Inter-rater and intra-rater reliability of the fluid goniometer for measuring active knee flexion in painful knees; correlations do not mean agreement

WILTON REMIGIO, MS, MPT, DSc1)*, NANCY TSAI, PT2), LEONIVIC LAYOS, MPT3), MICHELLE CHAVEZ, MPT4)

1) College of Health Professions and Education, Misericordia University: 301 Lake Street, Dallas, PA 18612, USA
2) ATP Specific Training, USA
3) Cedars-Sinai Medical Center, USA
4) Long Beach Memorial Medical Center, USA

Abstract. [Purpose] The fluid goniometer is an instrument for measuring range of motion. Reliability of the fluid goniometer has not been established for subjects with painful knee joints. The purpose of this study was to determine the inter-rater and intra-rater reliability of the fluid goniometer in measuring active knee flexion of painful knees and to test its agreement with the gold standard ruler goniometer. [Subjects and Methods] Twenty-five individuals with either unilateral or bilateral painful knees participated in the study. Two raters each took three measurements with the same Baseline® fluid goniometer on 35 knees. [Results] Intraclass correlation coefficients (ICC) were 0.97 for both intra-rater and inter-rater measurements, denoting high relative reliability. The large standard error of measurement (SEM) value of 6.6 degrees, and the 95% limits of agreement, which revealed a potential difference of 18.4 degrees between raters of similar subjects, however, revealed poor absolute reliability. The smallest detectable difference (SDD) of 18 degrees was also large. [Conclusion] The results revealed excellent relative reliability, but a large amount of variability between the raters’ measurements. The sensitivity of the fluid level of the goniometer to end range tremors of the lower leg flexed against gravity in the obligatory prone position may contribute significantly to the large variability in knee ROM values. Key words: Altman and Bland limits of agreement plot, Fluid goniometer, Inter-rater reliability

INTRODUCTION

Clinicians use goniometry extensively as a portable, inexpensive and practical way to assess joint motion. The universal goniometer (UG) is the most widely accepted standard instrument used in clinics for quantifying joint range of motion. It is much more precise than to visual estimates1,3. Measuring joint range of motion is a fundamental and objective part of assessing joint mobility and is an important component of functional outcome scores. Up to twenty five percent of disability scores are attributed to the range of motion deficit in knee and hip flexion3).

The widely accepted use of the UG as a standard range of motion measuring tool is supported by past studies confirming high intra-rater reliability for the instrument. According to Currier4), a correlation coefficient of 0.8 to 1.00 represents a high degree of reliability. In a study done in 1997, Brosseau et al.5) found high intra-rater reliability for the UG of 0.86–0.94 for

*Corresponding author. Wilton Remigio (E-mail: wremigio@misericordia.edu)
©2017 The Society of Physical Therapy Science. Published by IPEC Inc.
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial No Derivatives (by-nc-nd) License <https://creativecommons.org/licenses/by-nc-nd/4.0/>.
small angle positions and 0.95–0.97 for large-angle positions in healthy knees. A few years later, they found high intra-reliability value of 0.997 for knee flexion in subjects with knee impairments\(^6\). Rothstein et al.\(^7\) found high intra-rater reliability for pain-free elbow and knee range of motion (r=0.91 and 0.99, respectively).

The UG is associated with moderate to high inter-rater reliability values. In 1997, Brosseau\(^5\) found inter-rater reliability of 0.62–0.71 for small knee angle positions and 0.91–1.00 for large knee angle positions. In a different study, they found very high inter-rater reliability ranging from 0.977 to 0.982 in subjects with knee impairments\(^6\). Rothstein et al.\(^7\) studied the inter-rater reliability for elbow and knee flexion and the Pearson’s product-moment correlation was 0.88–0.977. Rheault et al.\(^8\) measured knee flexion in pain-free knees and reported the inter-rater reliability of the universal goniometer was r=0.87.

There are many devices developed that are easy to operate and that make the measurement process as accurate as possible. The more sophisticated electro-goniometer, however, is too expensive for small privately owned facilities to afford. The fluid goniometer (FG) is an instrument introduced by Buck et al.\(^9\) in 1959. Similar to the UG, it is compact, lightweight, but even easier to operate with one hand. We found one study by Rheault et al.\(^8\) who measured knee flexion in pain-free knees and reported inter-rater reliability coefficients of r=0.87 and r=0.83 for the universal and fluid-based goniometer respectively.

Despite the fact that therapists routinely treat and measure painful joints, most reliability studies are conducted on normal joints. Brosseau et al.\(^5\) performed one of the few studies done on knees with impairments. We did not find any study on measuring painful knee joints with the FG. The purpose of this study was to determine the inter- and intra-rater reliability of measuring active knee flexion in patients with pathological knee by using the fluid goniometer. Analyses of measurement error are reported to complement information provided by reliability coefficients, as recommended in more recent studies\(^10, 11\).

**SUBJECTS AND METHODS**

A total of 25 subjects were recruited by flyers and physical therapists at four clinics, three of which were outpatient and one transitional care facility. Subjects signed an informed consent form to take part on the study. The study was approved by the Loma Linda University Institution Research board, protocol No. HS 53089. All subjects 1) were 21 years of age or older, 2) had at least one painful knee, 3) were receiving physical therapy at the clinics, but not necessarily for the knee, 4) spoke and understood English fluently, and 5) were able to lie prone for at least ten minutes without pillow support under the abdomen.

The fluid-based goniometer is a fluid-filled chamber with a 360-degree scale divided into 1-degree increments. The flat base is placed on the back of the calf. Measurement is read at the point where the bottom of the meniscus aligns with the degree scale.

None of three individuals, who did the FG measurement, had previous experience with using a fluid goniometer. The raters followed the instructions on using the fluid goniometer posted on the MIE Medical Research Ltd. website\(^12\). Both raters practiced with the fluid goniometer for two days in the clinics where they worked prior to the first day of the data collection. The reader practiced reading the measurements from the goniometer on one of the days the raters practiced in their clinics. One rater was a licensed physical therapist with three years of clinical experience. The second rater and the reader were foreign trained physical therapists, who each had worked as physical therapy aides for three years.

Flyers regarding the study were posted two to three weeks prior to the commencement of data collection. Volunteers that met the study criteria were scheduled for the measurements to be taken after they completed their routine physical therapy treatment. The researchers measured the subjects’ knees at their respective clinics, where they received routine physical therapy treatment.

At each clinic, subjects were shown to a private treatment room with a height adjustable table, where the study’s purpose and procedures were explained, and measurements were taken. Subjects were allowed as much time as necessary to satisfy their questions and to sign the informed consent. The same researcher collected information from all the subjects regarding age, which knee was painful and the onset of knee pain, and the specific diagnoses was determined through chart review after the data was collected for the day.

There were three researchers involved with the measuring procedure. Two served as raters to position the subject and the fluid goniometer, and the third served as the reader to read the measurements from the goniometer and record the measurements.

The subject lay down on the hydraulic table with feet off the end of the table. A pen mark was made by the first rater on the lateral aspect of the foot to be measured at the point, where the bottom of the treatment table met the foot and a second mark on the calf ten inches distal from the middle popliteal crease of the painful joint. A 3-inch high towel was placed under the distal femur, above the superior border of the patella, and the center of the goniometer’s flat base was aligned with the mark on the calf. The rater zeroed the goniometer prior to each measurement by rotating the plastic dial until the arrow aligned with the zero mark. The zeroing procedure occurred while the knee was in the maximal knee extension position.

The rater asked the subject to bend the knee as far as he could tolerate the pain. As the subject flexed his knee, the rater maintained a slight pressure with one hand on the goniometer to ensure full skin contact with the base evenly during the entire knee flexion. When the subject reached the maximum knee flexion within tolerable pain, he said “Stop. The subject then was asked to rate his pain on a 10 point pain scale. The reader read and recorded the measurements. Whether the subjects were able or not to achieve full knee extension, the knee flexion was recorded as the value obtained at maximal knee flexion.
Both the rater and reader were at the same eye level with the fluid goniometer. A five second pause was allowed between each of the three measurements for the subject to return to the initial knee position. The same procedure was repeated for the measuring the second knee joint if the subject has two painful knee joints. The first rater erased the pen marks completely without leaving any redness on the skin to give away where the marks were made with a single-use alcohol pad. The second rater entered the room within a minute and repeated the same procedure. When both raters had completed the measurements, the subject received a copy of the informed consent and shown out of the room.

To determine inter-rater and intra-rater reproducibility, interclass correlation coefficients (ICC) were used. We judged a correlation above 0.7 to indicate acceptable reliability. The square root of the residual variance from the ANOVA table reported as part of the ICC procedure was used as the SEM.

Bland and Altman’s plot was used to display the size, direction and range of differences between raters in the same unit as the original measurement. Agreement between the raters was calculated as the mean difference between the two raters. Differences between the raters for individual subjects were plotted against the mean of the measurements obtained by the two raters. The Altman and Bland plot was obtained using a free trial version of MedCalc software Version 7.4.3.0 — © 2004 by Frank Schoonjans. The plot was printed using a print screen keyboard command. All other calculations were performed using SPPS® version 11.0.

The limit of agreement, which estimates total error (bias and random error) with 95% confidence, was obtained as the mean difference between the measurements of the two raters (bias), ± 2 SD (random error) of the differences. To determine a threshold for real change in knee ROM when different raters record measurements over a period of time, the smallest detectable difference (SDD) was used. The SDD was calculated as $1.96 \times \sqrt{E \times \text{SEM}}$.

**RESULTS**

The volunteer sample of 25 patients, with 35 involved knees that were studied, presented many different kinds of painful knee pathologies (Table 1). Sixty-nine percent of the subjects reported having knee pain that was less than or equal to 4 out of 10. The mean age was 62.9 years old (Table 2).

The distribution of mean knee flexion measurements (22 to 140 degrees), as measured by each rater, was bimodal. Further investigation revealed the presence of two distinct groups, namely, subjects who had knee surgery and those who did not.

The ICCs for inter- and intra-rater reliability were high, 0.97 (p<0.0005) for both. The standard error of measurement, which accounts for the typical random variability within repeated measurements, was 6.6°. Absolute rater agreement was also assessed by analyzing differences in measurements between the two raters. The difference between the measurements represents the amount of error or deviance from a perfect agreement. The mean difference was −0.93° indicating that the measurements of one of the raters were slightly higher on the average, than those of the other rater. Bland and Altman’s plot showed no relationship between the error values and the magnitude of the measurements.

**DISCUSSION**

Baumgarter describes reliability as being both relative and absolute. Relative reliability shows the degree to which subjects maintain their position within a sample. Absolute measures of reliability focuses on the amount of error to expect in a person’s score, and it makes it possible to consider the practical significance of the reliability results as well. Relative reliability for the FG was determined by the ICC statistic. Our results showed excellent intra- and inter-rater correlation

<table>
<thead>
<tr>
<th>Table 1. Characteristics of the subjects (n=25)</th>
<th>Table 2. Diagnoses related to individual knees (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>n</td>
</tr>
<tr>
<td>Age (yrs)</td>
<td>Post-surgical knees</td>
</tr>
<tr>
<td>Gender</td>
<td>Meniscal tear</td>
</tr>
<tr>
<td>Male</td>
<td>Total knee replacement</td>
</tr>
<tr>
<td>Female</td>
<td>Knee arthroscopy</td>
</tr>
<tr>
<td>Knee involvement</td>
<td>Non-surgical knees</td>
</tr>
<tr>
<td>Unilateral</td>
<td>Osteoarthritis</td>
</tr>
<tr>
<td>Bilateral</td>
<td>Overuse syndrome</td>
</tr>
<tr>
<td>Clinical Setting</td>
<td>Trauma</td>
</tr>
<tr>
<td>Inpatient</td>
<td>Fibromyalgia</td>
</tr>
<tr>
<td>Outpatient</td>
<td>Parkinson’s disease</td>
</tr>
<tr>
<td></td>
<td>Patellar tendinitis</td>
</tr>
<tr>
<td></td>
<td>Rheumatoid arthritis</td>
</tr>
</tbody>
</table>
coefficients (0.97 and 0.97 respectively) for the measurements. On the other hand, the standard error of measurement (SEM) value, a measure of absolute reliability, was 6.6° for the two raters. The SEM in a measurement study such as this, reflects not only the disagreement among raters but also the imprecision with which each of the raters makes their measurement according to Eliaziszw et al\textsuperscript{11}. LaStayo and Wheeler\textsuperscript{26} believe the SEM, which is expressed in the unit of the measurement, may be the best indicator of reliability.

These statistics show that one can have excellent correlation values in the presence of a substantial amount of error. Others have made the same observation\textsuperscript{19}. Such an apparent contradiction is related to the fact that relative reliability measures are strongly influenced by the range of measured values. In our study, the measurements among subjects ranged from 22° to 140° for knee flexion. Although the ICC accurately reflects the degree to which individuals maintain their position in a sample on repeated measurements, it cannot be used to infer anything with regard to variability in the repeated measurements. Additionally some have pointed out that the ICC value is a dimensionless ratio and cannot be viewed as an index for judging instrument usefulness for individual patients\textsuperscript{18, 19}.

Altman’s and Bland’s protocol was used to estimate the agreement between raters, magnitude of bias, and the possibility of error varying according to the size of the angle being measured. The difference between raters’ mean measurements was less than 2 degrees. The limits of agreement, however, indicate that the difference between two raters’ measurements for a new individual from a similar population is expected 95% of the time to fall approximately between +19.3° and –17.5°, a range of 36.8°. Indryan\textsuperscript{20} calls these the limits of disagreement. The difference encountered between raters may have been partially related to the degree of forcefulness commanded by the rater when subjects were asked to actively flex their knees. Such large differences in joint measurements suggest poor agreement between raters. According to this analysis we cannot conclude that the measurements of the two raters were interchangeable.

A margin of error of $\pm 5\ degree$ has been deemed acceptable for measuring joint range of knee motion\textsuperscript{44}. Although cited in some studies, this value seems to be based on an arbitrary decision. Nevertheless, it has been proposed as a rule of thumb for clinicians. Estimations such as this have been criticized as devoid of statistical basis\textsuperscript{11}. The smallest detectable difference (SDD) has been used as an index to estimate with 95% confidence that an observed change on a subsequent measurement exceeds irrelevant, typical fluctuation. Though its adoption has been recommended in goniometric studies, its use is still new\textsuperscript{17}. The SDD value of 18° for the fluid goniometer was unexpectedly large. It means that a patient’s change in joint motion subsequent to an intervention would have to exceed a value of 18° to be accounted as a real change, if measured by two different therapists. The SDD statistic has been suggested as a cut off point for true change based on probability rules. The acceptance of the SDD as a valid criterion to measure change is still under debate\textsuperscript{19, 21}. Smith\textsuperscript{22} has suggested that an SDD less than 10% of the total range of measured values is acceptable. In our study the SDD was 10% of the range for intra-observer and 21% for inter-observer reliability Schreuders et al.\textsuperscript{19} pointed out that since the range includes the extreme values of a distribution, a more useful interpretation of the SDD value would be obtained by examining it in relation to the SD, which is less influenced by the range of measurements. According to this author, the SDD/SD ratio is more informative of whether measurements can be used for detecting clinical change. Adopting this ratio calls for determining another cut off, for which there is no suggestion at this time. It has been recognized, that adopting cut-off points to define change is more of a clinical than a statistical decision\textsuperscript{15}.

Our measurements using the fluid goniometer showed greater variability across repeated measurements for both raters than expected. Recent articles have reported values for the SEM of knee flexion ranging from 1.6 to 15 degrees measuring with the universal goniometer, but comparison between studies are difficult, since sample sizes and methodologies vary greatly\textsuperscript{21, 23, 24}. Many factors could account for variability in knee ROM within subjects, including the complexities inherent to knee motion, such as the large amount of soft tissue around the joint and the triaxial planes of motion or presence of pain. A study conducted on normals analyzing error would help identify how much knee pathology actually is responsible for variability in measurement. Similarly, a parallel study using the UG and FG may also help sort out differences inherent in measurement process with the two instruments and allow for more realistic instrument comparison.

In this particular study, we feel the greatest source of variation to be related to the lower leg oscillations evident during maximal active knee flexion. Measuring knee flexion with the fluid goniometer requires lying in the prone position. In this position, the knee flexors are less stable than in supine with the hip flexed, even though the therapist stabilized, the hip during the measurements. Because the lower leg is not externally stabilized it is free to oscillate at the end range, making the fluid level in this instrument vary greatly. Though the actual angular displacement of the limb may be minor, the linear displacement of the limb with each oscillation is appreciable, because the goniometer is placed mid-thigh, away from the fulcrum of the joint.

Though the fluid goniometer has been studied and used by some, its effectiveness for yielding precise measurement (reliability) of knee flexion in patients has not been established. In goniometer measurement studies, good correlation coefficients are not an indication of how free of measurement error a measurement can be. The amount of measurement variability found in our study makes the use of the fluid goniometer on the same patient by different therapists unadvisable. The prone position for measurement and the effects of gravity on the fluid element may interfere with the precision of measurement.
ACKNOWLEDGEMENTS

We want to express our appreciation to the following people for providing the sites and for permitting us to recruit their patients. Chris and Traci Mallett, owners of ATP Physical Therapy, Los Angeles, California, Joe Borland, owner of Borland Physical Therapy, Los Angeles, California, Paula Steeper and Romel Salcido, Director and Physical Therapy Assistant of Foothill Nursing and Rehabilitation, Los Angeles, California, and Michel Vincent, owner of Vincent Physical Therapy, Los Angeles, California. Lastly, we want to deeply thank all the subjects that willingly collaborated with this project taking time from their therapy sessions for the measurements, without which this study would not have been possible.

REFERENCES