercial devices to measure limb circumference in dogs. Vet Comp Orthop Traumatol 2010;23(6):406-410.

[93] Maylia M, Fairclough J, Nokes L, Jones M. Can thigh girth be measured accurately? A preliminary investigation. Journal of Sport Rehabilitation, J Sport Rehabil 1999;8(1):43-49.

[94] Geil MD. Consistency and accuracy of measurement of lower-limb amputee anthropometrics. J Rehabil Res Dev 2005 Mar-Apr;42(2):131-140.

[95] Yoo JH, Yi SR, Kim JH. The geometry of patella and patellar tendon measured on knee MRI. Surg Radiol Anat 2007 Dec;29(8):623-628.

[96] Introna F,Jr, Di Vella G, Campobasso CP. Sex determination by discriminant analysis of patella measurements. Forensic Sci Int 1998 Jul 6;95(1):39-45.

A Appendix

A.1 Results of the second measuring day

A.1.1 Inter-observer reproducibility of girth measurements (t2)

Table 29 and Table 30 show the results of the inter-observer agreement and reliability calculations, respectively, for the second measuring day. The corresponding Bland and Altman plots for the measurement sites at 7 cm proximal of mid-patella, mid-patella, and 7 cm distal of mid-patella are presented in Figure 36, Figure 37, and Figure 38, respectively. The SDD ranged from 0.4 to 1.8 cm and was highest for the measurement site at 7 cm proximal of mid-patella. The Waegener tape measure showed the highest agreement (SDD range, 0.6 to 1.0 cm), followed by the standard tape measures (SDD range, 0.7 to 1.2 cm) and the Gulick I tape measure (0.4 to 1.3 cm). As for the first measuring day, the SDD was highest for the Gulick II plus measure (range, 0.5 to 1.8 cm). The inter-observer reliability was generally high, with ICCs ranging from 0.94 to 0.99. Again, the lowest reliability was found for the Gulick II tape measure (ICC range, 0.94 to 0.99)

		O1	O2	Agreement t2: O1-O2			
		Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	Lower limit [95% CI]	Upper limit [95% CI]
PP	GI	38.8 ± 2.3	38.6 ± 2.5	0.2 [-0.0 to 0.4]	0.6	-1.0 [-1.4 to -0.6]	1.3 [1.0 to 1.7]
	GII	40.7 ± 2.4	40.5 ± 2.5	0.2 [-0.1 to 0.5]	0.8	-1.4 [-1.8 to -0.9]	1.8 [1.3 to 2.3]
	S	40.3 ± 2.5	40.2 ± 2.6	0.1 [0.0 to 0.3]	0.6	-0.9 [-1.3 to -0.6]	1.2 [0.9 to 1.5]
	W	38.8 ± 2.4	38.8 ± 2.6	0.1 [-0.1 to 0.2]	0.5	-0.8 [-1.1 to -0.6]	1.0 [0.7 to 1.2]
MP	GI	36.6 ± 2.2	36.6 ± 2.2	0.0 [-0.1 to 0.2]	0.4	-0.8 [-1.1 to -0.5]	0.8 [0.5 to 1.1]
	GII	37.6 ± 2.0	37.7 ± 2.1	-0.1 [-0.2 to -0.0]	0.3	-0.8 [-1.0 to -0.6]	0.5 [0.3 to 0.7]
	S	37.7 ± 2.0	37.5 ± 2.1	0.2 [0.1 to 0.3]	0.4	-0.5 [-0.7 to -0.3]	0.9 [0.7 to 1.1]
	W	37.0 ± 2.0	36.9 ± 2.1	0.1 [0.0 to 0.2]	0.3	-0.5 [-0.6 to -0.3]	0.7 [0.5 to 0.8]
DP	GI	33.4 ± 2.3	33.5 ± 2.4	-0.1 [-0.2 to 0.0]	0.3	-0.7 [-0.9 to -0.5]	0.4 [0.2 to 0.6]
	GII	34.2 ± 2.2	34.3 ± 2.3	-0.1 [-0.3 to 0.0]	0.4	-0.9 [-1.1 to -0.7]	0.7 [0.4 to 0.9]
	S	34.5 ± 2.1	34.4 ± 2.2	0.1 [-0.0 to 0.2]	0.3	-0.5 [-0.7 to -0.3]	0.7 [0.5 to 0.9]
	W	34.0 ± 2.1	34.2 ± 2.1	-0.2 [-0.3 to -0.0]	0.4	-0.9 [-1.2 to -0.7]	0.6 [0.4 to 0.8]

Table 29: Girth - Inter-observer agreement on second measuring day (O1 and O2)

Tape	ICC [95% CI]		
	PP	MP	DP
GI	0.97 [0.93 to 0.99]	0.98 [0.96 to 0.99]	0.99 [0.98 to 1.00]
GII	0.94 [0.89 to 0.97]	0.99 [0.97 to 0.99]	0.98 [0.97 to 0.99]
S	0.98 [0.95 to 0.99]	0.98 [0.95 to 0.99]	0.99 [0.98 to 0.99]
W	0.98 [0.97 to 0.99]	0.99 [0.98 to 0.99]	0.98 [0.96 to 0.99]

Table 30: Girth - Inter-observer reliability (ICC) on second measuring day (O1 and O2)

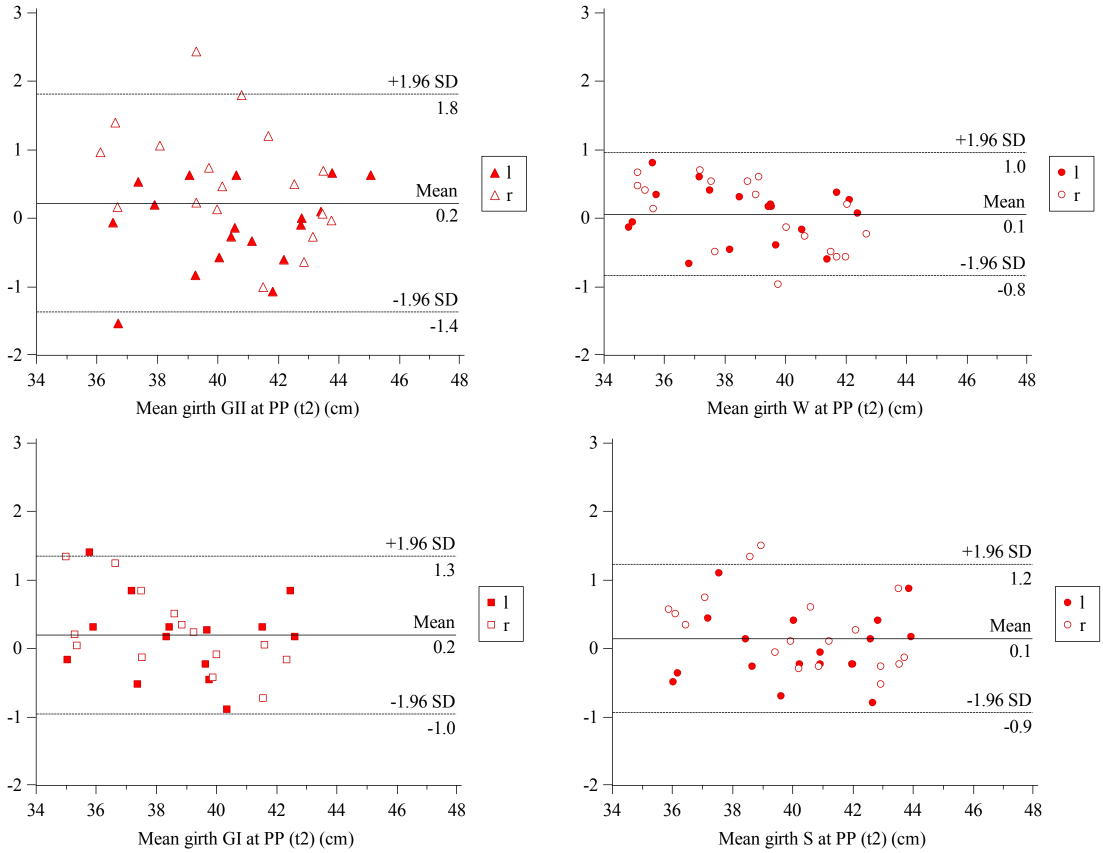


Figure 36: Girth - Inter-observer B-A plots at PP for the observers O1 and O2 (t2)
with mean difference between observers O1 and O2 (solid black line) and limits of agreement (broken black lines);

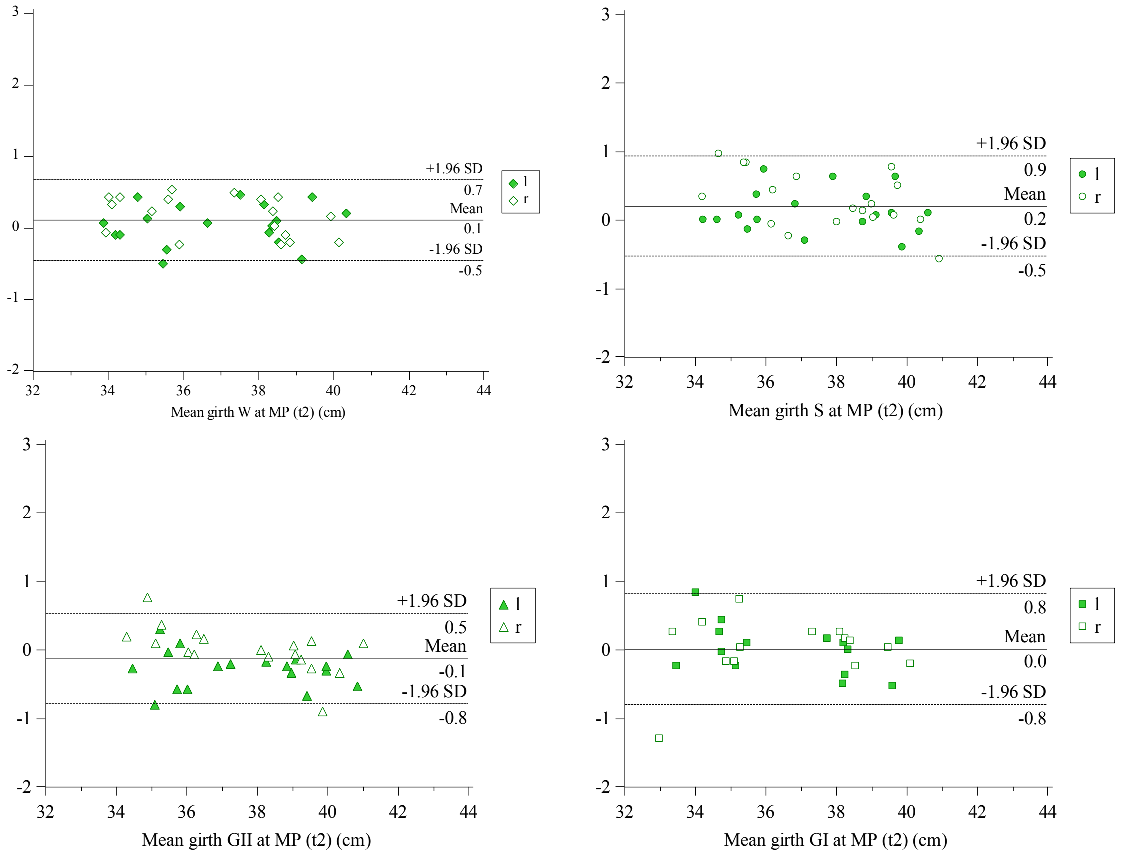


Figure 37: Girth - Inter-observer B-A plots at MP for the observers O1 and O2 (t2)

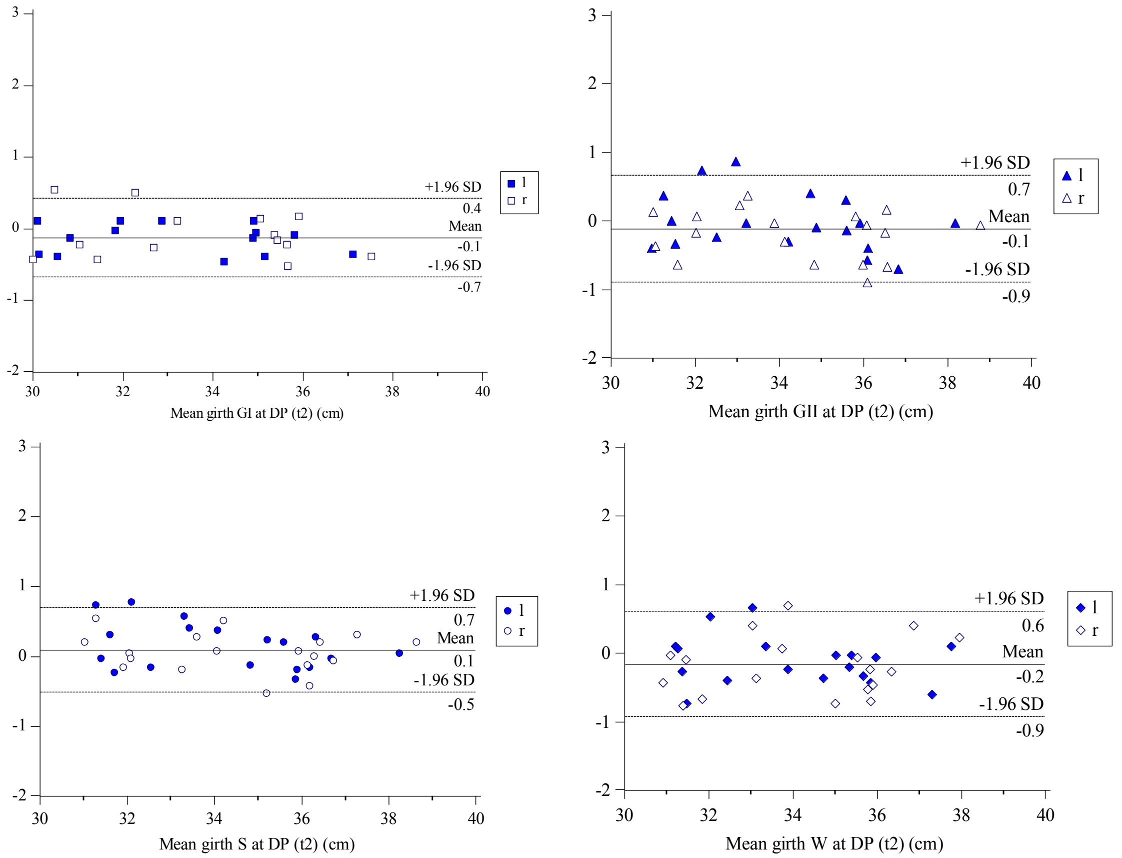


Figure 38: Girth - Inter-observer B-A plots at DP for the observers O1 and O2 (t2)

A.1.2 Inter-observer reproducibility of goniometric measurements (t2)

Table 31 shows the results of the inter-observer reproducibility for the second measuring day. The corresponding Bland and Altman plots are shown in Figure 39. Reliability and agreement were higher for the knee joint position P1 than P2. The SDD ranged from 7.6 to 9.0°, and the ICC from 0.85 to 0.98. Inter-observer reproducibility was higher for test position P1 than P2.

Position	O1 (°)	O2 (°)	Agreement: O1-O2 (°)				Reliability
	Mean ±SD	Mean ±SD	mD [95% CI]	SD_{diff}	Lower limit [95% CI]	Upper limit [95% CI]	ICC [95% CI]
P1	113.8 ±18.04	112.8 ±17.3	1.0 [-0.2 ; 2.1]	3.4	-5.6 [-7.7 ; -3.6]	7.6 [5.5 ; 9.6]	0.98 [0.96 ; 0.99]
P2	84.0 ±6.32	82.4 ±8.3	1.1 [-0.3 ; 2.5]	4.1	-6.9 [-9.3 ; -4.4]	9.0 [6.6 ; 11.5]	0.85 [0.73 ; 0.92]

Table 31: Flexion - Inter-observer reproducibility on second measuring day (O1 and O2)
mD, mean difference between measuring days; sample size, n=19

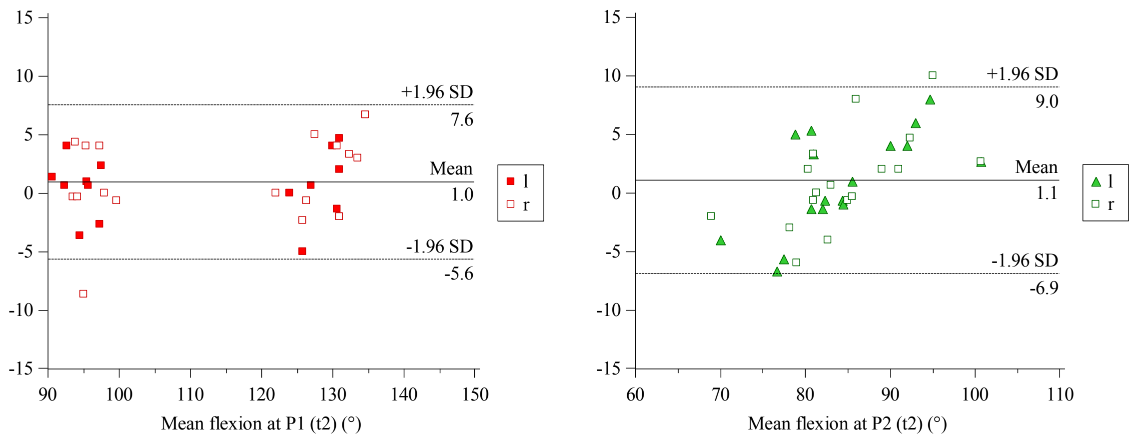


Figure 39: Flexion - Inter-observer B-A plots for second measuring day (O1 and O2)
for the knee joint positions P1 and P2; with mean difference between observers O1 and O2 (solid black line) and limits of agreement (broken black lines); l, left leg; r, right leg;

A.2 Results of the observers O1, O2, O3, O4 and O5

A.2.1 Reproducibility of girth measurements (O1-O5)

A.2.1.1 Inter-observer reliability (O1-O5)

A.2.1.1.1 First measuring day

The inter-observer reliability (ICC) measured by the observers O1, O2, O3, O4 and O5 on the first measuring day ranged from 0.94 to 0.98 (mean ICC 0.97) across three measurement sites and four tape measures (Table 32). The ICCs were higher at the mid-patella site (range, 0.97 to 0.98) than at 7 cm distal of mid-patella (0.97) and 7 cm proximal of mid-patella (range 0.94 to 0.97). Considering the different measuring tapes, the ICC values were slightly higher for the Gulick I and Gulick II tape measures (mean ICC 0.97) than for the standard and Waegener tape measures (mean ICC 0.96).
Comparing these results with those of the observers O1 and O2 measuring 20 subjects, reliability was slightly lower (mean ICC, 0.98 and 0.97, respectively).
Because the Bland and Altman limits of agreement were designed for pairwise comparison of measurements, this analysis could not be applied to the data of the five observers for inter-observer agreement analysis purposes.

Tape	Reliability: ICC [95% CI]		
	PP	MP	DP
GI	0.96 [0.91 to 0.99]	0.98 [0.94 to 0.99]	0.97 [0.92 to 0.99]
GII	0.97 [0.93 to 0.99]	0.97 [0.94 to 0.99]	0.97 [0.92 to 0.99]
S	0.94 [0.83 to 0.98]	0.98 [0.94 to 0.99]	0.97 [0.91 to 0.99]
W	0.94 [0.83 to 0.99]	0.98 [0.94 to 0.99]	0.97 [0.93 to 0.99]

Table 32: Girth - Inter-observer reliability for the observers O1-O5 (n=10 legs)

A.2.1.1.2 Second measuring day

Table 33 show the results of the reliability analysis for the second measuring day. The ICC ranged from 0.91 to 0.99 (mean ICC 0.97). Thus, inter-observer reliability was comparable between the first and second measuring day. Reliability was higher for the measurement sites at mid-patella (range, 0.97 to 0.98) and 7 cm distal of mid-patella (range 0.98 to 0.99) than at the measurement site at 7 cm proximal of mid-patella (range 0.91 to 0.98). The Gulick I tape measure showed the highest reliability (range 0.98 to 0.99), followed by the

Waegener tape measure (range 0.95 to 0.99) and the standard tape measure (range, 0.93 to 0.98). Level of reliability was lowest for the Gulick II plus tape measure (ICC range, 0.91 to 0.98).

Tape	ICC [95 CI]		
	PP	MP	DP
GI	0.98 [0.94 to 0.99]	0.98 [0.95 to 0.99]	0.99 [0.96 to 1.00]
GII	0.91 [0.74 to 0.97]	0.97 [0.89 to 0.99]	0.98 [0.92 to 0.99]
S	0.93 [0.80 to 0.98]	0.97 [0.92 to 0.99]	0.98 [0.96 to 1.00]
W	0.95 [0.82 to 0.99]	0.98 [0.93 to 0.99]	0.98 [0.95 to 1.00]

Table 33: Girth - Inter-observer reliability (ICC) on second measuring day (O1-O5)

A.2.1.2 Intra-observer reproducibility (O1-O5)

The results of intra-observer agreement for the observers O1, O2, O3, O4, and O5 are presented in Table 34, Table 35, Table 37 and Table 38, respectively. Table 39 shows the corresponding intra-observer reliability (ICC).

- **Tables**

Site	Tape	t1 (cm)	t2 (cm)	t1-t2 (cm)		Lower limit (cm)	Upper limit (cm)
		Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
PP	GI	38.7 ± 2.7	38.4 ± 2.7	0.3 [0.0 to 0.5]	0.4	-0.5 [-0.9 to 0.0]	1.0 [0.5 to 1.5]
	GII	40.2 ± 2.8	39.9 ± 2.8	0.3 [-0.1 to 0.8]	0.6	-0.9 [-1.7 to -0.1]	1.5 [0.7 to 2.3]
	S	39.8 ± 3.0	39.6 ± 2.8	0.3 [0.0 to 0.5]	0.4	-0.4 [-0.9 to 0.0]	1.0 [0.5 to 1.4]
	W	38.7 ± 2.9	38.5 ± 2.8	0.2 [-0.1 to 0.4]	0.3	-0.5 [-0.9 to -0.0]	0.9 [0.4 to 1.3]
MP	GI	37.2 ± 2.5	36.9 ± 2.5	0.3 [0.1 to 0.4]	0.2	-0.1 [-0.4 to 0.1]	0.7 [0.4 to 1.0]
	GII	38.0 ± 2.7	37.8 ± 2.5	0.2 [0.0 to 0.4]	0.3	-0.4 [-0.7 to 0.0]	0.8 [0.4 to 1.2]
	S	38.0 ± 2.6	37.7 ± 2.4	0.3 [0.0 to 0.5]	0.3	-0.4 [-0.8 to 0.0]	0.9 [0.5 to 1.3]
	W	37.5 ± 2.4	37.3 ± 2.3	0.2 [-0.1 to 0.4]	0.4	-0.5 [-0.9 to -0.0]	0.9 [0.4 to 1.3]
DP	GI	34.2 ± 2.4	33.8 ± 2.7	0.3 [0.1 to 0.6]	0.3	-0.3 [-0.8 to 0.1]	1.0 [0.6 to 1.4]
	GII	35.0 ± 2.5	34.7 ± 2.7	0.3 [0.0 to 0.6]	0.4	-0.5 [-1.1 to -0.0]	1.1 [0.6 to 1.6]
	S	35.1 ± 2.5	34.8 ± 2.7	0.3 [0.1 to 0.5]	0.3	-0.3 [-0.7 to 0.1]	0.9 [0.5 to 1.2]
	W	34.7 ± 2.3	34.4 ± 2.6	0.3 [0.0 to 0.5]	0.3	-0.4 [-0.8 to 0.0]	0.9 [0.5 to 1.3]

Table 34: Girth - Intra-observer agreement of observer O1 (n=10 legs)

Site	Tape	t1 (cm)	t2 (cm)	t1-t2 (cm)		Lower limit (cm)	Upper limit (cm)
		Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
PP	GI	38.5 ± 2.9	38.3 ± 2.9	0.2 [-0.2 to 0.7]	0.6	-1.0 [-1.7 to -0.2]	1.4 [0.7 to 2.2]
	GII	40.1 ± 2.8	39.7 ± 2.9	0.5 [0.1 to 0.8]	0.5	-0.5 [-1.1 to 0.1]	1.4 [0.8 to 2.0]
	S	39.6 ± 2.8	39.5 ± 2.9	0.1 [-0.3 to 0.5]	0.6	-1.1 [-1.9 to -0.3]	1.3 [0.5 to 2.1]
	W	38.6 ± 2.8	38.3 ± 2.9	0.3 [0.0 to 0.5]	0.4	-0.4 [-0.9 to 0.0]	1.0 [0.5 to 1.4]
MP	GI	37.0 ± 2.5	37.0 ± 2.6	0.0 [-0.3 to 0.3]	0.4	-0.8 [-1.3 to -0.3]	0.7 [0.2 to 1.3]
	GII	38.0 ± 2.4	38.0 ± 2.5	0.1 [-0.2 to 0.3]	0.3	-0.6 [-1.0 to -0.2]	0.7 [0.3 to 1.2]
	S	37.9 ± 2.4	37.7 ± 2.5	0.1 [0.0 to 0.3]	0.2	-0.3 [-0.6 to -0.0]	0.6 [0.3 to 0.9]
	W	37.4 ± 2.1	37.3 ± 2.2	0.2 [- 0.1 to 0.4]	0.3	-0.4 [-0.8 to -0.0]	0.7 [0.4 to 1.1]
DP	GI	34.1 ± 2.5	34.1 ± 2.7	0.0 [-0.3 to 0.3]	0.4	-0.8 [-1.3 to -0.3]	0.7 [0.3 to 1.2]
	GII	34.9 ± 2.5	34.8 ± 2.8	0.1 [-0.1 to 0.4]	0.4	-0.6 [-1.0 to -0.1]	0.8 [0.4 to 1.3]
	S	34.9 ± 2.4	34.9 ± 2.6	0.0 [-0.2 to 0.2]	0.2	-0.4 [-0.7 to -0.1]	0.5 [0.2 to 0.8]
	W	34.9 ± 2.4	34.7 ± 2.4	0.2 [-0.1 to 0.5]	0.5	-0.7 [-1.3 to -0.1]	1.1 [0.5 to 1.7]

Table 35: Girth - Intra-observer agreement of observer O2 (n=10 legs)

Site	Tape	t1 (cm)	t2 (cm)	t1-t2 (cm)		Lower limit (cm)	Upper limit (cm)
		Mean ± SD	Mean ±SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
PP	GI	38.1 ± 2.8	37.9 ± 2.8	0.2 [-0.2 to 0.6]	0.5	-0.9 [-1.5 to -0.2]	1.2 [0.6 to 1.9]
	GII	39.6 ± 3.1	38.6 ± 2.9	1.0 [0.2 to 1.9]	1.2	-1.3 [-2.8 to 0.2]	3.3 [1.8 to 4.8]
	S	40.0 ± 3.0	39.4 ± 2.7	0.6 [0.2 to 1.0]	0.6	-0.6 [-1.4 to 0.2]	1.8 [1.0 to 2.5]
	W	39.0 ± 2.9	38.8 ± 2.6	0.3 [-0.2 to 0.8]	0.7	-1.1 [-2.1 to -0.2]	1.7 [0.8 to 2.7]
MP	GI	36.7 ± 2.6	36.5 ± 2.5	0.1 [-0.2 to 0.5]	0.5	-0.8 [-1.5 to -0.2]	1.1 [0.5 to 1.8]
	GII	37.6 ± 2.6	37.1 ± 2.5	0.4 [0.0 to 0.9]	0.6	-0.7 [-1.4 to 0.0]	1.6 [0.9 to 2.38]
	S	37.7 ± 2.6	37.3 ± 2.4	0.3 [-0.1 to 0.7]	0.5	-0.7 [-1.4 to -0.0]	1.3 [0.7 to 2.08]
	W	37.5 ± 2.5	37.1 ± 2.2	0.3 [-0.1 to 0.8]	0.6	-0.9 [-1.7 to -0.1]	1.5 [0.7 to 2.3]
DP	GI	33.8 ± 2.7	33.6 ± 2.6	0.2 [-0.1 to 0.6]	0.5	-0.8 [-1.4 to -0.1]	1.2 [0.6 to 1.9]
	GII	34.5 ± 2.8	34.2 ± 2.6	0.4 [0.0 to 0.7]	0.5	-0.6 [-1.2 to 0.0]	1.3 [0.7 to 2.0]
	S	34.6 ± 2.8	34.4 ± 2.3	0.2 [-0.3 to 0.7]	0.7	-1.1 [-1.9 to -0.2]	1.5 [0.7 to 2.4]
	W	34.8 ± 2.6	34.7 ± 2.4	0.1 [-0.2 to 0.5]	0.5	-0.8 [-1.4 to -0.2]	1.1 [0.5 to 1.7]

Table 36: Girth - Intra-observer agreement of observer O3 (n=10 legs)

Site	Tape	t1 (cm)	t2 (cm)	t1-t2 (cm)		Lower limit (cm)	Upper limit (cm)
		mean ± SD	mean ±SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
PP	GI	38.4 ± 3.3	38.1 ± 3.1	0.3 [-0.1 to 0.8]	0.6	-0.8 [-1.5 to -0.1]	1.5 [0.7 to 2.2]
	GII	40.2 ± 3.4	40.0 ± 3.2	0.3 [-0.3 to 0.8]	0.8	-1.3 [-2.3 to -0.3]	1.8 [0.8 to 2.8]
	S	40.6 ± 3.5	40.2 ± 3.5	0.4 [-0.1 to 0.8]	0.6	-0.8 [-1.6 to -0.1]	1.6 [0.8 to 2.3]
	W	39.4 ± 3.2	39.0 ± 3.1	0.4 [0.1 to 0.6]	0.4	-0.4 [-0.9 to 0.1]	1.1 [0.6 to 1.6]
MP	GI	36.9 ± 2.8	36.8 ± 2.5	0.1 [-0.2 to 0.5]	0.5	-0.8 [-1.4 to -0.2]	1.1 [0.5 to 1.7]
	GII	38.1 ± 2.8	37.9 ± 2.6	0.1 [-0.3 to 0.5]	0.5	-0.9 [-1.6 to -0.2]	1.2 [0.5 to 1.8]
	S	38.2 ± 3.0	38.0 ± 2.7	0.2 [-0.2 to 0.6]	0.6	-0.9 [-1.6 to -0.2]	1.3 [0.6 to 2.0]
	W	37.7 ± 2.6	37.4 ± 2.4	0.3 [0.0 to 0.6]	0.4	-0.4 [-0.9 to 0.0]	1.0 [0.6 to 1.5]
DP	GI	34.0 ± 2.7	33.7 ± 2.6	0.2 [-0.0 to 0.4]	0.3	-0.5 [-0.9 to -0.0]	0.9 [0.4 to 1.3]
	GII	35.0 ± 2.7	34.8 ± 2.6	0.2 [-0.1 to 0.5]	0.4	-0.7 [-1.2 to -0.1]	1.1[0.5 to 1.6]
	S	35.1 ± 2.8	34.8 ± 2.7	0.3 [0.0 to 0.6]	0.4	-0.5 [-1.0 to 0.0]	1.1[0.6 to 1.6]
	W	35.2 ± 2.5	35.0 ± 2.3	0.2 [-0.0 to 0.5]	0.3	-0.4 [-0.8 to 0.0]	0.9 [0.5 to 1.3]

Table 37: Girth - Intra-observer agreement of observer O4 (n=10 legs)

Site	Tape	t1(cm)	t2 (cm)	t1-t2 (cm)		Lower limit (cm)	Upper limit (cm)
		mean ± SD	Mean ±SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
PP	GI	38.5 ± 3.4	38.2 ± 2.7	0.3 [-0.3 to 0.9]	0.8	-1.4 [-2.4 to -0.3]	1.9 [0.8 to 2.9]
	GII	40.5 ± 3.4	40.4 ± 2.6	0.0 [-0.9 to 0.9]	1.3	-2.4 [-4.0 to -0.8]	2.5 [0.9 to 4.1]
	S	40.9 ± 3.3	40.8 ± 2.8	0.2 [-0.4 to 0.7]	0.8	-1.3 [-2.3 to -0.4]	1.6 [0.7 to 2.6]
	W	39.9 ± 3.1	39.6 ± 2.8	0.3 [-0.3 to 0.9]	0.9	-1.4 [-2.5 to -0.3]	2.0 [0.9 to 3.1]
MP	GI	37.2 ± 2.7	37.1 ± 2.3	0.1 [-0.3 to 0.5]	0.5	-0.9 [-1.6 to -0.3]	1.1 [0.5 to 1.8]
	GII	38.1 ± 2.8	38.0 ± 2.6	0.1 [-0.3 to 0.5]	0.6	-1.0 [-1.7 to -0.3]	1.2 [0.5 to 1.9]
	S	38.2 ± 2.7	38.1 ± 2.5	0.1 [-0.1 to 0.3]	0.2	-0.4 [-0.7 to -0.1]	0.6 [0.3 to 0.9]
	W	37.8 ± 2.7	37.8 ± 2.5	0.0 [-0.2 to 0.3]	0.3	-0.6 [-1.0 to -0.2]	0.7 [0.3 to 1.1]
DP	GI	34.5 ± 2.7	34.0 ± 2.5	0.5 [0.1 to 1.0]	0.6	-0.7 [-1.5 to 0.1]	1.7 [1.0 to 2.5]
	GII	35.3 ± 2.8	34.9 ± 2.4	0.4 [-0.2 to 1.0]	0.8	-1.3 [-2.3 to -0.2]	2.0 [1.0 to 3.1]
	S	35.4 ± 2.7	34.9 ± 2.7	0.6 [0.0 to 1.1]	0.7	-0.9 [-1.8 to 0.1]	2.0 [1.0 to 2.9]
	W	35.0 ± 2.7	34.5 ± 2.5	0.5 [0.0 to 1.0]	0.7	-0.8 [-1.6 to 0.1]	1.8 [1.0 to 2.7]

Table 38: Girth - Intra-observer agreement of observer O5 (n=10 legs)

Site	Tape	Reliability t1-t2: ICC [95% CI]				
		O1	O2	O3	O4	O5
PP	GI	0.99 [0.92 to 1.00]	0.98 [0.91 to 0.99]	0.98 [0.93 to 1.00]	0.98 [0.91 to 1.00]	0.96 [0.87 to 0.99]
	GII	0.97 [0.89 to 0.99]	0.97 [0.76 to 0.99]	0.88 [0.42 to 0.97]	0.97 [0.89 to 0.99]	0.92 [0.72 to 0.98]
	S	0.99 [0.93 to 1.00]	0.98 [0.92 to 0.99]	0.96 [0.64 to 0.99]	0.98 [0.92 to 1.00]	0.97 [0.89 to 0.99]
	W	0.99 [0.96 to 1.00]	0.99 [0.92 to 1.00]	0.96 [0.86 to 0.99]	0.99 [0.89 to 1.00]	0.95 [0.84 to 0.99]
MP	GI	0.99 [0.80 to 1.00]	0.99 [0.96 to 1.00]	0.98 [0.93 to 1.00]	0.98 [0.94 to 1.00]	0.98 [0.92 to 0.99]
	GII	0.99 [0.94 to 1.00]	0.99 [0.97 to 1.00]	0.96 [0.79 to 0.99]	0.98 [0.93 to 1.00]	0.98 [0.93 to .100]
	S	0.99 [0.90 to 1.00]	0.99 [0.98 to 1.00]	0.97 [0.88 to 0.99]	0.98 [0.93 to 1.00]	1.00 [0.98 to 1.00]
	W	0.99 [0.94 to 1.00]	0.99 [0.96 to 1.00]	0.96 [0.85 to 0.99]	0.98 [0.89 to 1.00]	0.99 [0.97 to 1.00]
DP	GI	0.98 [0.81 to 1.00]	0.99 [0.96 to 1.00]	0.98 [0.93to 1.00]	0.99 [0.95 to 1.00]	0.95 [0.70 to 1.00]
	GII	0.98 [0.92 to 1.00]	0.98 [0.87 to 0.99]	0.99 [0.96 to 1.00]	0.98 [0.94 to 1.00]	0.94 [0.79 to 0.99]
	S	0.99 [0.88 to 1.00]	1.00 [0.99 to 1.00]	0.97 [0.88 to 0.99]	0.98 [0.90 to 1.00]	0.95 [0.73 to 0.99]
	W	0.99 [0.90 to 1.00]	0.98 [0.92 to 0.99]	0.98 [0.93 to 1.00]	0.99 [0.94 to 1.00]	0.95 [0.73 to 0.99]

Table 39: Girth - Intra-observer reliability for the observers O1-O5

- **Bland and Altman plots**

The B-A plots for the observers O1, O1, O3, O4 and O5 at the measurement site 7 cm proximal of mid-patella are shown in Figure 40, Figure 41, Figure 42, Figure 43 and Figure 44, respectively. It has to be taken into account that the scaling of the y-axis changes differs from observer to observer. In Figure 45, Figure 46, Figure 47, Figure 48 and Figure 49, the B-A plots for the measurement site at mid-patella are presented. Finally, the B-A plots for the measurement site at 7 cm distal of mid-patella are displayed in Figure 50, Figure 51, Figure 52, Figure 53 and Figure 54. In these plots, again, the scaling of the y-axis differs from observer to observer.

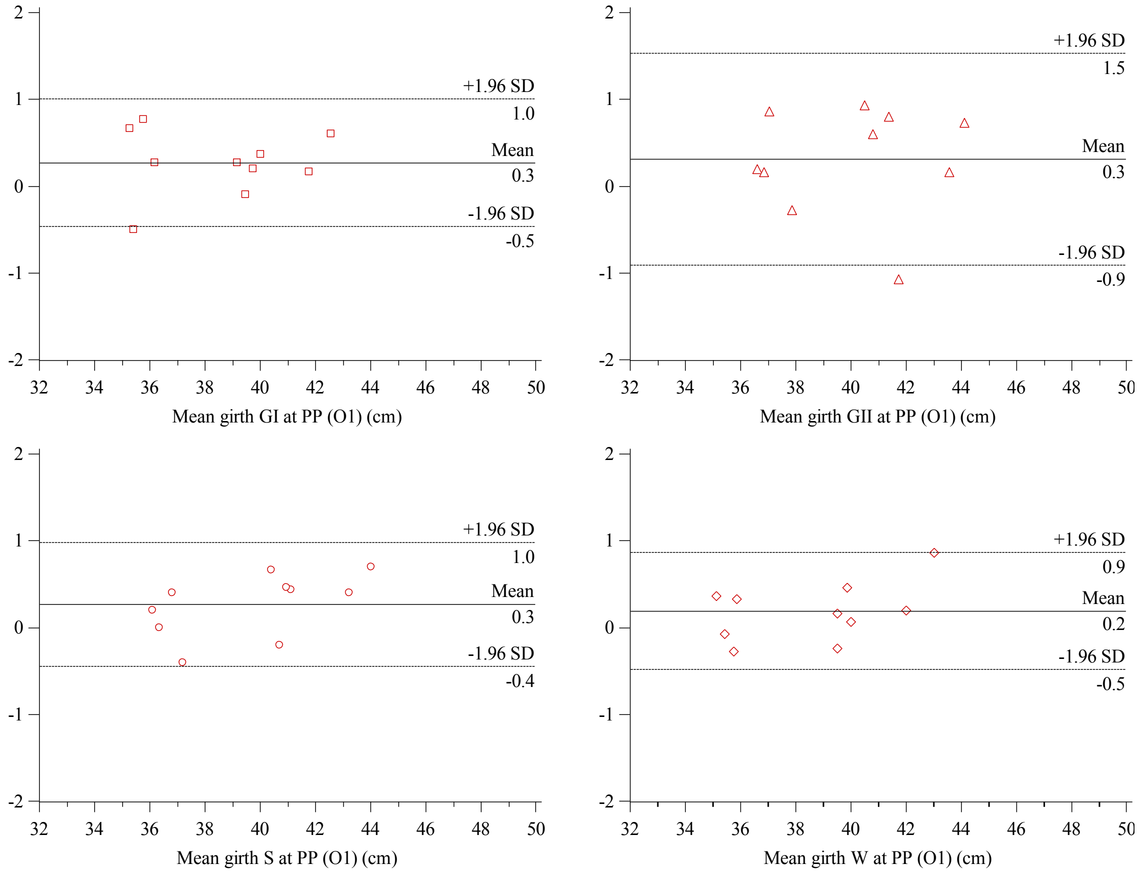


Figure 40: Girth - Intra-observer B-A plots at PP for the observer O1 (n=10 legs)

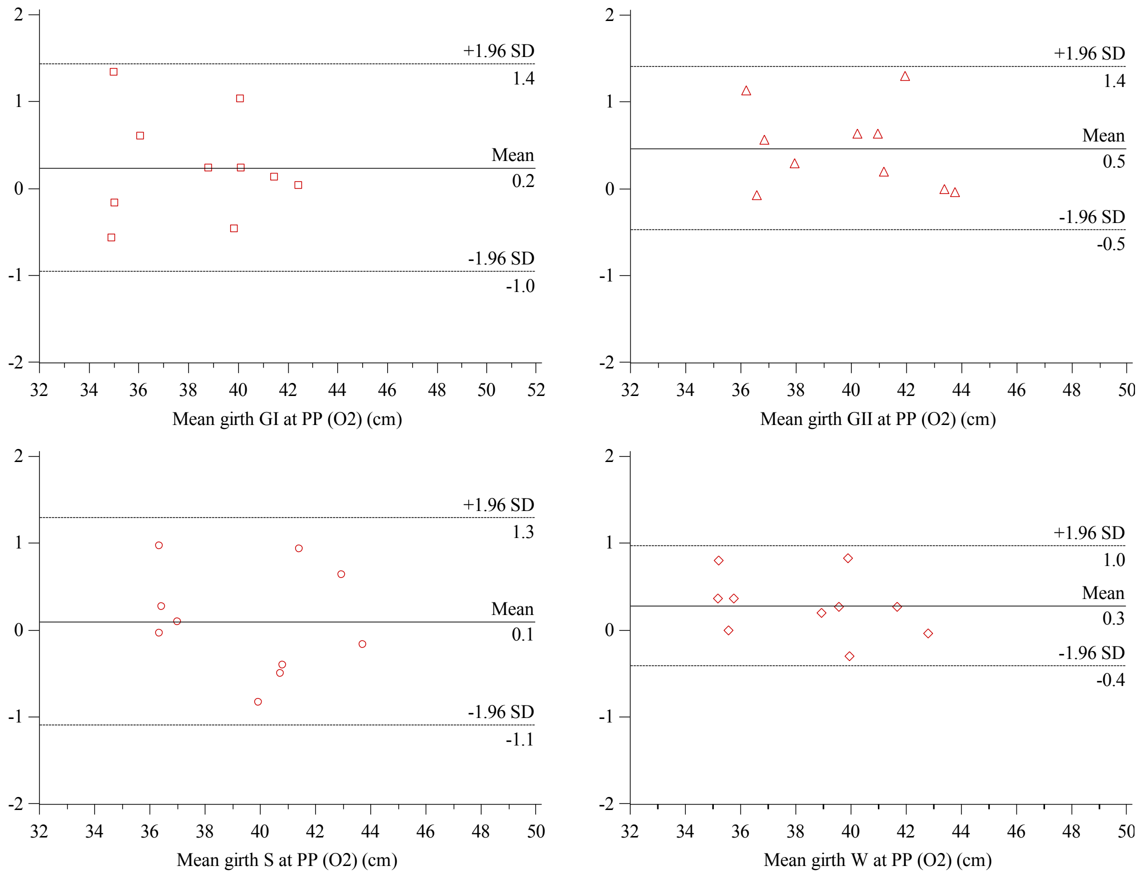


Figure 41: Girth - Intra-observer B-A plots at PP for the observer O2 (n=10 legs)

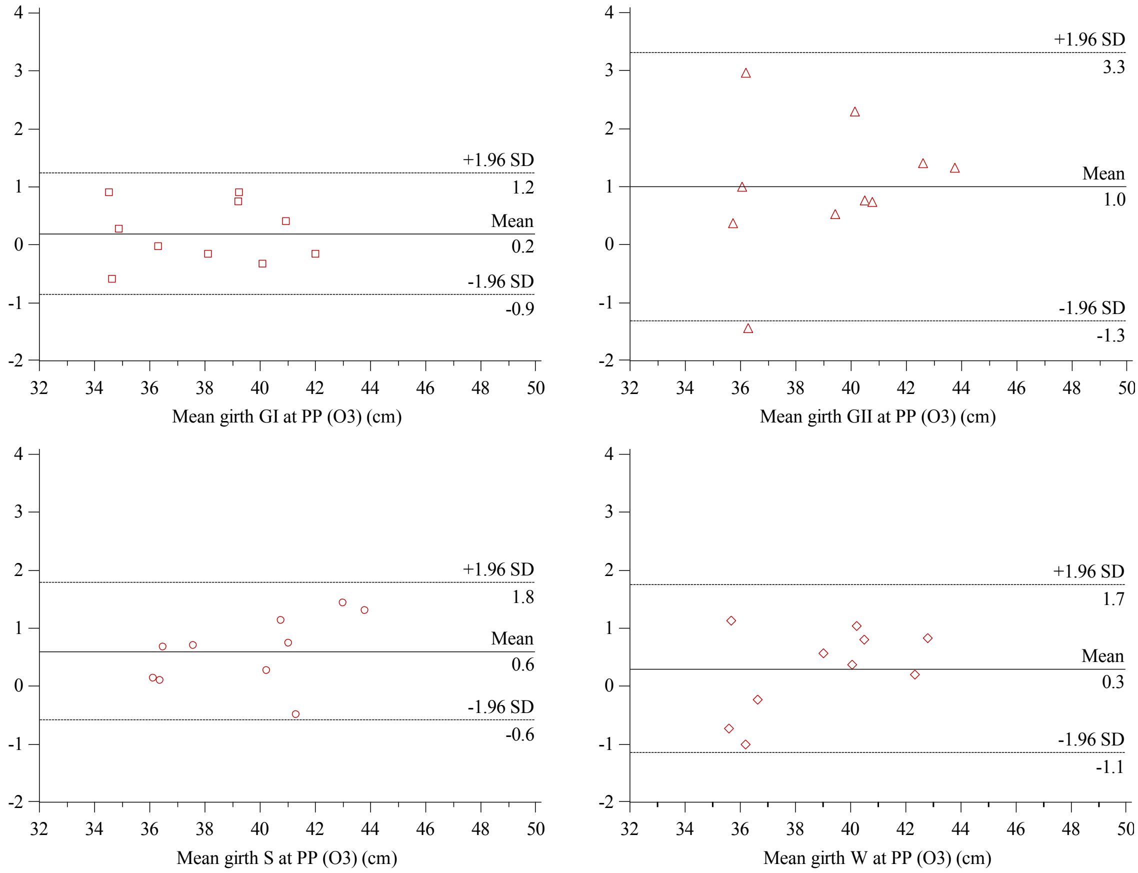


Figure 42: Girth - Intra-observer B-A plots at PP for the observer O3 (n=10 legs)

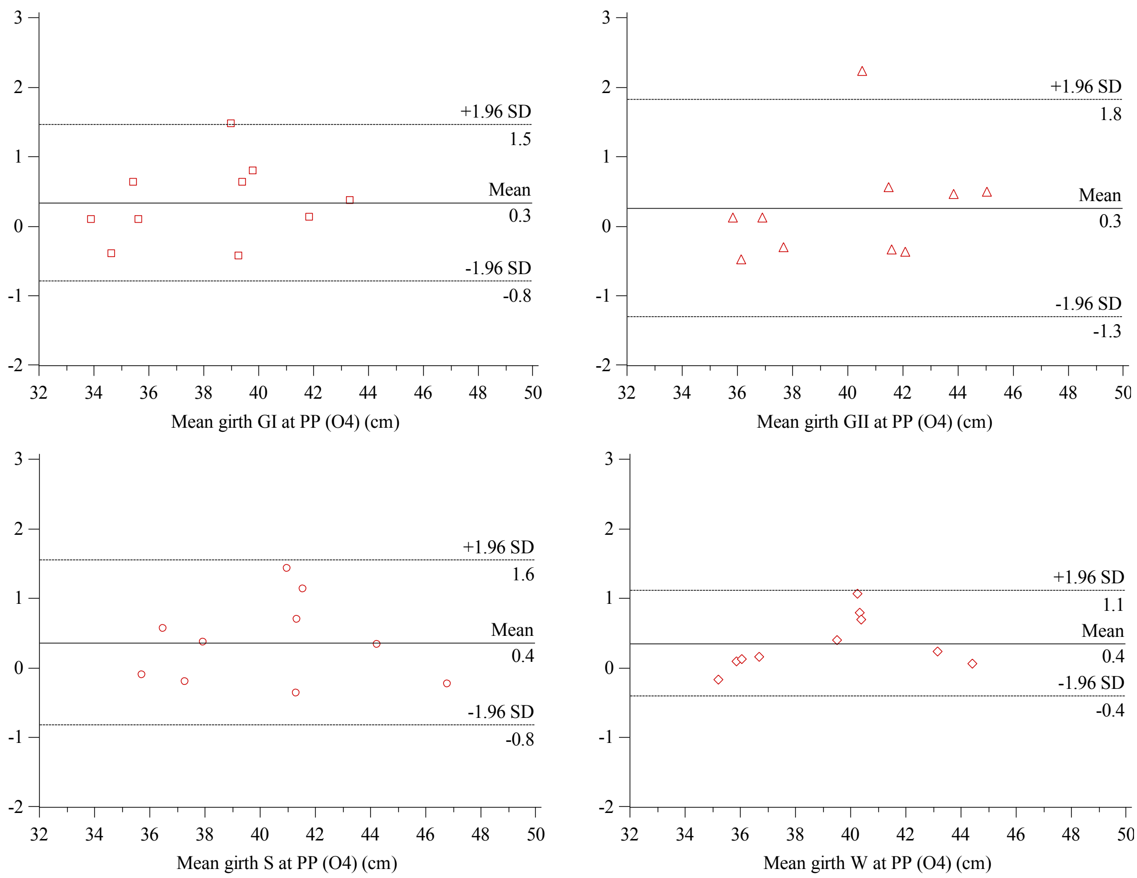


Figure 43: Girth - Intra-observer B-A plots at PP for the observer O4 (n=10 legs)

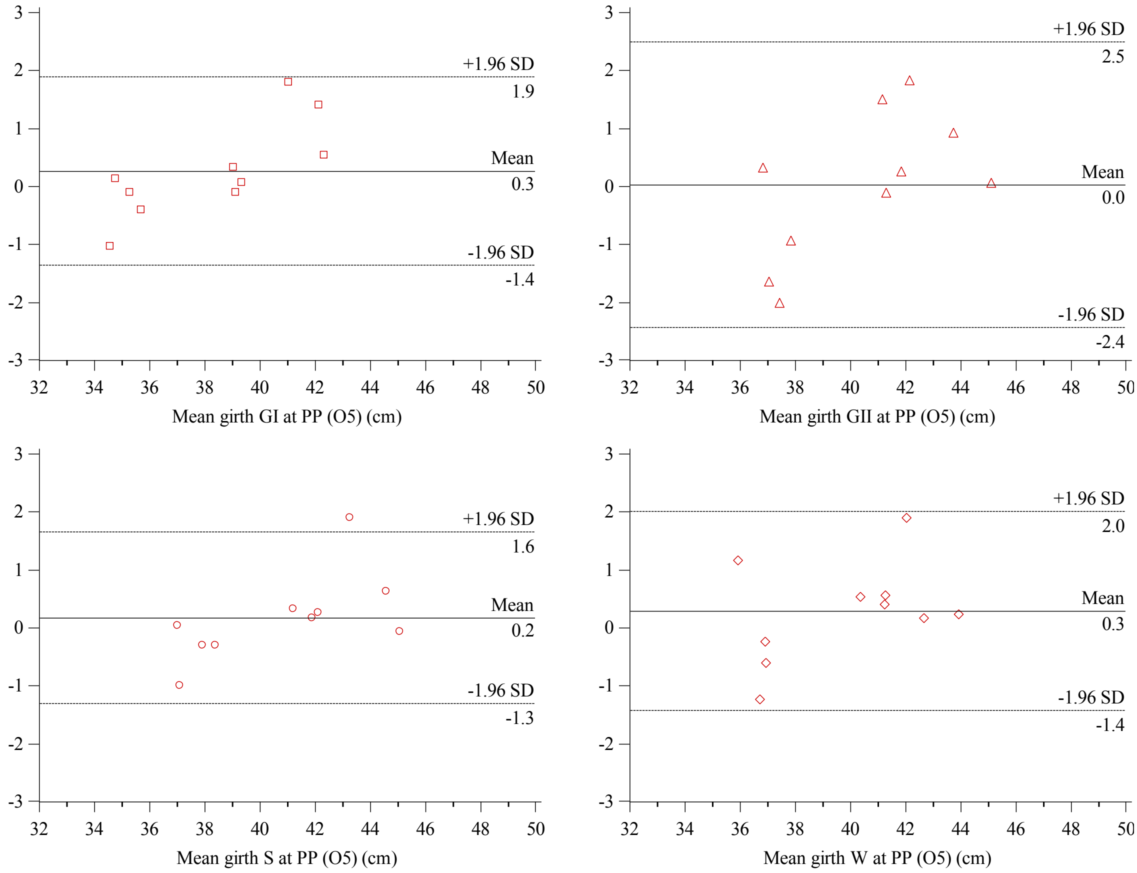


Figure 44: Girth - Intra-observer B-A plots at PP for the observer O5 (n=10 legs)

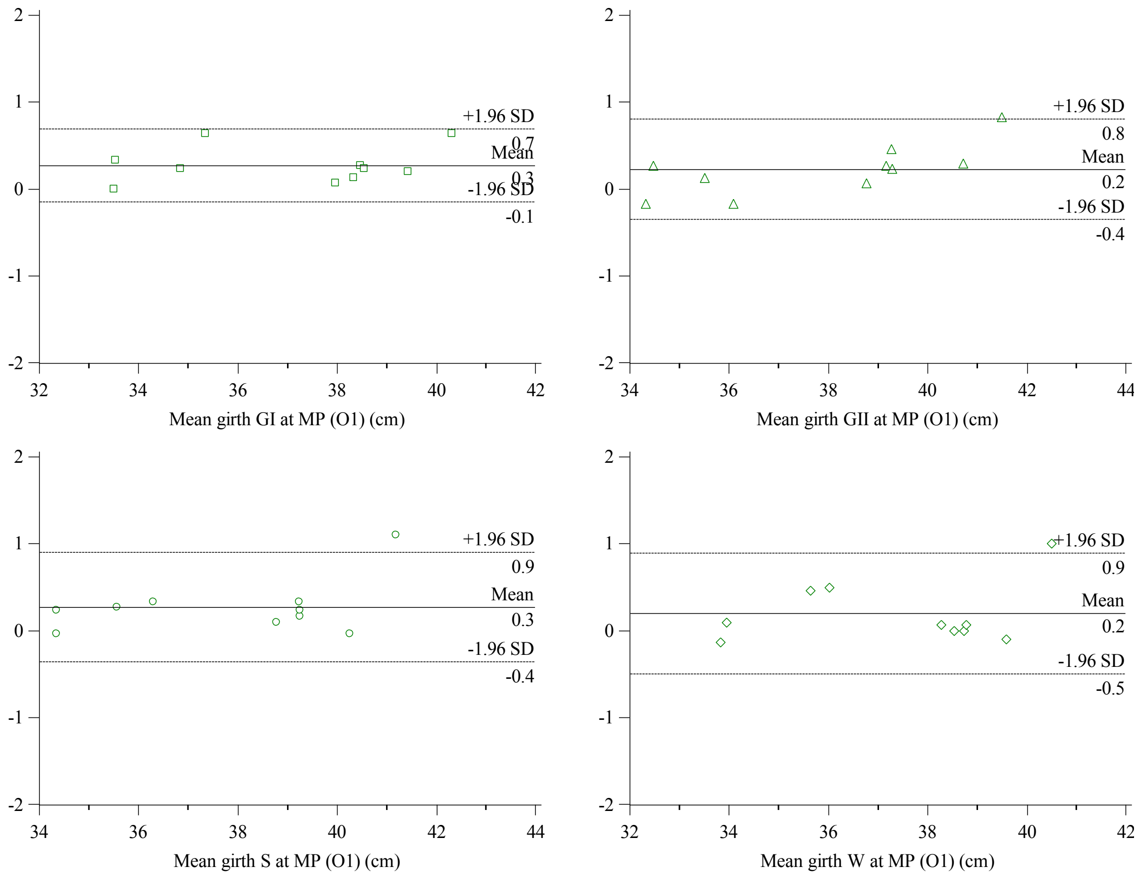


Figure 45: Girth - Intra-observer B-A plots at MP for observer O1 (n=10 legs)

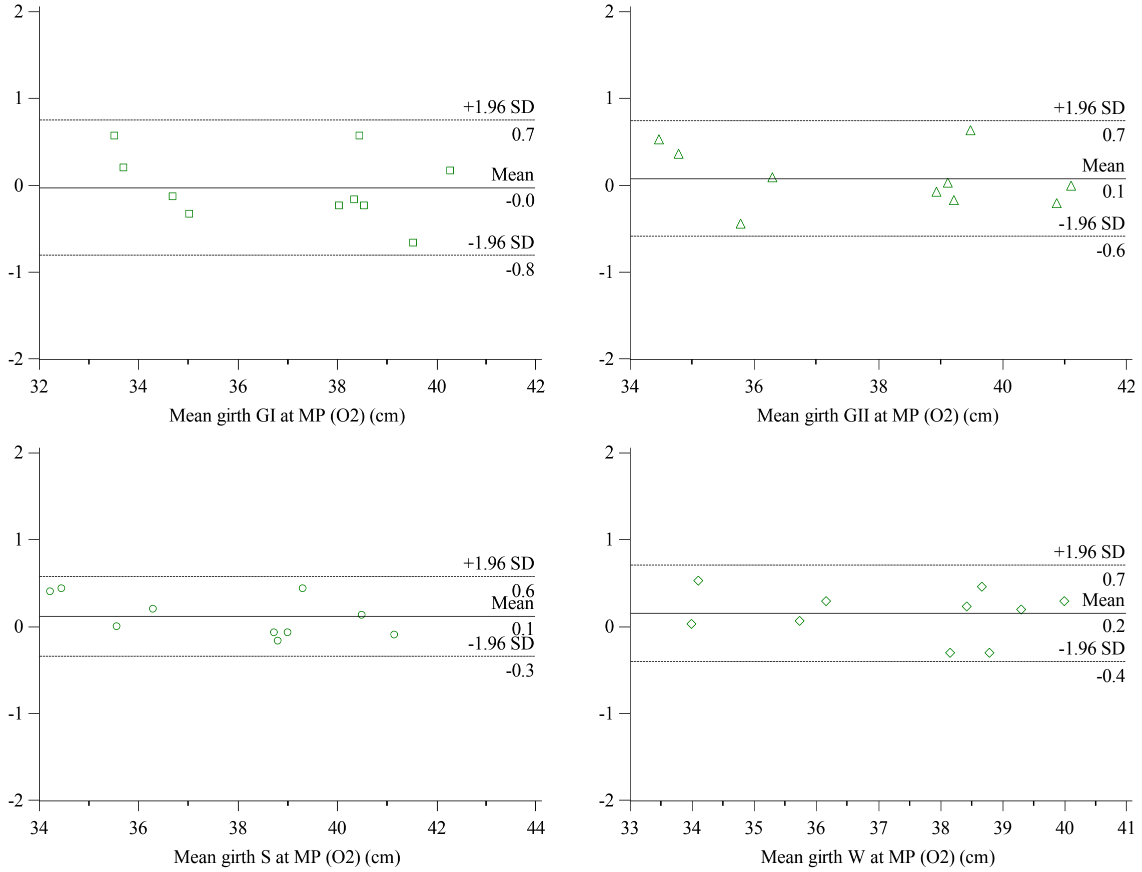


Figure 46: Girth - Intra-observer B-A plots at MP for observer O2 (n=10 legs)

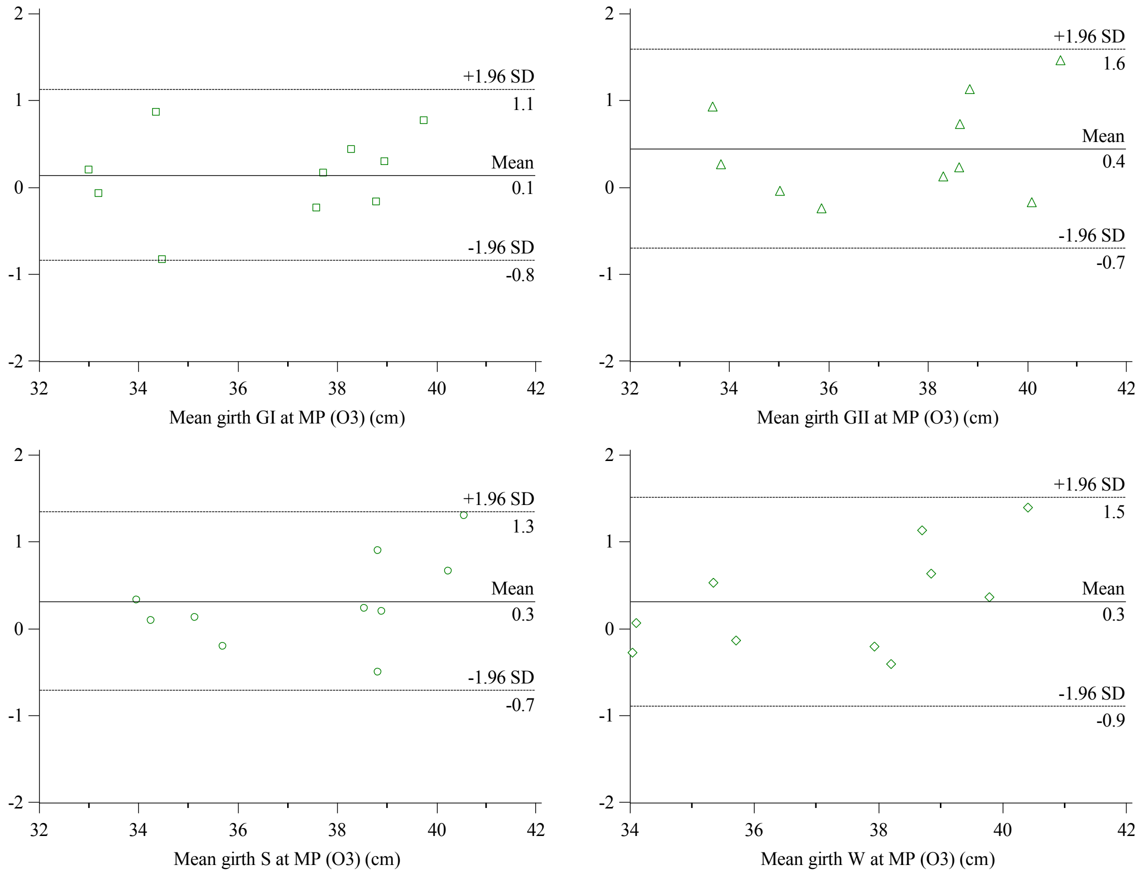


Figure 47: Girth - Intra-observer B-A plots at MP for observer O3 (n=10 legs)

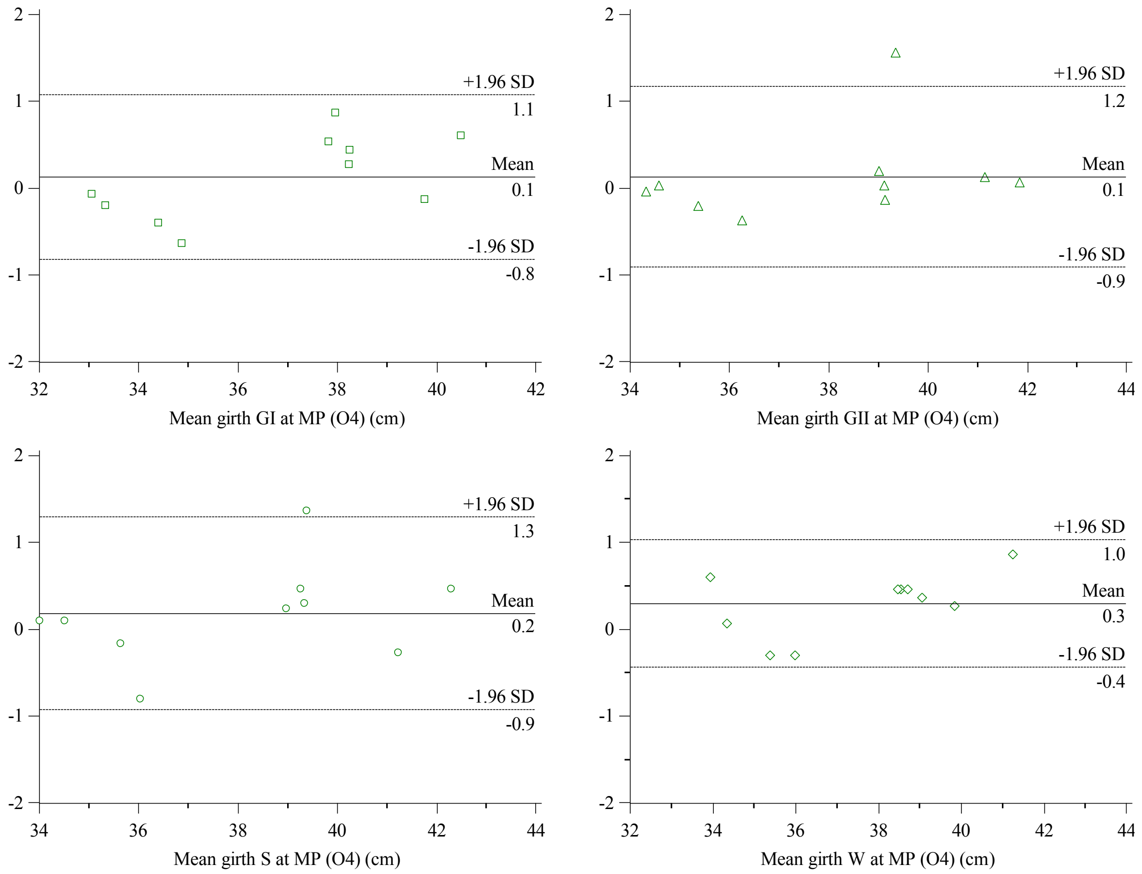


Figure 48: Girth - Intra-observer B-A plots at MP for observer O4 (n=10 legs)

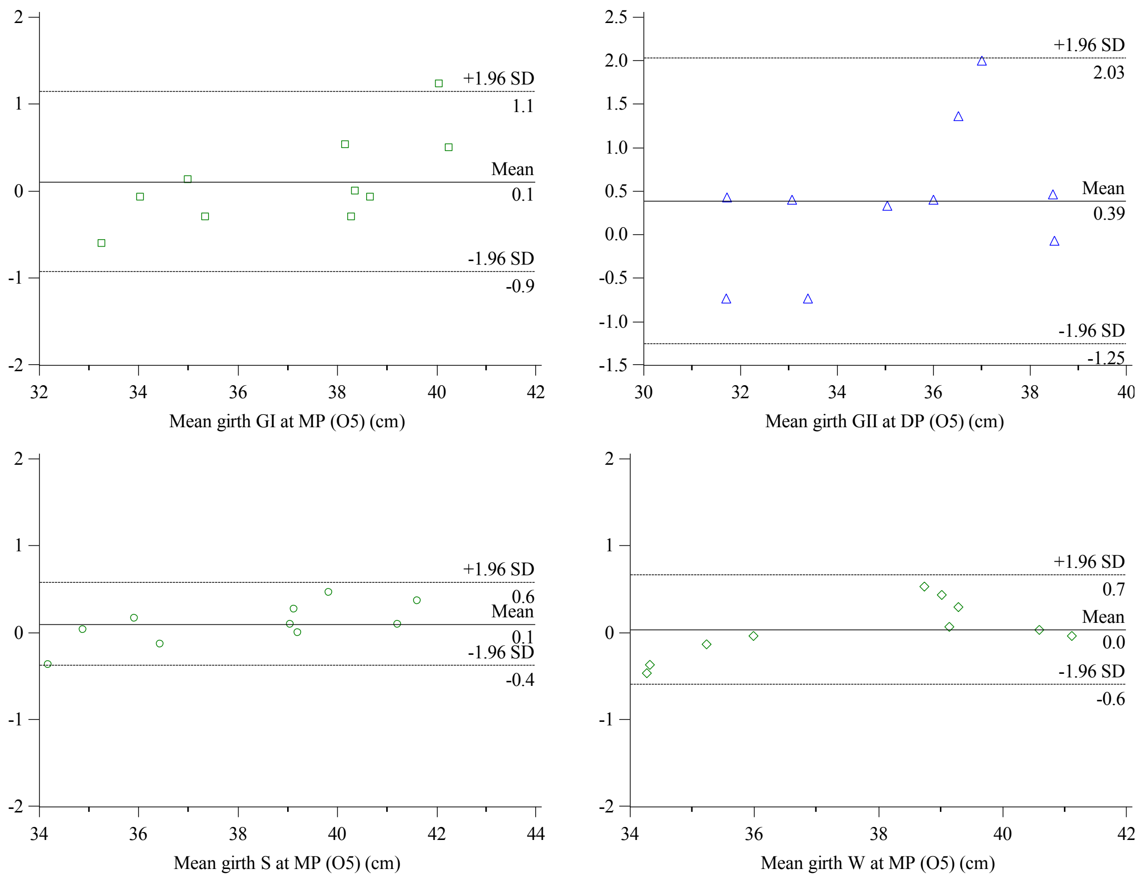


Figure 49: Girth - Intra-observer B-A plots at MP for observer O5 (n=10 legs)

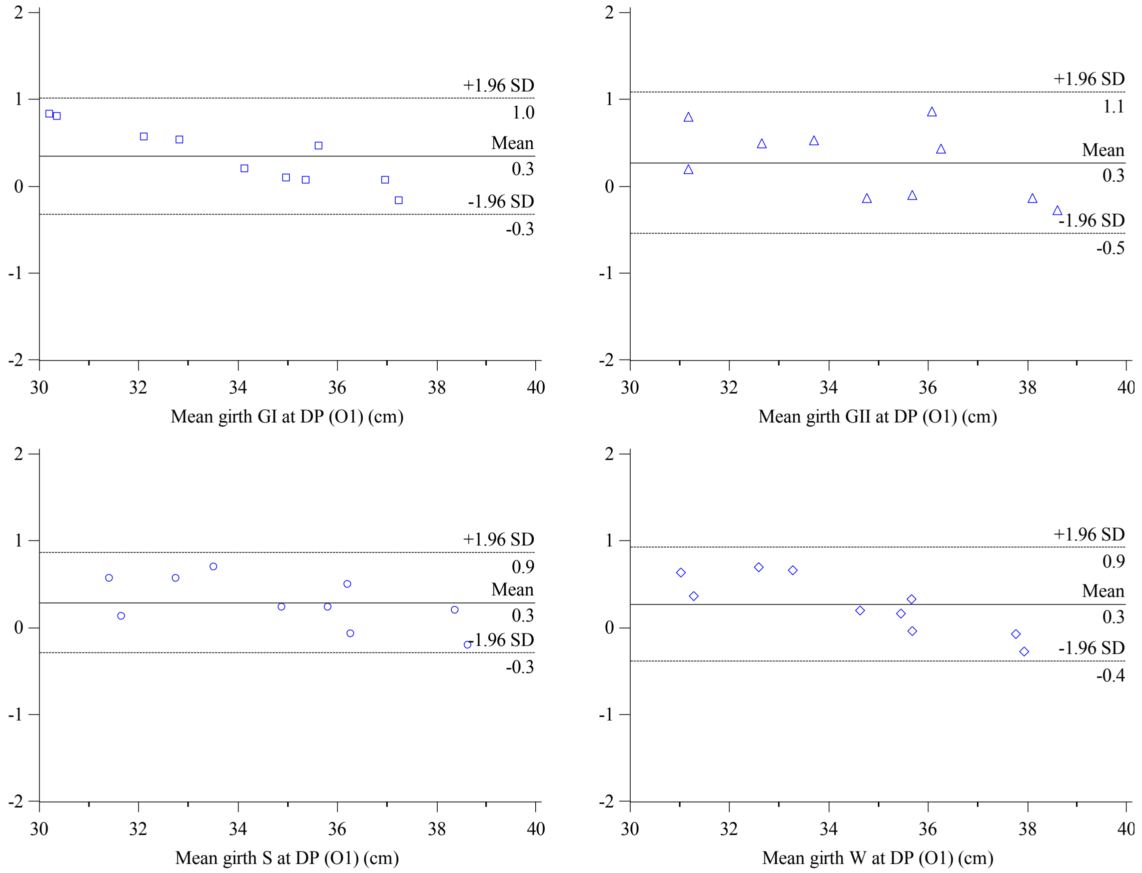


Figure 50: Girth - Intra-observer B-A plots at DP for observer O1 (n=10 legs)

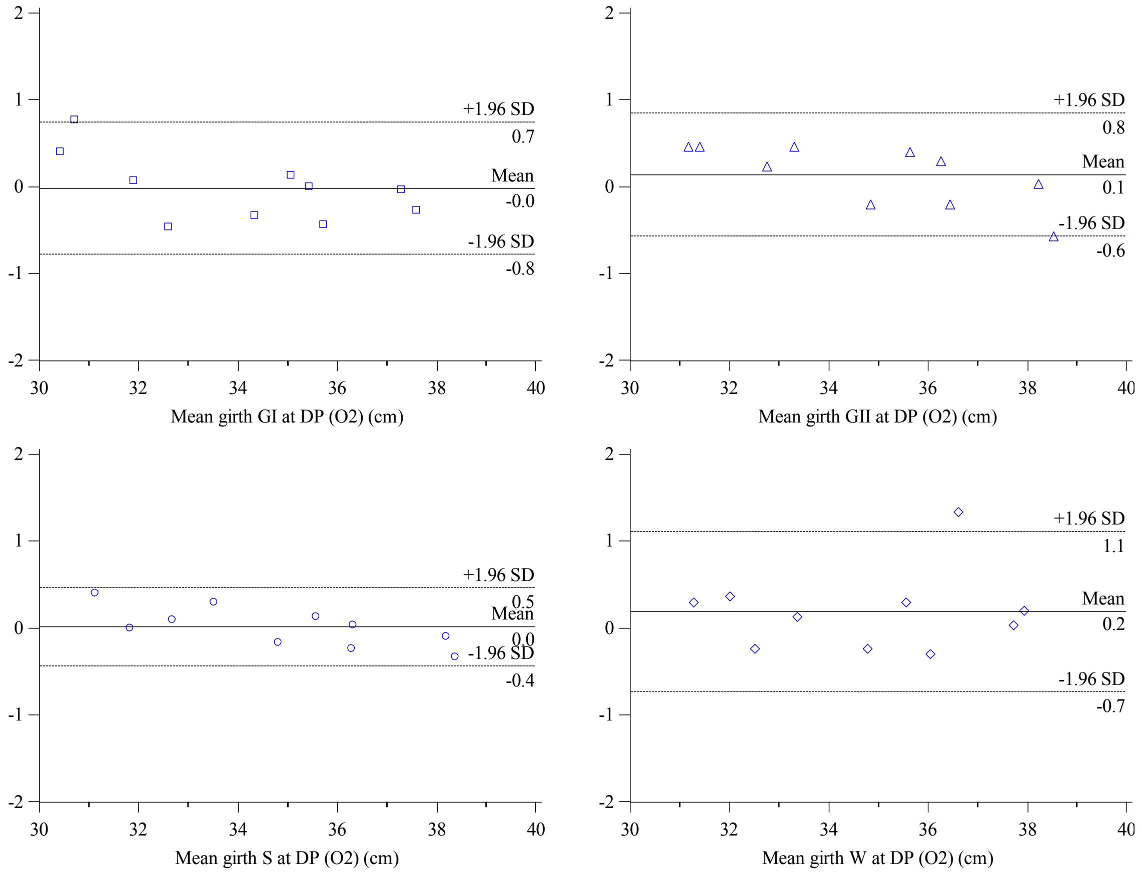


Figure 51: Girth - Intra-observer B-A plots at DP for observer O2 (n=10 legs)

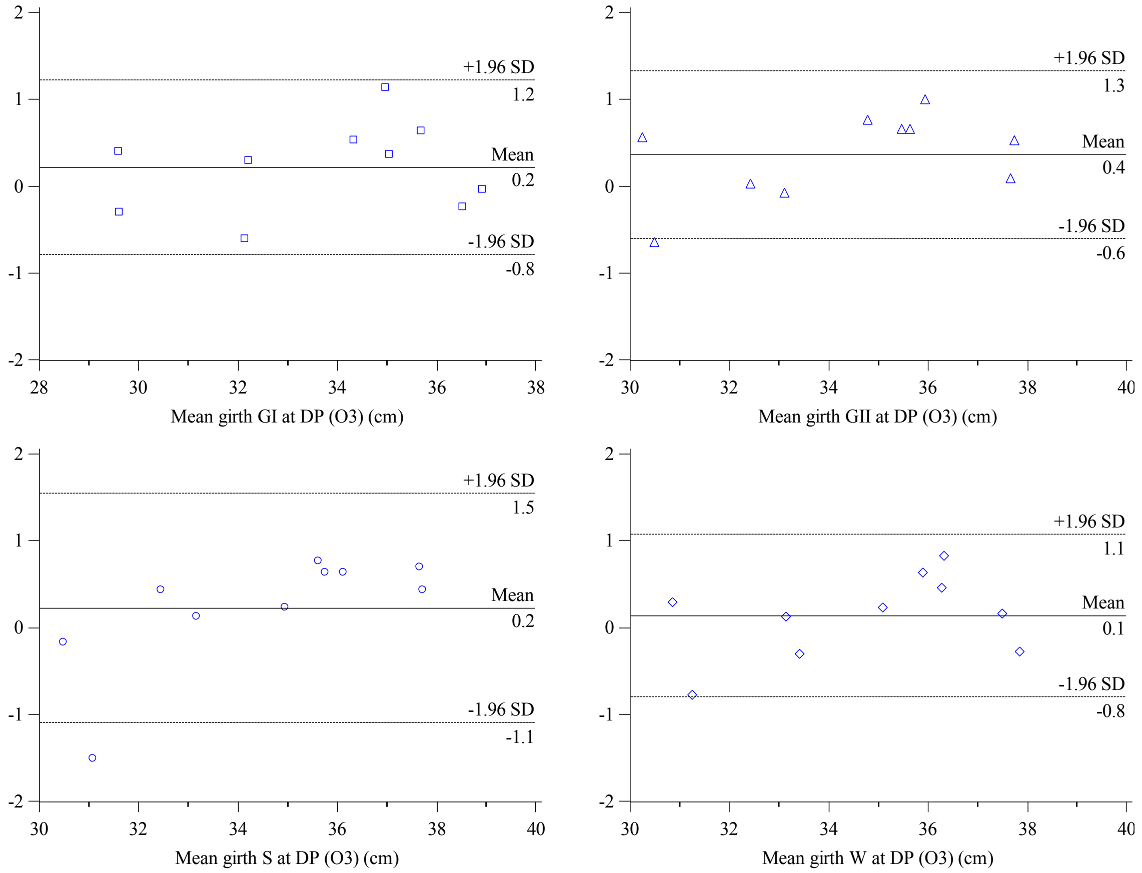


Figure 52: Girth - Intra-observer B-A plots at DP for observer O2 (n=10 legs)

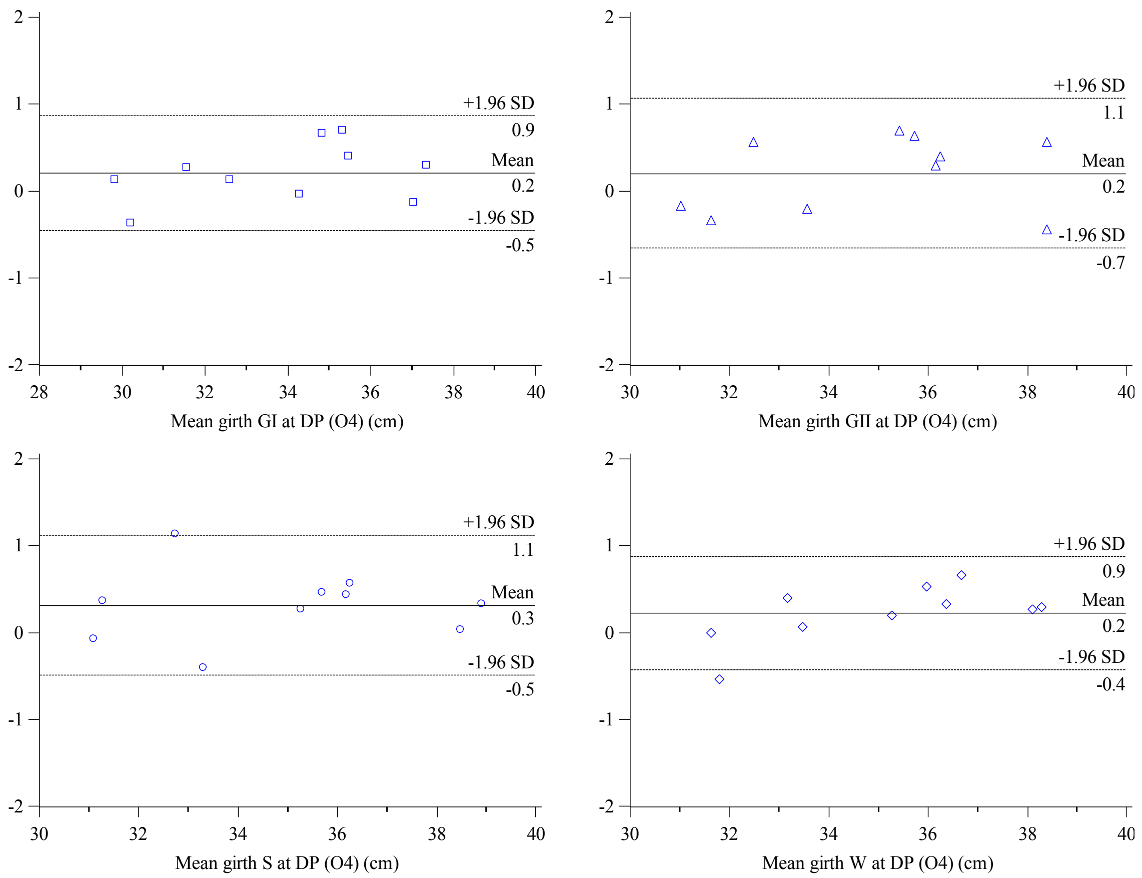


Figure 53: Girth - Intra-observer B-A plots at DP for observer O4 (n=10 legs)

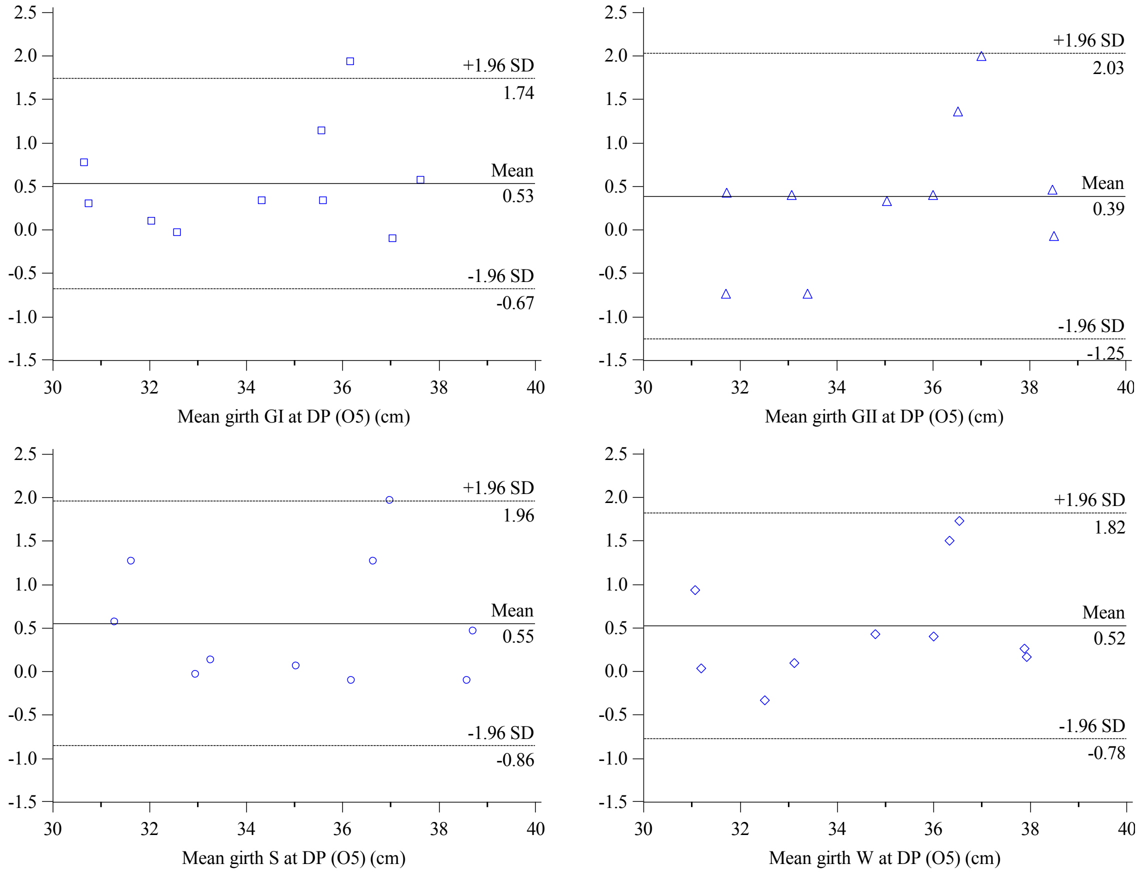


Figure 54: Girth - Intra-observer B-A plots at DP for observer O5 (n=10 legs)

A.2.2 Reproducibility of goniometric measurements (O1-O5)

A.2.2.1 Inter-observer reliability (O1-O5)

Figure 55 shows the inter-observer intraclass correlation coefficients and 95% confidence intervals for the observers O1, O2, O3, O4 and O5 on the first and second measuring day. On the first measuring day, the ICC ranged from 0.68 to 0.88, while it ranged from 0.60 to 0.82 on the second measuring day. Considering the measuring position, the ICC for test position P1 was lower than for position P2.

Position	Reliability: ICC [95% CI]	
	t1	t2
P1	0.68 [0.35 , 0.90]	0.60 [0.27 , 0.86]
P2	0.88 [0.59 , 0.97]	0.82 [0.51 , 0.95]

Table 40: Flexion - Inter-observer reliability (ICC) for observers O1-O5 (n=5)
ICC, intraclass correlation coefficient; CI, confidence interval; P1, first knee joint position; P2, second knee joint position;

A.2.2.2 Intra-observer reproducibility (O1-O5)

Table 41 and Table 42summarize the results of the intra-observer agreement and reliability for the observers O1, O2, O3, O4 and O5 and five subjects, respectively. The corresponding Bland and Altman plots are shown in Figure 55.

Position	Observer	t1 (°)	t2 (°)	t1-t2 (°)		Lower limit (°)	Upper limit (°)
		Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
P1	O1	128.0 ± 4.7	130.7 ±5.3	-2.7 [-4.2 to -1.2]	2.1	-6.9 [-9.6 to -4.2]	1.4 [-1.3 to 4.1]
	O2	125.2 ±3.8	129.1 ±1.9	-3.8 [-5.5 to -2.2]	2.3	-8.3 [-11.1 to -5.4]	0.6 [-2.3 to 3.4]
	O3	125.4 ±5.0	126.1 ±5.3	-0.7 [-2.2 to 0.8]	2.2	-4.9 [-7.6 to -2.2]	3.5 [0.8 to 6.2]
	O4	124.6 ±5.8	126.2 ±3.7	-1.6 [-3.9 to 0.7]	3.2	-7.8 [-11.9 to -3.8]	4.7 [0.6 to 8.7]
	O5	121.3 ±5.1	124.6 ±4.9	-3.2 [-5.4 to -1.1]	3.0	-9.1 [-12.9 to -5.3]	2.7 [-1.1 to 6.5]
P2	O1	85.5 ±11.0	89.4 ±12.5	-3.9 [-5.5 to -2.4]	2.2	-8.2 [-11.0 to -5.4]	0.4 [-2.4 to 3.2]
	O2	87.3 ±8.3	88.1 ±7.5	-0.9 [-3.2 to 1.4]	3.2	-7.2 [-11.2 to -3.1]	5.4 [1.3 to 9.4]
	O3	81.0 ±11.6	81.9 ±12.2	-0.9 [-3.9 to 2.1]	4.2	-9.2 [-14.5 to -3.9]	7.3 [2.0 to 12.7]
	O4	79.0 ±12.3	80.4 ±11.0	-1.5 [-3.4 to 0.5]	2.7	-6.7 [-10.1 to -3.3]	3.8 [0.4 to 7.2]
	O5	81.6 ±10.2	82.8 ±9.9	-1.1 [-2.6 to 0.3]	2.1	-5.2 [-7.8 to -2.6]	2.9 [0.3 to 5.5]

Table 41: Flexion - Intra-observer agreement for observers O1–O5
with mean girth ±SD measured on first and second measuring day, (n=5);

Observer	Reliability: ICC [95% CI]	
	P1	P2
O1	0.80 [0.00 to 0.96]	0.93 [0.06 to 0.99]
O2	0.40 [-0.11 to 0.81]	0.92 [0.73 to 0.98]
O3	0.91 [0.70 to 0.98]	0.94 [0.79 to 0.98]
O4	0.76 [0.31 to 0.93]	0.97 [0.87 to 0.99]
O5	0.69 [-0.01 to 0.92]	0.97 [0.90 to 0.99]

Table 42: Flexion - Intra-observer reliability (ICC) for observers O1-O5 (n=5)
ICC, intraclass correlation coefficient; CI, confidence interval; P1, first knee joint position; P2, second knee joint position;

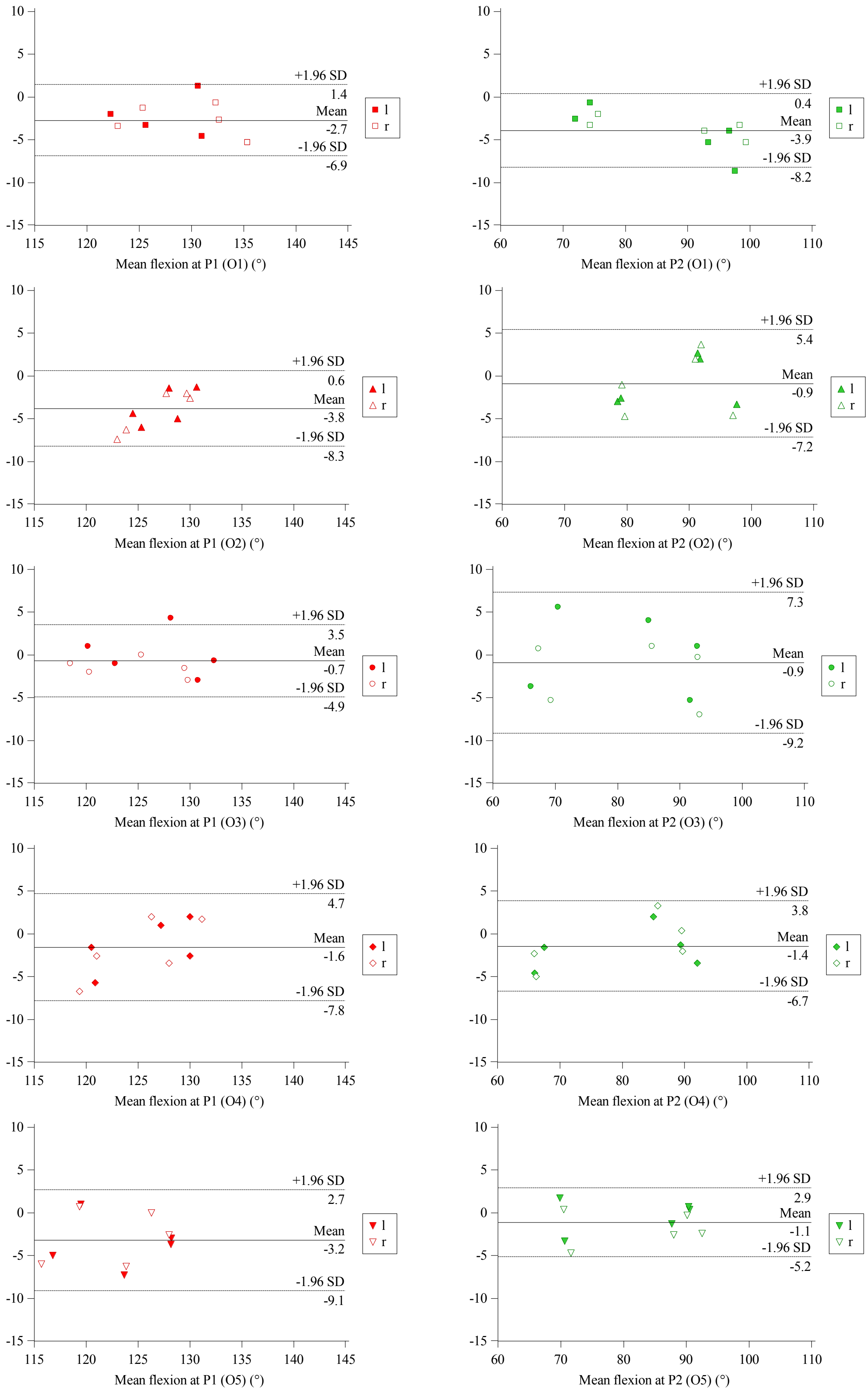


Figure 55: Flexion - Intra-observer B-A plots for the observers O1-O5

A.3 Additional tables and figures of observers O1 and O2

A.3.1 Reproducibility of girth measurements (first measuring day)

A.1.1.1 Inter-observer reproducibility

Table 43 shows in detail the results of the Bland and Altman analysis with confidence intervals for the mean difference mD and the lower and upper limits of agreement.

Site	Tape	O1 (cm)	O2 (cm)	Agreement: O1-O2 (cm)			
		Mean ±SD	Mean ±SD	mD [95% CI]	SD_{diff}	Lower limit [95% CI]	Upper limit [95% CI]
PP	GI	39.1 ± 2.4	38.9 ± 2.5	0.2 [-0.0 to 0.3]	0.5	-0.9 [-1.2 to -0.6]	1.2 [0.9 to 1.5]
	GII	40.9 ± 2.5	40.5 ± 2.5	0.4 [0.2 to 0.7]	0.8	-1.1 [-1.7 to -0.8]	2.1 [1.6 to 2.5]
	S	40.7 ± 2.5	40.4 ± 2.55	0.3 [0.1 to 0.6]	0.7	-1.0 [-1.4 to -0.6]	1.7 [1.3 to 2.1]
	W	39.1 ± 2.4	39.1 ± 2.4	0.0 [-0.1 to 0.2]	0.4	-0.8 [-1.1 to -0.6]	0.9 [0.6 to 1.1]
MP	GI	36.7 ± 2.1	36.7 ± 2.0	0.05 [-0.1to 0.2]	0.5	-0.9 [-1.2 to -0.7]	1.0 [0.7 to 1.3]
	GII	37.8 ± 2.2	37.8 ± 2.1	-0.1 [-0.3 to 0.1]	0.6	-1.2 [-1.5 to -0.9]	1.0 [0.7 to 1.3]
	S	37.8 ± 2.2	37.7 ± 2.1	0.2 [0.05 to 0.3]	0.5	-0.7 [-0.9 to -0.4]	1.1 [0.8 to 1.3]
	W	37.1 ± 2.1	37.1 ± 2.0	0.0 [-0.1 to 0.2]	0.4	-0.8 [-1.0 to -0.6]	0.9 [0.6 to 1.1]
DP	GI	33.4 ± 2.1	33.6 ± 2.1	-0.2 [-0.3 to -0.1]	0.4	-0.9 [-1.1 to -0.7]	0.5 [0.3 to 0.7]
	GII	34.3 ± 2.1	34.5 ± 2.0	-0.2 [-0.4 to -0.1]	0.4	-1.0 [-1.2 to -0.8]	0.6 [0.3 to 0.8]
	S	34.5 ± 2.1	34.5 ± 2.0	0.0 [-0.1 to 0.1]	0.4	-0.7 [-1.0 to -0.5]	0.7 [0.5 to 1.0]
	W	34.1 ± 2.0	34.3 ± 2.0	-0.2 [-0.3 to 0.0]	0.4	-1.0 [-1.2 to -0.8]	0.6 [0.4 to 0.9]

Table 43: Girth - Inter-observer agreement for first measuring day (O1 and O2)

A.1.1.2 Intra-observer reproducibility

The detailed results of inter-observer agreement analysis for the observers O1 and O2 are presented in Table 44 and Table 45, respectively. The corresponding Bland and Altman plots for the measurement sites at mid-patella and 7 cm distal of mid-patella, which were not shown in the results chapter, are displayed in Figure 56 and Figure 57, respectively.

Site	Tape	t1	t2	Agreement O1: t1-t2 (cm)			
		Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	Lower Limit [95% CI]	Upper limit [95% CI]
PP	GI	39.1 ± 2.4	38.8 ± 2.3	0.1 [-0.2 to 0.3]	0.6	-1.1 [-1.5 to -0.7]	1.3 [0.9 to 1.7]
	GII	40.9 ± 2.5	40.7 ± 2.4	0.1[-0.2 to 0.4]	0.8	-1.4 [-1.9 to -1.0]	1.7 [1.2 to 2.2]
	S	40.7 ± 2.5	40.3 ± 2.5	0.2 [0.0 to 0.3]	0.5	-0.9 [-1.2 to -0.6]	1.2 [0.9 to 1.5]
	W	39.1 ± 2.4	38.8 ± 2.4	0.1 [-0.1 to 0.3]	0.5	-0.9 [-1.2 to -0.6]	1.1 [0.8 to 1.4]
MP	GI	36.7 ± 2.1	36.6 ± 2.2	0.1 [0.0 to 0.3]	0.4	-0.7 [-0.9 to -0.4]	0.9 [0.7 to 1.2]
	GII	37.8 ± 2.2	37.6 ± 2.0	0.0 [-0.1 to 0.2]	0.5	-0.9 [-1.2 to -0.6]	1.0 [0.7 to 1.2]
	S	37.8 ± 2.1	37.7 ± 2.0	0.0 [-0.2 to 0.2]	0.5	-1.1 [-1.4 to -0.8]	1.0 [0.7 to 1.3]
	W	37.1 ± 2.1	37.0 ± 2.0	0.0 [-0.2 to 0.2]	0.5	-1.0 [-1.3 to -0.7]	1.0 [0.7 to 1.3]
DP	GI	33.4 ± 2.1	33.4 ± 2.3	0.1 [0.0 to 0.3]	0.5	-0.8 [-1.2 to -0.5]	1.1[0.8 to 1.4]
	GII	34.3 ± 2.1	34.2 ± 2.2	0.1 [-0.1 to 0.3]	0.5	-0.9 [-1.2 to -0.6]	1.2 [0.9 to 1.5]
	S	34.5 ± 2.1	34.5 ± 2.1	0.0 [-0.1 to 0.2]	0.5	-0.9 [-1.2 to -0.6]	1.0 [0.7 to 1.3]
	W	34.1 ± 2.0	34.0 ± 2.1	0.1 [-0.1 to 0.3]	0.5	-0.8 [-1.1 to -0.6]	1.1 [0.8 to 1.4]

Table 44: Girth - Intra-observer agreement for observer O1

Site	Tape	t1	t2	Agreement O2: t1-t2			
		Mean ± SD (cm)	Mean ± SD (cm)	mD [95% CI] (cm)	SD_{diff}	Lower limit [95% CI] (cm)	Upper limit [95% CI] (cm)
PP	GI	38.9 ± 2.5	38.6 ± 2.5	0.1 [-0.2 to 0.3]	0.6	-1.2 [-1.6 to -0.8]	1.3 [0.9 to 1.7]
	GII	40.5 ± 2.5	40.5 ± 2.5	-0.1 [-0.4 to 0.2]	0.8	-1.6 [-2.0 to -1.2]	1.4 [0.9 to 1.8]
	S	40.4 ± 2.6	40.2 ± 2.6	0.0 [-0.2 to 0.2]	0.6	-1.1 [-1.5 to -0.8]	1.1 [0.8 to 1.5]
	W	39.1 ± 2.4	38.8 ± 2.6	0.1 [0.0 to 0.3]	0.6	-0.9 [-1.3 to -0.6]	1.2 [0.9 to 1.6]
MP	GI	36.7 ± 2.0	36.6 ± 2.2	0.0 [-0.1 to 0.2]	0.4	-0.8 [-1.1 to -0.5]	0.9 [0.6 to 1.2]
	GII	37.8 ± 2.1	37.7 ± 2.1	0.0 [-0.2 to 0.2]	0.5	-1.1 [-1.4 to -0.8]	1.0 [0.7 to 1.3]
	S	37.7 ± 2.1	37.5 ± 2.1	0.0 [-0.1 to 0.2]	0.4	-0.9 [-1.1 to -0.6]	0.9 [0.6 to 1.2]
	W	37.1 ± 2.0	36.9 ± 2.1	0.1 [-0.1 to 0.2]	0.4	-0.8 [-1.0 to -0.5]	0.9 [0.7 to 1.2]
DP	GI	33.6 ± 2.1	33.5 ± 2.4	0.2 [0.1 to 0.4]	0.4	-0.6 [-0.8 to -0.3]	1.0 [0.8 to 1.3]
	GII	34.5 ± 2.0	34.3 ± 2.3	0.2 [0.1 to 0.4]	0.5	-0.8 [-1.1 to -0.5]	1.3 [1.0 to 1.6]
	S	34.5 ± 2.0	34.4 ± 2.2	0.2 [0.0 to 0.3]	0.4	-0.7 [-1.0 to -0.5]	1.0 [0.8 to 1.3]
	W	34.3 ± 2.0	34.2 ± 2.1	0.1 [0.0 to 0.3]	0.5	-0.8 [-1.1 to -0.5]	1.1 [0.8 to 1.4]

Table 45: Girth - Intra-observer agreement for observer O2

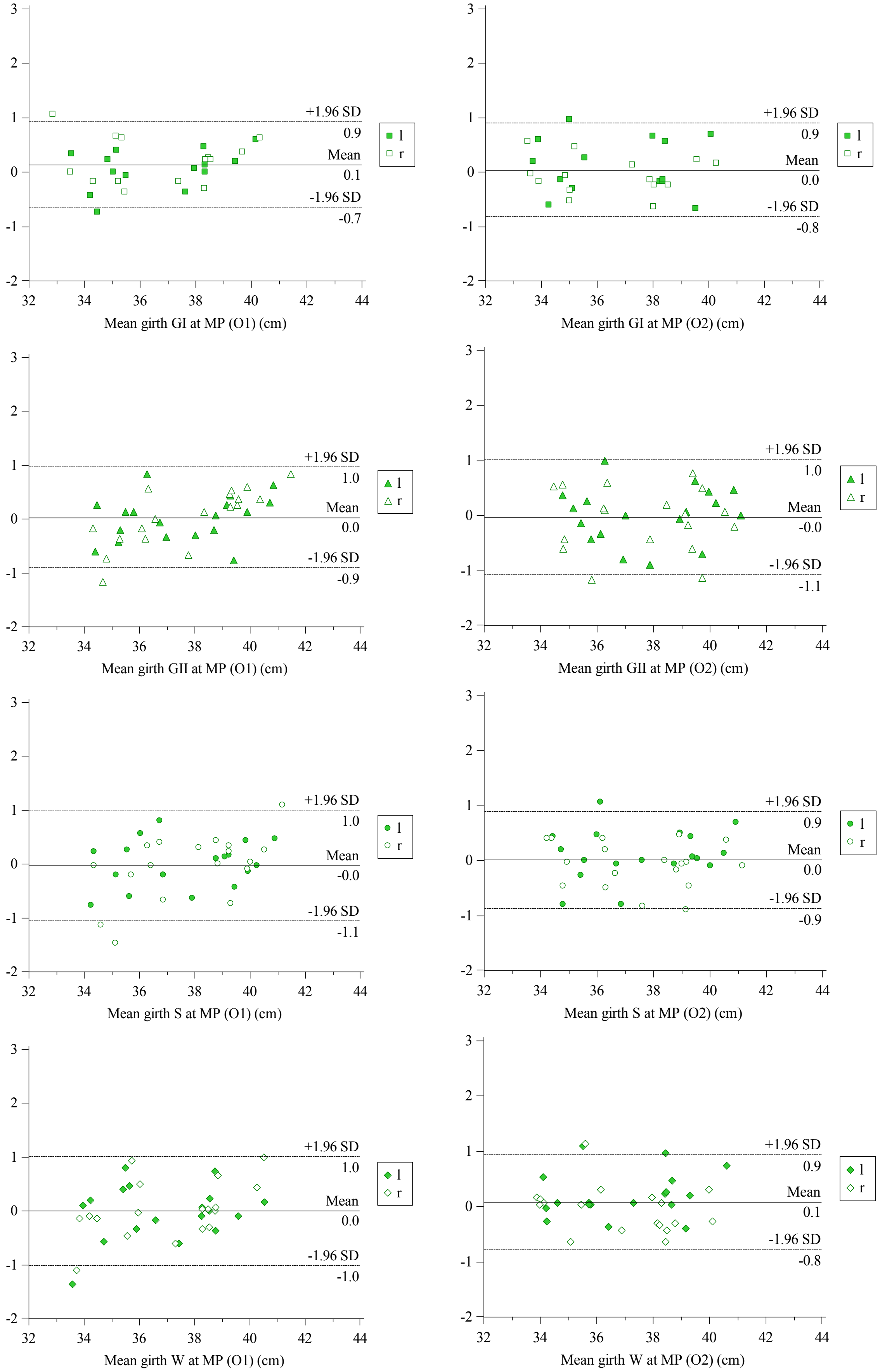


Figure 56: Girth - Intra-observer B-A plots at MP for the observers O1 and O2

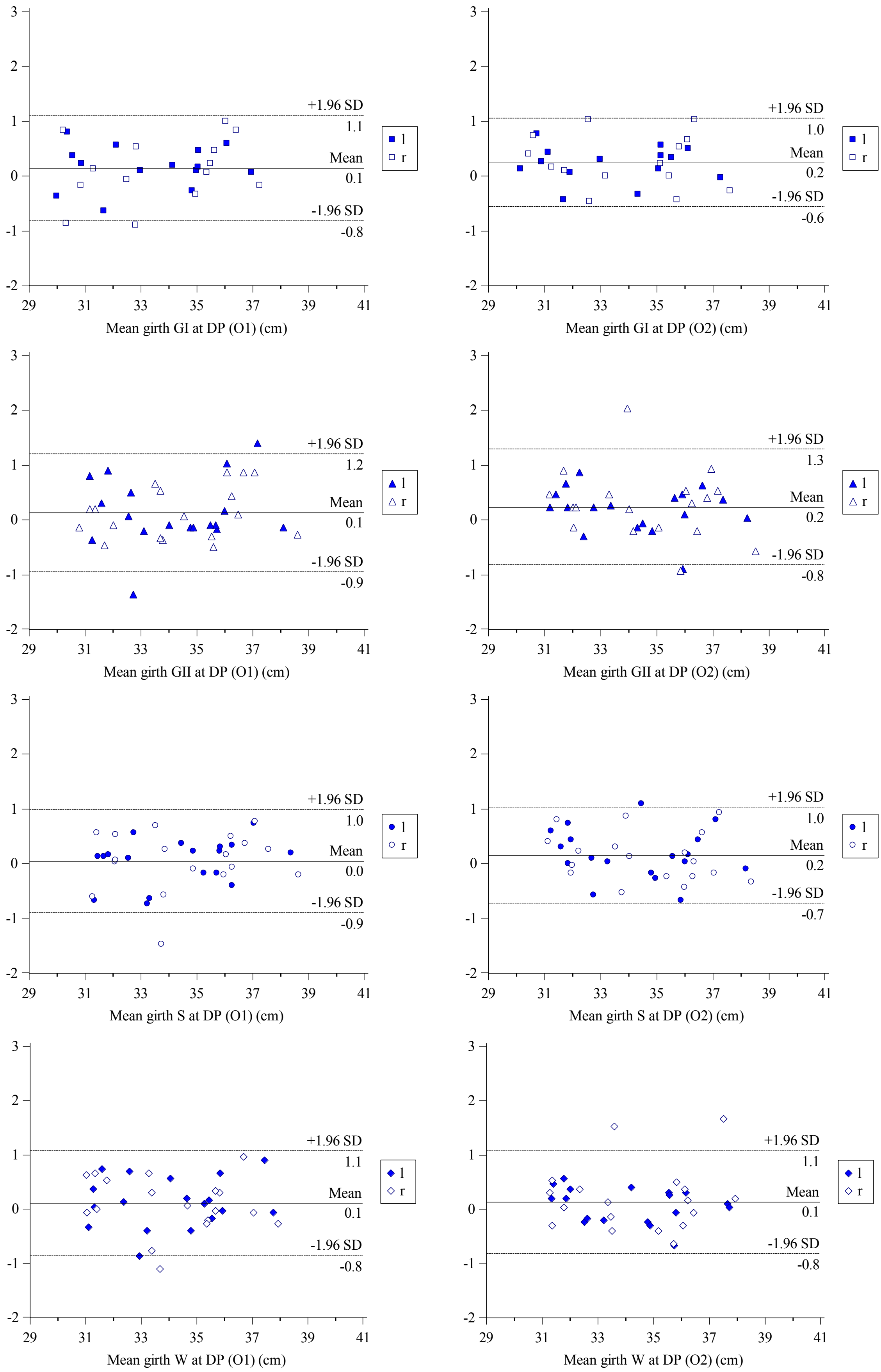


Figure 57: Girth - Intra-observer B-A plots at DP for the observers O1 and O2

A.3.2 Reproducibility of knee flexion measurements

A.3.2.1 Inter-observer reliability (first measuring day)

Position	O1 (°)	O2 (°)	O1-O2 (°)		Lower limit (°)	Upper limit (°)
	Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	[95% CI]	[95% CI]
P1	112.0 ± 18.1	111.1 ± 17.1	1.0 [0.1 to 1.8]	2.5	-4.0 [-5.4 to -2.5	5.9 [4.5 to 7.3
P2	82.4 ± 8.3	85.1 ± 8.8	-1.1 [-2.3 to 0.1]	3.6	-8.2 [-10.3 to -6.2]	6.0 [3.9 to 8.0]

Table 46: Flexion - Inter-observer reproducibility for O1 and O2 (t1, n=38 legs)

A.1.1.1 Intra-observer reproducibility

Position	t1 (°)	t2 (°)	Agreement: t1-t2 (°)			
	Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	Lower limit [95% CI]	Upper limit [95% CI]
P1	112.0 ± 18.1	113.8 ± 18.0	-1.8 [-2.8 to -0.8]	2.7	-7.1 [-8.8 to -5.5]	3.6 [1.9 to 5.2]
P2	82.4 ± 8.3	85.1 ± 8.8	-3.2 [-4.0 to -2.4]	2.3	-7.7 [-9.1 to -6.3]	1.2 [-0.1 to 2.6]

Table 47: Flexion - Intra-observer reproducibility for observer O1 (n=34 legs)

Position	t1 (°)	t2 (°)	Agreement: t1-t2 (°)			
	Mean ± SD	Mean ± SD	mD [95% CI]	SD_{diff}	Lower limit [95% CI]	Upper limit [95% CI]
P1	111.1 ± 17.1	112.8 ± 17.3	-1.9 [-3.0 to -0.8]	3.2	-8.1 [-10.0 to -6.2]	4.3 [2.4 to 6.2]
P2	83.6 ± 6.58	84.0 ± 6.3	-0.8 [-1.9 to 0.4]	3.3	-7.1 [-9.1 to -5.2]	5.6 [3.6 to 7.6]

Table 48: Flexion - Intra-observer reproducibility for observer O2 (n=34 legs)

A.4 Additional figures of clinical course measurements

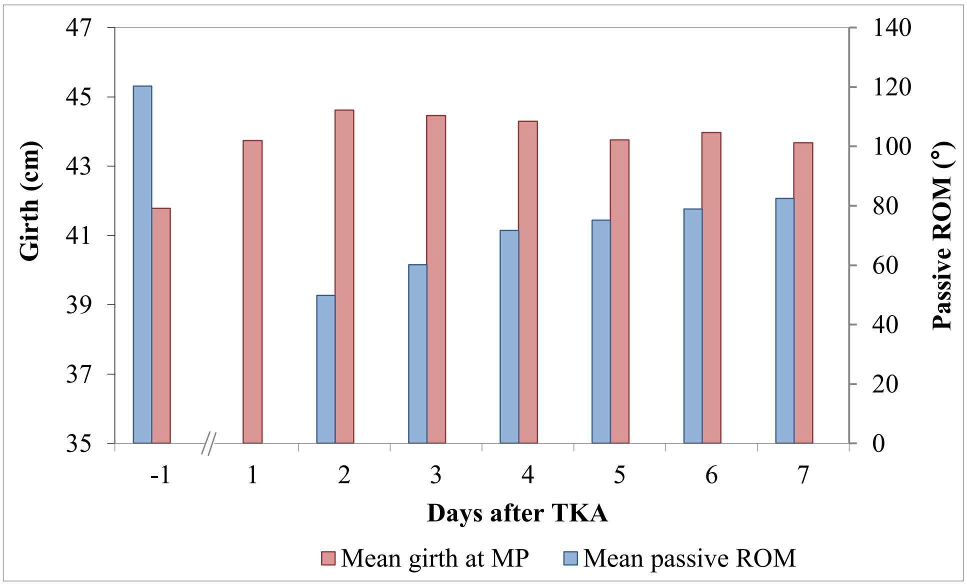


Figure 58: Relationship between mean girth at MP and mean passive ROM

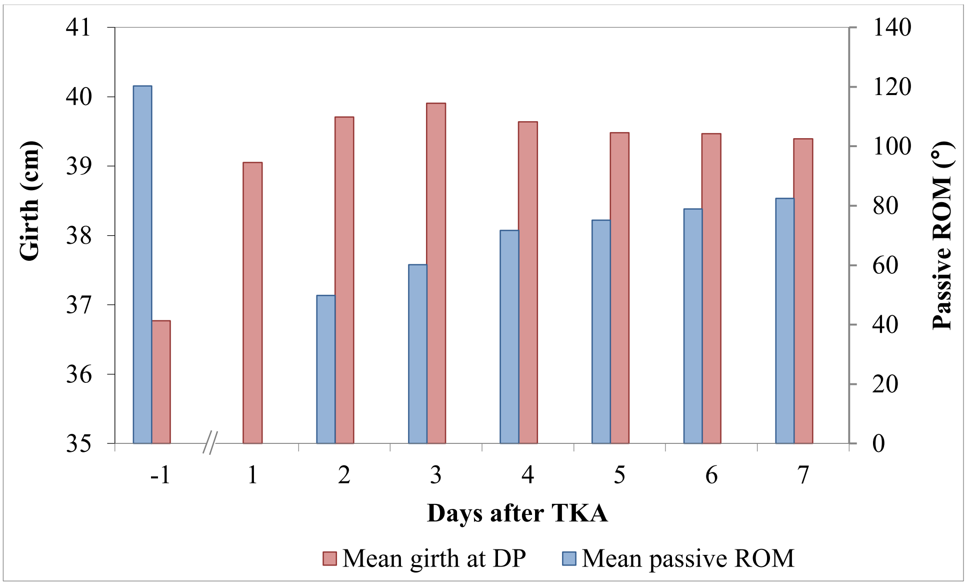


Figure 59: Relationship between mean girth at DP and mean passive ROM

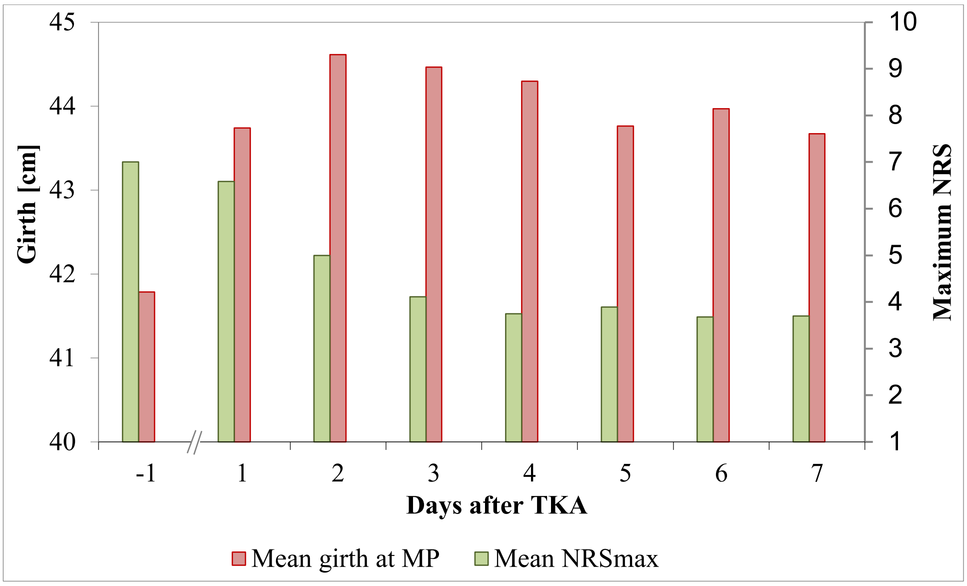


Figure 60: Relationship between mean Girth at MP and maximum reported NRS

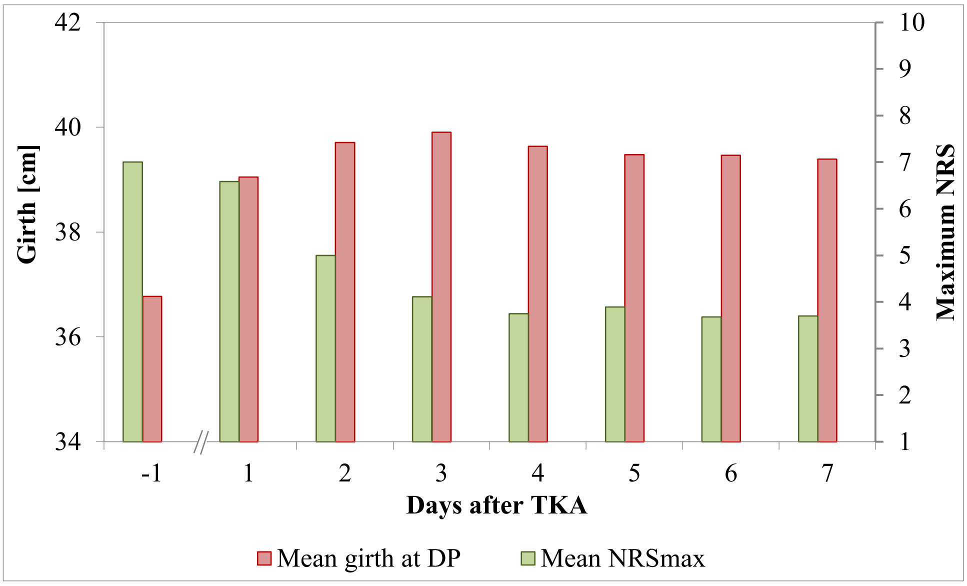


Figure 61: Relationship between mean Girth at DP and maximum reported NRS

Questionnaire on the usability of the measuring tapes used in this study (Gulick I, Gulick II plus, standard, and Waegener tape measures) (translated from German)

Please take some time to answer the following questions on the measuring tapes used!

Observer number:

1. What is your gender?

☐ male ☐ female

2. Did you have any experience in circumferential leg measurements prior to this study?

☐ yes ☐ no

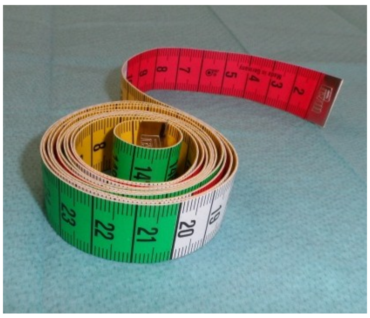
3. Please list the measuring tapes used in this study according to their usability!

Note: 1. Most user-friendly
4. Least user-friendly

1.	
2.	
3.	
4.	

Gulick I

Gulick II plus



Standard

Waegener

Please mark with a cross where applicable!

4. What did you like most about the tape measure you ranked first?

	Does apply	Does not apply
Easy handling	☐	☐
Precise measurement	☐	☐
Measurements taken quickly	☐	☐
Thick tape	☐	☐
Thin tape	☐	☐
Numbers easy to read	☐	☐
Zero line clearly visible	☐	☐
Tape easy to position	☐	☐
Tape doesn´t slip	☐	☐

Space for comments:

5. What did you dislike about the measuring tape you ranked lowest?

	Does apply	Does not apply
Complicated handling	☐	☐
Inaccurate measurement	☐	☐
Measuring takes long	☐	☐
Thick tape	☐	☐
Thin tape	☐	☐
Numbers difficult to read	☐	☐
Zero line difficult to identify	☐	☐
Difficult positioning of tape	☐	☐
Tape slips easily	☐	☐

Space for comments:

6. Which is the most accurate tape measure in your opinion?

☐ Gulick I ☐ Gulick II plus ☐ Standard ☐ Waegener

7. Which do you think is the least accurate tape measure?

☐ Gulick I ☐ Gulick II plus ☐ Standard ☐ Waegener

8. With which tape measure can the measurements be taken fastest?

☐ Gulick I ☐ Gulick II plus ☐ Standard ☐ Waegener

9. Which tape measure takes the longest?

☐ Gulick I ☐ Gulick II plus ☐ Standard ☐ Waegener

10. Do you feel that your measuring accuracy decreased with the duration of the measurement procedure (because you got tired)?

☐ Yes ☐ No

11. Do you think that your measuring accuracy increased with the duration of the measurement procedure (because you got used to handling the measuring tapes)?

☐ Yes ☐ No

12. Space for personal remarks

	Pros	Cons
Standard tape measure		
Waegener tape measure		
Gulick I tape measure		
Gulick II plus tape measure		

Thank you for your time!

Probandeninformation/Einwilligungserklärung

zur Teilnahme an der Studie

Umfangs- und Winkelmessungen im Kniebereich

Sehr geehrte Teilnehmerin, sehr geehrter Teilnehmer!

Ich lade Sie ein an der oben genannten Studie im Rahmen meiner Diplomarbeit teilzunehmen.
Ihre Teilnahme an dieser Studie erfolgt freiwillig. Sie können jederzeit ohne Angabe von Gründen aus der Studie ausscheiden.
Studien sind notwendig, um verlässliche neue medizinische Forschungsergebnisse zu gewinnen. Unverzichtbare Voraussetzung für die Durchführung einer Studie ist jedoch, dass Sie Ihr Einverständnis zur Teilnahme an dieser Studie schriftlich erklären.
Bitte unterschreiben Sie die Einwilligungserklärung nur

- wenn Sie Art und Ablauf der Studie vollständig verstanden haben,
- wenn Sie bereit sind, der Teilnahme zuzustimmen und
- wenn Sie sich über Ihre Rechte als Teilnehmer an dieser Studie im Klaren sind.

1. Was ist der Zweck der Studie?

Der Zweck dieser Studie ist es, eine Messmethode zur Bestimmung des Beinumfanges im Kniebereich zu evaluieren, sowie die Messmethode zur Bestimmung der Kniegelenksbeugung mittels Goniometer auf ihre Genauigkeit zu überprüfen.

2. Wie läuft die Studie ab?

Der Beinumfang jedes Beins wird an drei verschiedenen Messpunkten im Bereich des Kniegelenks mit vier verschiedenen Maßbändern gemessen. Für die Umfangsmessung wird das jeweilige Bein auf einer Papierrolle gelagert. Als Referenz für die Messpunkte wird die Kniescheibe herangezogen. Die Position der Kniescheibe wird zunächst durch Ertasten ermittelt und die obere und untere Begrenzung auf einem zuvor geklebten Pflaster markiert. Dann wird die Mitte der Kniescheibe ermittelt. Dies ist der erste Messpunkt. Die beiden weiteren Messpunkte liegen 7 cm über bzw. 7 cm unter dem ersten Messpunkt. Da die drei Messpunkte von jedem Prüfer neu auf einem Pflaster markiert werden und daher das Pflaster nach den Messungen jedes Prüfers entfernt wird, wird den Teilnehmern vor Beginn der Messungen ein Strumpf angezogen, sodass das Pflaster nicht direkt auf die

Haut geklebt werden muss. Um die Genauigkeit der Methode zu überprüfen, wird an jeder Messstelle drei Mal gemessen. Die Beugung im Kniegelenk wird mit einem Goniometer in drei verschiedenen Beugepositionen bestimmt. Dazu wird das jeweilige Bein in einer speziellen Vorrichtung gelagert. Auch hier wird die Messung drei Mal wiederholt, um die Genauigkeit der Methode zu überprüfen.

3. Gibt es Risiken, Beschwerden und Begleiterscheinungen?

Die Messungen sind für die/den (kniegesunden) TeilnehmerIn schmerzfrei.

4. Hat die Teilnahme an der Studie sonstige Auswirkungen auf die Lebensführung und welche Verpflichtungen ergeben sich daraus?

Keine

5. In welcher Weise werden die im Rahmen dieser Studie gesammelten Daten verwendet?

Nur die Prüfer und deren Mitarbeiter haben Zugang zu den vertraulichen Daten, in denen Sie namentlich genannt werden. Diese Personen unterliegen der Schweigepflicht.
Die Auswertung und- wenn dafür nötig- die Weitergabe der erhobenen Daten erfolgt ausschließlich zu in anonymisierter Form. Auch zur etwaigen Veröffentlichung der Ergebnisse dieser Studie werden nur anonymisierte Daten verwendet und Sie nicht namentlich erwähnt.

6. Möglichkeit zur Diskussion weiterer Fragen

Für weitere Fragen im Zusammenhang mit dieser Studie stehen Ihnen Ihre Prüfer gern zur Verfügung. Auch Fragen, die Ihre Rechte als Teilnehmer an dieser Studie betreffen, werden Ihnen gerne beantwortet.

Namen der Kontaktpersonen:

Daniela Hirzberger
Universitätsklinik für Orthopädie, Graz
Erreichbar unter: 0676 67 20 464

Ass. Prof. Dr. Mathias Glehr
Universitätsklinik für Orthopädie, Graz
Erreichbar unter: 0316 385-81756

7. Einwilligungserklärung

Name des Teilnehmers in Druckbuchstaben:

....

Geb.Datum: .... Code: ....

Ich erkläre mich bereit, an der Studie „Umfangs- und Bewegungsmessungen“ teilzunehmen.

Ich bin von Herrn Ass.-Prof. Dr. Mathias Glehr ausführlich und verständlich über mögliche Belastungen und Risiken, sowie über Wesen, Bedeutung und Tragweite der Studie und sich für mich daraus ergebenden Anforderungen aufgeklärt worden. Ich habe darüber hinaus den Text dieser Patientenaufklärung und Einwilligungserklärung, die insgesamt 3 Seiten umfasst, gelesen. Aufgetretene Fragen wurden mir vom Prüfer verständlich und genügend beantwortet. Ich hatte ausreichend Zeit, mich zu entscheiden. Ich habe zurzeit keine weiteren Fragen mehr.

Ich werde den Anordnungen, die für die Durchführung der Studie erforderlich sind, Folge leisten, behalte mir jedoch das Recht vor, meine freiwillige Mitwirkung jederzeit zu beenden.

Ich bin zugleich damit einverstanden, dass meine im Rahmen dieser Studie ermittelten Daten aufgezeichnet werden. Um die Richtigkeit der Datenaufzeichnung zu überprüfen, dürfen Beauftragte des Auftraggebers und der zuständigen Behörden (z.B. Medizinische Universität Graz) beim Prüfer Einblick in meine personenbezogenen Daten nehmen. Beim Umgang mit den Daten werden die Bestimmungen des Datenschutzgesetzes beachtet. Eine Kopie dieser Patienteninformation und Einwilligungserklärung können Sie auf Wunsch erhalten. Das Original verbleibt beim Prüfer.

....

(Datum und Unterschrift der Teilnehmerin/des Teilnehmers)

....

(Datum, Name und Unterschrift des verantwortlichen Prüfers)