The Thesis Committee for Keith Browning Macon
Certifies that this is the approved version of the following Thesis

Accuracy and Reliability of Single Camera Measurements of Ankle Clonus and Quadriceps Hyperreflexia

APPROVED BY
SUPERVISING COMMITTEE:

James Sulzer, Supervisor

Kathleen Manella

Accuracy and Reliability of Single Camera Measurements of Ankle Clonus and Quadriceps Hyperreflexia

by

Keith Browning Macon

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

August 2020

Dedication

To my family, who sparked and encouraged my childhood interest in science, and to my wife, whose support has kept me sane and whose companionship continues to help me grow.

Acknowledgements

My deepest gratitude is to Dr. James Sulzer for advising me throughout my master's program and for all the lessons, constructive criticism, and humor he shared. I could not have done this work without his help in first building up my foundations as an engineer and scientist.

I am indebted to Dr. Kathleen Manella for sharing her ideas, time, and perspective that made this thesis possible. It was her project that reignited my interest in research that led to where I am today, and undoubtedly where I will head in the future.

Special thanks to Dr. Sung Yul Shin, Jeonghwan Lee, and Dustin Hoang for their contributions to this work. Thanks as well to my colleagues in the Rewire Lab and elsewhere at UT, who provided excellent ideas, feedback, and friendship over the years. Thanks to my friends outside of UT, too, for providing me with social interactions and comfort outside of academia.

To my parents, Judy and Charlie Macon, thank you for providing me with a loving home and shaping me into who I am today. To my brother, Sam, thank you for helping me set ever increasing goals. Finally, to Brigid, thank you for enduring alongside me all this time. You make every day brighter than the last.

Abstract

Accuracy and Reliability of Single Camera Measurements of Ankle Clonus and Quadriceps Hyperreflexia

Keith Browning Macon, M.S.E.

The University of Texas at Austin, 2020

Supervisor: James Sulzer

In people with stroke, spinal cord injury, multiple sclerosis, and other upper motorneuron lesions, ankle clonus and quadriceps spasms may limit self-care and mobility tasks. The ankle clonus drop test, which measures the plantar flexor reflex threshold angle (PFRTA), and the pendulum test, which measures the quadriceps reflex threshold angle (QRTA), provide valid and reproducible measurements of ankle clonus and quadriceps hyperreflexia. However, measuring the PFRTA and QRTA requires high fidelity motion capture systems that are limited to laboratory settings by cost and complexity. The aim of this study was to evaluate a simple, single-camera based method of measuring ankle clonus and quadriceps spasticity in clinical settings. With synchronous 3-D inertial motion capture to provide a high fidelity reference, we used a smartphone camera and green stickers to measure the PFRTA and QRTA of 14 individuals with ankle clonus or quadriceps hyperreflexia in one or both legs. This resulted in test sessions on 22 impaired legs with four repetitions of each test on each leg conducted by a student physical therapist and an experienced physical therapist. We hypothesized that the smartphone camera measurements would provide clinically useful outcome measures for assessing ankle clonus and quadriceps spasticity. To assess accuracy of the camera-measurements, we computed the bias and limits of agreement between the camera and the inertial motion capture measurements. For reliability, we computed intra-rater and inter-sensor reliability coefficients in addition to the minimum detectable change. The smartphone PFRTA biases were smaller than 0.2° and the QRTA biases smaller than 1.2°. The limits of agreement for the PFRTA were ±4.66°/ ±7.49° (student/expert), and for the QRTA were

±4.40°/±4.67°. Reliability was similar between the camera and inertial measurements of tests by both rater types: intra-rater reliability ranged from 0.85-0.90 for the PFRTA and ranged from 0.96-0.98 for the QRTA. The inter-sensor reliability when measuring the PFRTA and QRTA was 0.97 and 0.99. The minimum detectable change for the PFRTA ranged from 7.10°-8.70°, while for the QRTA ranged from 7.65°-8.27°. Based on prior research, the limits of agreement and minimum detectable change were sufficiently low for purposes of interindividual, repeatable measurement. These data show that student and experienced physical therapists using ubiquitous existing hardware such as a smartphone can produce accurate, reliable assessments of ankle clonus and quadriceps hyperreflexia in a clinical environment.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

Stroke, spinal cord injury, multiple sclerosis, and other causes of upper motor neuron lesions result in loss of supraspinal modulation of spinal reflexes which manifests as hyperreflexia (Gracies, 2005). Ankle clonus and quadriceps spasm, forms of hyperreflexia, may limit performance of self-care and mobility tasks, thereby restricting independence and quality of life (Fee & Miller, 2004; Mayo et al., 2017). Precise, reproducible measures of ankle clonus and quadriceps hyperreflexia are necessary to evaluate the effectiveness of interventions directed at normalizing reflex excitability (Adams & Hicks, 2005; Patrick & Ada, 2016). The Modified Tardieu Scale utilizes kinematic measurements of "catch angles" or reflex threshold angles (RTAs) that predict functional impairment (Mehrholz et al., 2005). However, this scale has been shown to suffer from poor reliability when testing plantar flexor and quadriceps spasticity (Mehrholz et al., 2005; Yam & Leung, 2006). This has been attributed to the inaccuracy of goniometric measurements as well as the difficulty of applying consistent input kinematics to the test when manipulating the lower limbs (Ben-Shabat et al., 2013; Choi et al., 2018). Drop tests, where an object is released at rest from a prescribed height, have long been used to provide consistent initial conditions for experiments. Two such leg drop tests have been shown to reliably induce RTAs that quantify lower limb spasticity: the ankle clonus drop test, which elicits the plantar flexor reflex threshold angle (PFRTA), and the pendulum test, which elicits the quadriceps reflex threshold angle (QRTA). Both the PFRTA and QRTA have been shown to be valid and reliable quantifiers of spasticity (Bohannon et al., 2009; Manella & Field-Fote, 2013; Manella et al., 2017). However, PFRTA and QRTA measurements require high fidelity motion capture and are thus limited to laboratory settings or require expensive equipment.

Joint kinematics can be measured accurately with a variety of existing technologies. Optical motion capture systems provide the most accurate measurements (van der Kruk et al., 2018), but are not feasible in many clinics due to high cost, lengthy calibration, and technical expertise required. Inertial motion capture systems are nearly as accurate as optical motion capture (Ricci et al., 2016; Choi et al., 2018; Lee et al., 2019), and while not as expensive as optical motion capture, may still be cost prohibitive. They also require calibration for each user with static poses in predefined postures, uni-axial movements, or with additional sensors (Liu et al., 2019). Differences in patient pathologies prevent a one-type-fits-all approach to calibration, as predefined calibration postures and movements are difficult or impossible for some patients to perform (Picerno et al., 2019). Other less expensive, off-the-shelf technologies include time-of-flight sensors such as the Microsoft Kinect (Microsoft, Redmond, WA). However, the Kinect has been

shown to have dubious accuracy and reliability when measuring hip and lower joint kinematics (Bonnechère et al., 2014; Guess et al., 2017) and is limited to a 30 Hz sampling rate that is insufficient for measuring fast joint kinematics. These existing technologies have issues that prevent their widespread adoption into physical therapy clinics. To enable the routine use of the ankle clonus drop test and quadriceps pendulum test in clinical settings, a new motion capture system that is accurate, affordable, and easy for both patients and clinicians to use is required.

Our goal was to use a ubiquitous device, a smartphone camera, to measure the RTAs associated with the ankle clonus drop test and the quadriceps pendulum test in individuals with hyperreflexia. We then assessed the accuracy of these measurements using high fidelity 3-D inertial motion capture. We synchronously recorded ankle and quadriceps tests in the clinic using both sensors on 20 limbs of 14 different individuals who presented with ankle clonus and quadriceps hyperreflexia. Custom tracking and signal processing software was developed to process the test data and extract each RTA. Using repeated tests by both student and experienced physical therapists, we also evaluated the reliability of RTAs measured by the tracking software and the 3-D inertial motion capture. This study demonstrates that smartphone camera motion capture provides accurate and reliable measurements of plantar flexor and quadriceps RTAs with minimal setup and no calibration required.

METHODS

Participants

Fourteen individuals with clinical presentations of ankle clonus or quadriceps hyperreflexia in one or both legs were recruited for the study. Each individual was informed of the details of the study according to guidelines approved by the University of St. Augustine's Institutional Review Board and provided written consent. The participants included 8 males and 6 females with mean and standard deviation age = 41 ± 7.8 years, with chronic pathologies including stroke, spinal cord injury, multiple sclerosis, and transverse myelitis. Due to presentation of bilateral clonus and hyperreflexia in some participants, the ankle clonus drop test and quadriceps pendulum test were performed on a total of 20 impaired legs. Table 1 presents the participant demographic information.

P	S	Leg	Gender	Age (years)	Diagnosis	Months from Onset	Functional Status	Assistive Device	Orthotic
1	L	001	M	36	Stroke	32	ambulatory	LBQC	AFO
2	R	002	M	44	Stroke	109	ambulatory	SPC	AFO
3	L	003	F	55	Stroke	113	ambulatory	SPC	AFO
4	L R	004 005	M	37	SCI	96	non-ambulatory	Wheelchair	None
5	R	006	M	54	Stroke	46	ambulatory	None	AFO
6	R L	007 008	F	45	TM	25	non-ambulatory	Wheelchair	None
7	L	009	M	48	Stroke	26	ambulatory	None	None
8	R	010	M	41	Stroke	22	ambulatory	None	None
9	L	011	F	33	Stroke	23	ambulatory	None	None
10	R	012	M	27	SCI	81	ambulatory	None	AFO
11	R L	013 014	F	36	SCI	129	non-ambulatory	Wheelchair	AFO
12	R L	015 016	F	36	MS	300	ambulatory	SPC	AFO
13	L R	017 018	M	37	MS	52	non-ambulatory	Wheelchair	None
14	R L	019 020	F	48	MS	9	non-ambulatory	Walker	None

Table 1: Participant demographic information.
Key: P = participant, S = side of leg tested, TM = transverse myelitis, SCI = spinal cord injury, MS = multiple sclerosis, LBQC = large base quad cane, SPC = single point cane, AFO = ankle foot orthosis.

Experimental Setup

Each experiment was conducted at one of four local outpatient physical therapy clinics: Spero Rehab Austin, Spero Rehab Central Austin, St. David's Rehabilitation Hospital, and the University of St. Augustine for Health Sciences. A set of wireless commercial IMUs, Xsens MTw Awinda (Twente, the Netherlands), was used as a portable, high-quality reference signal to compare the output of our single-camera reflex tracking system (RT). To conduct each test, only equipment typically found in clinics was used. This included an adjustable height bench, a box step, a non-slip sheet (Dycem, Warwick, RI), and pillows as needed for the patient's comfort. The test area only required enough space for the camera phone to be positioned 1 meter away from the subject. Setup for the IMUs involved applying velcro bands and attaching the 7 IMUs to the subjects' pelvis, thighs, shanks, and feet. A laptop running Xsens' MVN Studio software was used to operate the IMUs and record their data. The IMUs were recalibrated using MVN Studio before testing each leg. Setup for the camera-based motion capture involved applying green adhesive stickers to the test leg at the greater trochanter, lateral knee joint line, lateral malleolus, lateral posterior aspect of the heel, and the lateral aspect of the 5^{th} metatarsal head. A typical setup using both sensors for each test is shown in Figures 1A and 1B.

Experimental Procedure

The ankle clonus drop test is designed to elicit a reproducible clonic response of the leg (Manella, 2011). The test involves raising the patient's leg until that the foot is 2 inches above a platform and releasing it such that the ball of the foot drops onto the platform edge. As the number of sustained clonic oscillations are also a useful outcome measure, the clonic response is recorded for at least 15 seconds after the drop before the leg is replaced into a rest position. To conduct the quadriceps pendulum test, the patient must be elevated such that their foot will not contact the ground, the tester grasps the foot and raises the knee into full extension before releasing. To avoid fatigue or other effects of repetition during both tests, there is a 1-minute rest period between repetitions. The exact procedure for both tests that was provided to clinicians is shown in Appendix A.

We used a 2 x 2 x 2 factorial design, with factors of rater type conducting the test (experienced physical therapist and student), joint (knee and ankle) and repetition (2 per condition). Each leg was tested a total of 8 times: 4 ankle clonus drop tests and 4 quadriceps pendulum tests. The order of tests was randomized such that the quadriceps test occurred before the ankle test in half the legs tested. For participants with

two legs tested, the IMUs were recalibrated before beginning the second leg. Each test produced time series data of all IMU-measured joint angles and a video, which each required further analysis to report the RTA and number of clonic oscillations outcome measures.

Reference Motion Capture

The Xsens Awinda IMUs feature an internal sampling rate of 1000 Hz that is digitally downsampled to 100 Hz. The IMU software, MVN Studio, includes its own proprietary signal processing to calibrate and manage drift of the sensors. The time series data recorded by the IMUs includes 3-D joint angles of the pelvis, both hips, knees, and feet. This data was exported to an XML format for later analysis in MATLAB.

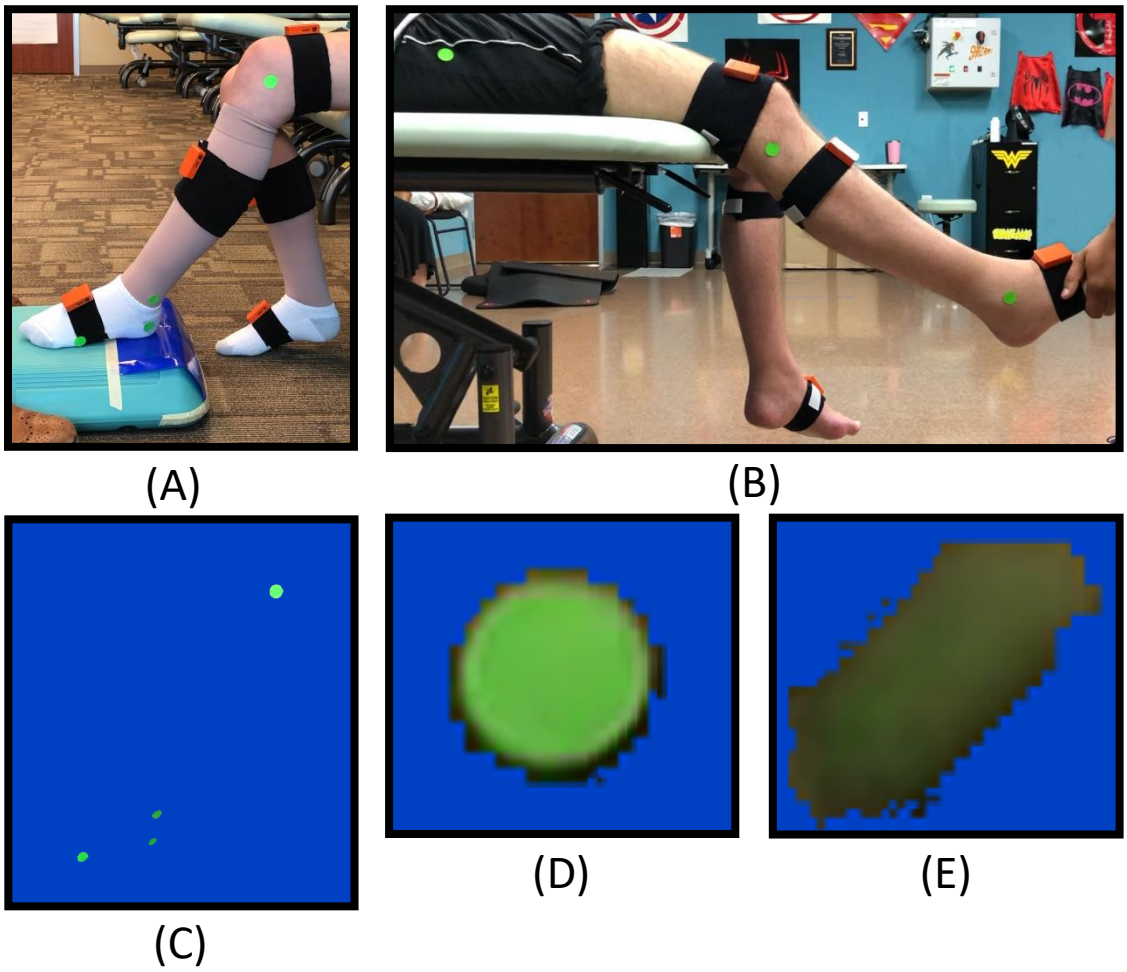

Figure 1: Experimental setups and single camera perspective. Ankle clonus drop test setup (A), quadriceps pendulum test setup (B), ankle drop test setup after applying color masking (C), stationary marker (D), marker experiencing motion blur (E).

Single-Camera Reflex Tracking System

The cameras used to record the drop tests were an iPhone 6 and iPhone 10 (Apple, Cupertino, CA) using their native camera app set to record at 60 frames per second. The native app was used both for familiarity to the videographer and to take advantage of its dynamic camera settings, which would adapt the camera ISO and shutter speed to each of the four test environments. Videos were then transferred onto a laptop for analysis. Each test video was processed using custom software written in Python, using several libraries including OpenCV (opencv-contrib-python), NumPy, and SciPy. For each test, the joint angle time series data were recorded together with the outcome measures. This section describes the algorithms used by the device to extract the RTA and number of oscillations of each test.

Tracking Algorithm: The reflex tracking software utilizes the Channel and Spatial Reliability Discriminative Correlation Filter (CSR-DCF) tracking algorithm (Lukežič et al., 2018) to obtain the pixel x-y coordinates of each green adhesive sticker in every frame of the test videos. When selecting the best tracking algorithm for this application, three primary factors were considered in order of importance: tracking accuracy (tendency to correctly hold the center of the marker), robustness (tendency to not lose track of the marker), and computational speed. Using pilot data from the first three subjects, three high-performing algorithms were identified: Kernelized Correlation Filters (KCF), Minimum Output Sum of Squared Error (MOSSE), and the SCR-DCF. The KCF and especially MOSSE algorithms were typically faster than CSR-DCF but were not as accurate or robust. Similarly, we found that converting each frame from RGB color space to the CIELAB color space, which has been shown to discriminate between colors better than RGB and other color spaces (Maldonado-Ramírez & Torres-Méndez, 2016), enabled more reliable tracking across the majority of test videos. Finally, in order to identify the markers associated with each joint, we developed a graphical user interface that allows the user to drag a box around each marker in order from hip to toe to identify each anatomical landmark and initialize the tracker's region of interest.

Masking: The lighting conditions at each test site were not always ideal; for example, some rooms were dimmer than others and lighting colors could vary. In some settings, the camera shutter speed would slow to ensure the camera received enough light which resulted in motion blur of the markers. Blurred markers sometimes interfered with the tracking algorithm performance (Figure 1D and 1E). To address this issue, we implemented a masking function that would threshold and subtract pixels that were not within the desired bandwidth of colors, demonstrated in Figure 1C. The color bandwidth of the masking function,

defined using three integers from 0-255, correspond to the CIELAB L* a* b* values that describe dark to lightness, green to redness, and blue to yellowness. Each value can be adjusted in the user interface to select what range of colors are to be kept after filtering. The masking function allows the user to remove all non-marker pixels that might lead to tracking confusion or failure. The result was improved tracking performance, especially in cases where the motion blur significantly distorted the shape and color of the markers.

Signal Processing: The ankle and knee joint angle data was extracted from the marker x-y coordinate time series data using the law of cosines. Additional signal processing was necessary to automate the extraction of the outcome measures: the RTA and the number of oscillations. We used the SciPy peak-finding function *find_peaks* to identify the important oscillations in the signal, which only required tuning the prominence, width, and window length to accommodate any expected data. However, while the RTA should be the first peak, we found that in some tests as the rater lifted the leg an unintentional but detectable oscillation in the joint angle could be generated through the leg's inertial dynamics. To rule out these potential false positives, we designed a detection algorithm around a minimum expected angular velocity of the test joint, as the peak joint velocities can always be expected during the initial leg release of the test. Through analysis of all test data collected, it was determined that a minimum velocity of 120 degrees/second correctly identified either the QRTA or PFRTA in 159/160 of all tests recorded. The algorithm defined the RTA as the first peak to occur after two subsequent frames that exceed the minimum angular velocity, which prevents false RTA detections due to noise or a later fast oscillation. The one failed detection had two frames with peak angular velocity of only 113.4 then 114.3 degrees/second, possibly due to excessive tone in the knee extensor muscles. The number of oscillations were then counted as the detected peaks that follow the RTA.

Statistical Methods

Measures of Agreement: Agreement between the RT and IMUs was evaluated using Bland-Altman plots. On a Bland-Altman plot, the difference (or error) of paired measurements are plotted on the vertical axis while the average of the two measurements are plotted on the horizontal. This allows one to reveal any fixed or proportional bias of one method relative to the other. The plot is developed further by calculating the 95% limits of agreement (LoA), which describe the expected range of error of a single measurement between the methods (Bland and Altman, 1986/2010). We generated four Bland-Altman plots to describe

the RT-IMU limits of agreement of different rater types (student/expert) conducting tests for each joint (knee/ankle). In order to focus the analysis solely on the agreement between each sensor, only the first repetition of each session was used. All Bland-Altman analyses were performed using the *BlandAltmanLeh* package in *R*. Agreement describes the relative accuracy of two methods. We defined acceptable LoA using the least significant difference (LSD) that we calculated using reported results from previous studies that demonstrated significantly different mean RTAs between impaired and control groups. If the LoA are smaller than the LSD, then the RT could diagnose the RTAs from the previous studies as impaired or normal with the same accuracy as the IMUs.

When an ANOVA reveals a significant F test that shows at least one group mean of a population is different from the others, Fisher's LSD can be used to define the minimum difference necessary for two group means to be considered significantly different. Fisher's LSD is known to have higher type-1 error compared to other post-hoc tests of group mean differences (Ramsey, 2007), and thus provides a conservative definition for acceptable LoA. Fisher's LSD is described by Equation 1 (Salkind, 2010, pp. 492-494):

$$LSD = t_{\alpha,df} \times \sqrt{MSW \times (1/n_A + 1/n_B)} \qquad (1)$$

where *t* represents a Student's t-value based on the within-groups degrees of freedom and a chosen significance level, *MSW* is the mean-squared-within error from a 1-way ANOVA, and n_A, n_B are the respective sample sizes of the groups being compared. We computed the 95% confidence LSD for the PFRTA and QRTA based on previous work (Fowler et al., 2000; Manella and Field-Fote, 2013) using Equation 1. We then hypothesized that the LoA of the ankle clonus drop test PFRTA and pendulum test QRTA would fall below their respective LSD, which would indicate that RT measurement error is small enough to classify RTA values as accurately as the IMUs.

Measures of Reliability: Reliability of the RT and IMUs were evaluated and compared using intraclass correlation coefficients (ICCs) and minimum detectable change (MDC_{95}). We used ICC(3,1) to describe the intra-rater reliability of both the RT and IMU measurements for each class of rater conducting each test. The first and second repetitions of each rater were used to calculate each sensor's ICC(3,1) for each rater type. The formula for ICC(3,1) for comparing two raters is given in Equation 2 (Portney and Watkins, 2000, p. 565):

$$ICC(3,1) = \frac{BMS - WMS}{BMS + WMS} \qquad (2)$$

where WMS is the within-subject mean sum squared error and BMS is the between-subject mean sum squared error. Equation 2 highlights that the ICC is based on the proportion of within-subject RTA variance and between-subject RTA variance, or the variance of the RTAs in the subject population. We also used ICC(3,k) to describe the inter-sensor reliability between RT and the IMUs. The inter-sensor reliability was calculated by comparing each subject's average IMU and RT measured RTAs. The equation for ICC(3,k), using the same definitions as Equation 2, is given in Equation 3 (Portney and Watkins, 2000, p.565).

$$ICC(3,k) = \frac{BMS - WMS}{BMS} \quad (3)$$

Here, inter-sensor reliability is based on the proportion of between-sensor RTA variance and between subject population RTA variance. Between-sensor variance is similarly used in the construction of the LoAs in Bland Altman plots. For both the intra-rater and inter-sensor ICCs, interpretations were defined as excellent for ICC ≥ 0.90, good for ICC ≥ 0.75, and moderate to poor ICC < 0.75 using guidelines provided by Portney and Watkins (2000, pp. 557-586).

In contrast to the ICCs, the MDC_{95} and the standard error of measurement (SEM) from which it is derived depend only on within-subject variance found from repeated measures. The SEM was estimated for each rater type using each sensor in each test by finding the within-subject standard deviation using 1-way ANOVAs. The MDC_{95} describes the smallest within-subject change of a single measurement that would indicate the quantity has changed enough to not be attributed to random noise, i.e. $MDC_{95} = 1.96\sqrt{2}\,SEM$. The MDC_{95} was compared to the pre-post changes in group mean PFRTA (Manella and Field-Fote, 2013) and QRTA (Ness and Field-Fote, 2009) values of previous studies that sought to evaluate treatment efficacy. We hypothesized that the MDC_{95} for RT and IMU measurements would not be significantly different from each other, indicating both methods are comparably reliable. Additionally, we hypothesized that each MDC_{95} would be smaller than the group average RTA changes measured in previous studies, indicating that both methods would detect the changes in a replication study.

Linear Mixed Modeling: We used linear mixed models to evaluate the fixed effects of test type, rater experience, and repetition on RTA disagreement between methods for both the ankle clonus drop test and quadriceps pendulum test data. Each model's random effects included the subject, the subject-rater interaction, and the subject-repetition interaction. It was hypothesized that the models would reveal no

statistically significant effect of rater experience or repetition on the RTA disagreement in both tests. The *R* package *afex* was used to construct the model and perform analyses ($\alpha < 0.05$). This process was repeated with RTA magnitude as a covariate to check for the presence of proportional bias, and again with non-flexion joint angles measured using the IMUs as covariates to quantify their effect on sensor disagreement. Each model was compared using the *anova()* function from the *stats* package to determine which best modeled the data.

Missing and Excluded Data: Some data was not included in the analysis due to rater error during testing. In two tests, the rater's hand made contact with the IMU attached to the participant's shank just as the RTA occurred. The ankle markers placed on Subject 003 for the ankle clonus drop test were attached to loose clothing that shifted between each test, which led to the exclusion of their ankle drop test data. We excluded the experienced rater QRTA measurements of Subject 007 due to the participant's inability to follow instructions. Due to the subjects' time constraints, four ankle clonus drop test repetitions (two from Subjects 001 and two from 002) and two pendulum test repetitions (Subject 001) were not recorded with IMUs. Finally, the data for Subject 018 was removed due to substantial magnetic interference affecting the IMUs at the test site. This interference was identified by MVN Studio and was avoided in future testing by relocating within the test site.

RESULTS

Representative Data

Representative time series of the aligned RT and IMU data are presented in Figures 2 and 3. PFRTA and QRTA errors were -1.1° and -1.4°, respectively. These figures show that the number of oscillations, another clinically relevant parameter, were identical between the two systems. However, Figure 2 shows a difference in the steady state error, with the RT ankle angle decreasing below the IMU data after 8 seconds. This sudden increase in sensor disagreement can be attributed to changes in the subject's hip abduction and internal rotation as they adjusted their leg posture.

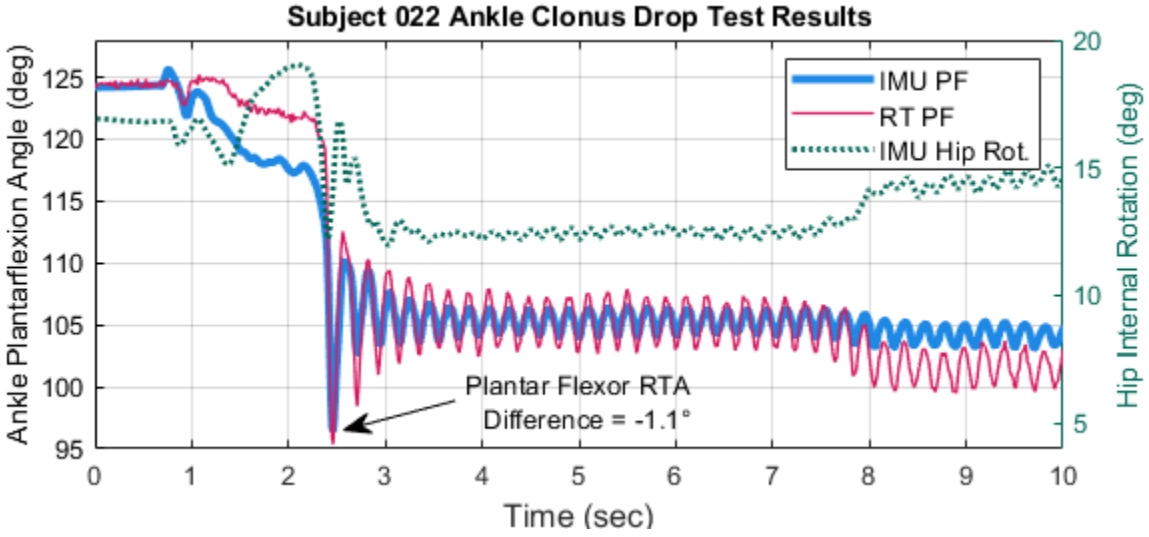

Figure 2: Representative results of the ankle clonus drop test measured by IMUs and our custom software. Hip internal rotation included to provide insight into mismatch between RT and IMU plantar flexion angle.

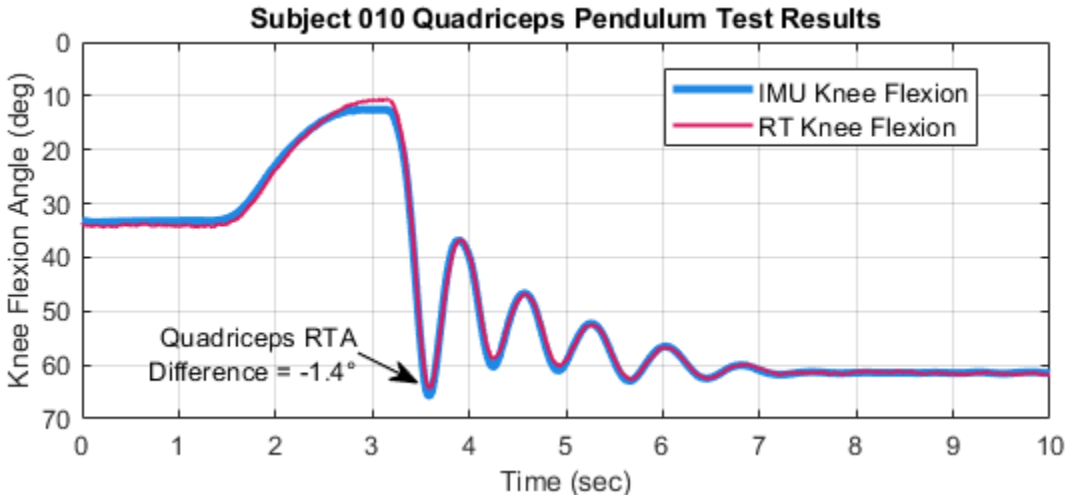

Figure 3: Representative results of the pendulum test measured by IMUs and our custom software.

Accuracy: Sensor Bias, Limits of Agreement, and Least Significant Difference

Four Bland-Altman plots for the PFRTA and QRTA measured by each type of rater using each method are presented in Figure 4. For the ankle clonus drop test PFRTA, the student rater tests had an RT-IMU bias of 0.18° ± 1.22° (mean ± 95% CI) and LoA from -4.48° ± 2.12° to 4.83° ± 2.12°. The experienced rater tests had a bias of -0.11° ± 2.04 and LoA from -7.60° ± 3.53° to 7.37° ± 3.53°. An F-test revealed no significant difference between the student and experienced LoA ($p = 0.084$). For the quadriceps pendulum test, the student rater tests had a bias of -0.47° ± 1.22° and LoA from -5.14° ± 2.12° to 4.20° ± 2.12°. The experienced rater tests had a bias of -1.15° ± 1.12° and LoA from -5.54° ± 1.93° to 3.25° ± 1.93°. Based on earlier work (Manella and Field-Fote, 2013; Fowler et al., 2000) we calculated LSD of the PFRTA and QRTA of 8.10° and 16.4°, respectively.

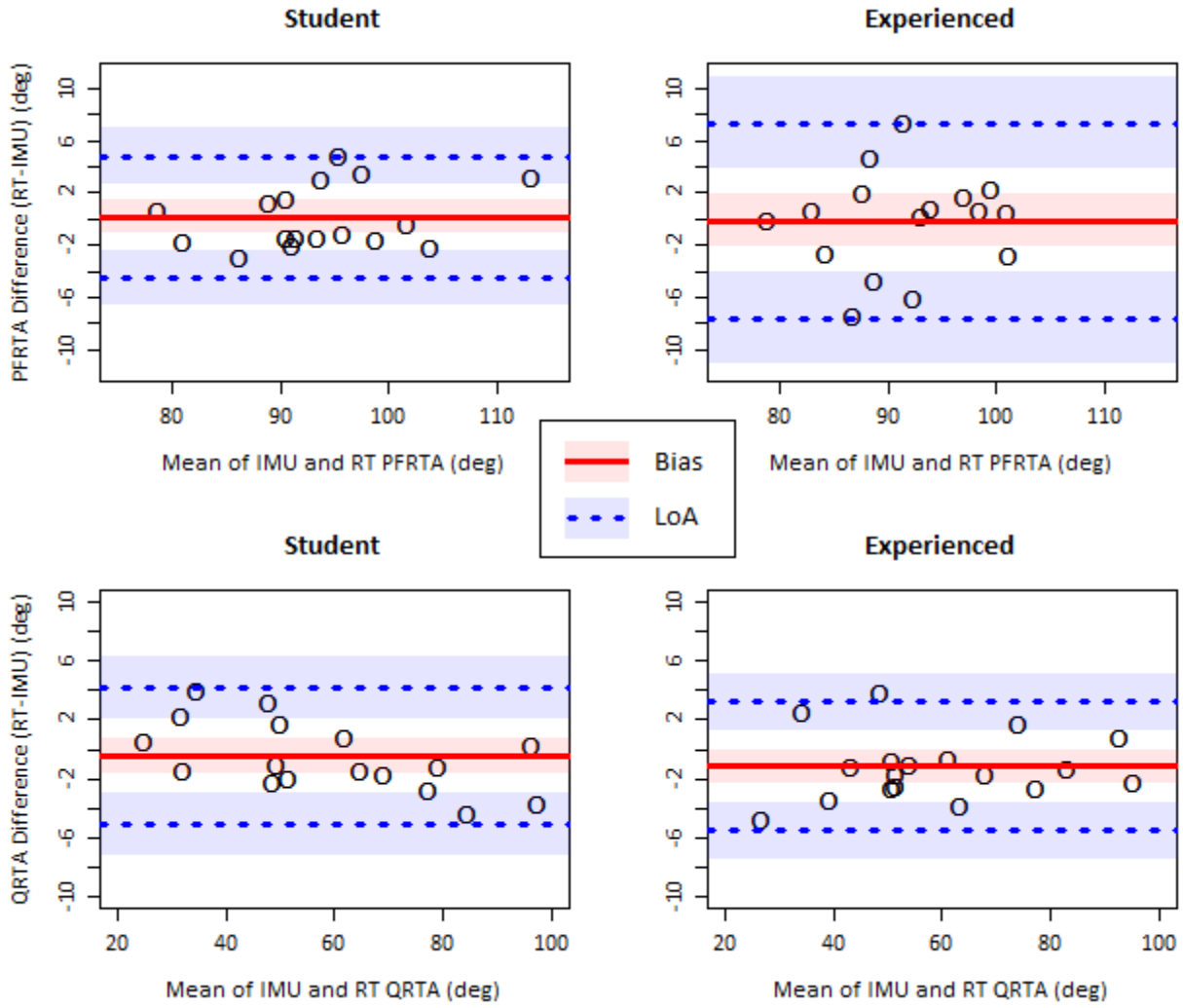

Figure 4: Bland Altman plots of RTA measurements made by the reflex tracking system (RT) and IMUs. The shaded regions represent 95% confidence intervals of the bias and LoA.

Reliability: Intraclass Correlation Coefficients and Minimum Detectable Change

The intra-rater reliability ICCs for each test, each sensor, and each rater type are presented in Table 2. These coefficients reflect the proportion of within-subject variance to between-subject variance of the RTA measurements, which are both confounded by both the sensor reliability and any variance in the initial conditions of the tests. The MDC_{95}, also reported in Table 2, is only affected by within-subject variance that is confounded by sensor reliability and any variance in the test initial conditions.

The inter-sensor reliability ICC(3,k) obtained by comparing the average measurements of RT and the IMUs were found to be 0.969 (95% CI from 0.931 to 0.986) for measurements of the PFRTA and 0.998 (95% CI from 0.995 to 0.999) for measurements of the QRTA. These ICCs reflect the proportion of between subject variance and between sensor variance.

Ankle Clonus Drop Test PFRTA Reliability

Rater Type	Sensor	ICC (95% CI)	MDC_{95}, deg. (95% CI)
Experienced	IMU	0.85 (0.70-0.93)	7.52 (5.38, 9.67)
Experienced	RT	0.90 (0.79-0.95)	7.10 (4.33, 9.87)
Student	IMU	0.87 (0.73-0.94)	7.75 (5.21, 10.30)
Student	RT	0.85 (0.69-0.93)	8.70 (5.37, 12.03)

Quadriceps Pendulum Test QRTA Reliability

Rater Type	Sensor	ICC (95% CI)	MDC_{95}, deg. (95% CI)
Experienced	IMU	0.96 (0.92-0.98)	6.10 (4.11, 8.08)
Experienced	RT	0.97 (0.93-0.99)	7.65 (4.82, 10.46)
Student	IMU	0.99 (0.97-0.99)	8.40 (5.82, 10.97)
Student	RT	0.98 (0.96-0.99)	8.27 (5.73, 10.81)

Table 2: Intra-rater reliability ICC(3,1) and the minimum detectable change (MDC_{95}) of each rater type using each sensor. Each pair of rows with the same rater type were parallel measurements made using the IMU and RT measurement methods.

Linear Mixed Effects Random Models

The PFRTA disagreement model that best described the data included fixed effects of rater type, repetition, and the ankle inversion angle, with an interaction effect of rater type and repetition. These final mixed model results are summarized in Table 3. The main effect of rater type was not statistically significant ($\beta = 0.65$, SE = 0.78, $p = 0.41$), repetition was not significant ($\beta = 1.02$, SE = 0.55, $p = 0.073$),

and their interaction was significant (β = -1.91, SE = 0.74, p = 0.02). Ankle inversion was determined to be a near-significant covariate (β = -0.066, SE = 0.034, p = 0.058) that likely improved the model's description of the data as indicated by X^2 test (p = 0.068). Ankle abduction was not a significant covariate and did not improve the model. Ankle plantar flexion angle, defined as the average of RT and IMU measured PFRTAs, was also not a significant covariate and did not improve the model.

For the model of the quadriceps pendulum test QRTA disagreement, we found no statistical significance in effects of rater type (β = 0.61, SE = 0.40, p = 0.13), repetition (β = 0.59, SE = 0.39, p = 0.14), or their interaction (β = -1.02, SE = 0.56, p = 0.074). The intercept was found to be significant (β = -1.15, SE = 0.52, p = 0.036). Modifying the model to include knee flexion magnitude, defined as the average of RT and IMU measured QRTA, showed no significance and did not improve the model. The final QRTA disagreement model included fixed effects of rater type, repetition, and their interaction effect, and is summarized in Table 3.

Linear Mixed Effects Random Model of PFRTA Disagreement

Fixed Effect	Estimate	Std. Error	DoF	P(>|t|)
Intercept	-0.039	0.693	36.0	0.956
Rater Type (student)	0.654	0.780	28.0	0.409
Repetition (rep2)	1.018	0.547	29.5	0.073
Ankle Inversion	-0.066	0.034	44.5	0.058
Rater Type*Repetition	-1.908	0.740	17.2	0.020

Linear Mixed Effects Random Model of QRTA Disagreement

Fixed Effect	Estimate	Std. Error	DoF	P(>|t|)
Intercept	-1.146	0.520	26.7	0.036
Rater Type (student)	0.613	0.398	50.1	0.130
Repetition (rep2)	0.589	0.391	50.0	0.138
Rater Type*Repetition	-1.020	0.558	50.1	0.074

Table 3: Fixed effect estimates for RT vs. IMU disagreement models of PFRTA (top) and QRTA (bottom)

DISCUSSION

The aim of this study was to evaluate a simple, single-camera based method of measuring ankle clonus and quadriceps spasticity in the clinic. This reflex tracking system used a smartphone to record video of the ankle clonus drop test and quadriceps pendulum test. The recorded video was analyzed with user-friendly, offline software. Comparing Reflex Tracker and inertial motion capture outcomes, we found measurement agreement for plantar flexor and quadriceps reflex threshold angles. Specifically, the limits of agreement and minimum detectable change were sufficiently low for purposes of interindividual, repeatable measurement. However, we also observed that out-of-plane motion has a small but measurable distorting effect on the accuracy of the Reflex Tracker. These data show that student and experienced physical therapists using ubiquitous existing hardware such as a smartphone can produce accurate and reliable assessments of ankle clonus and quadriceps spasticity in a clinical environment.

Agreement of RT and IMU Reflex Threshold Angle Measurements

PFRTA Agreement: The bias of RT-IMU disagreement was estimated to be near-zero in tests conducted by both student and experienced raters. The LoAs between the two rater classes did not significantly differ (student ± 4.66°, experienced ± 7.49°). The PFRTA LSD based on the findings of Manella and Field-Fote (2013) between normal and groups with ankle clonus was ± 8.10°, indicating both the student and professional PFRTA LoA are within the acceptable range. Thus, RT can assess clinically relevant differences in PFRTA with similar accuracy as inertial motion capture.

QRTA Agreement: The bias of RT-IMU disagreement was near-zero for the student tests at -0.47° and -1.15° for the experienced rater tests. This can be interpreted as RT tends to underestimate the QRTA by a small, clinically insignificant amount. The LoA between the rater classes did not significantly differ (student ±4.67°, experienced ±4.40°). The QRTA LSD based on the findings of Fowler et al. (2000) was ±16.41°, indicating both the student and professional QRTA LoA are well within the acceptable range. Thus, in addition to PFRTA, RT can assess clinically relevant differences in QRTA with similar accuracy as inertial motion capture.

Reliability of PFRTA and QRTA Measurements

Intra-Rater and Inter-Sensor ICCs: Intra-rater reliability of the pendulum test was shown to be excellent (ICC ≥ 0.90) for both rater types measured by both sensors, while the intra-rater reliability of the ankle

clonus drop test ranged from good (ICC ≥ 0.75) to excellent. The higher pendulum test ICCs highlight the presence of intrinsic differences between the ankle clonus drop test and the quadriceps pendulum test. As the MDC_{95} of each test were similar, this difference can be attributed to the greater between-subject variance of the QRTA compared to PFRTA measurements.

The inter-sensor reliability when comparing average measurements of the PFRTA and QRTA were both excellent with estimates of 0.969 and 0.998, respectively. This can be interpreted as the between-subject RTA variance was substantially higher than the variance between the sensor measurements. As between-sensor variance is also a key component in finding the LoA, this result shows both that RT and IMU have good reliability as well as good agreement relative to the differences between subjects.

MDC_{95} of the PFRTA: Estimating minimum detectable change is useful for categorizing changes in an individual's scores as significant (Rábago et al., 2015). The PFRTA MDC_{95} using RT were 8.7° and 7.10° for the student and experienced raters, respectively, while the IMU MDC_{95} were 7.75° and 7.52°. Manella and Field-Fote (2013) evaluated changes in 10 clonic individuals' PFRTA in response to 12 weeks of locomotor training. Of the 7 individuals who reduced their PFRTA, the average change after training was 10.7°. Due to a lack of available data, we cannot conclude how well the MDC of the RT can facilitate early evaluation of an intervention. However, these data suggest that the RT is capable of measuring clinically relevant within-patient changes in PFRTA and that there is no clinically significant difference between the IMU and RT systems when measuring either experienced or student raters conducting the tests.

MDC_{95} of the QRTA: The QRTA MDC_{95} using RT were 8.27° and 7.65° for the student and experienced raters, respectively, while the IMU MDC_{95} were 8.40° and 6.10°. Ness and Field-Fote (2009) evaluated changes in the QRTA (referred to as first swing excursion, or FSE) of 16 individuals with chronic spinal cord injury in response to a 3 day/week, 12-session whole-body vibration intervention. The significant average change in QRTA after four weeks of treatment was 12.05°. Analysis of their reported data revealed that 8 of the 13 participants who improved demonstrated QRTA changes that exceeded our highest MDC_{95}. The authors also measured the QRTA twice weekly, testing participants once 5 minutes after concluding a session and then again after 15 minutes. They observed significant mean changes of 7.07°, 8.73°, and 8.13° degrees between the 5-minute and 15-minute tests on weeks 1, 2, and 4. The individual patient scores were not reported for the weekly measurements, and so we are unable to use the MDC_{95} to

conclude how many individuals' improvement exceeded each MDC_{95}. As the average change in week 2 was larger than the $MDC_{95}$for both RT and IMUs, multiple subjects would have experienced changes detectable by both RT and IMUs. This response, measured once a week, could be used to facilitate early evaluation of patients' responsiveness to the whole-body vibration intervention. These data suggest that the RT is capable of measuring clinically relevant within-patient changes in QRTA before, during, and after treatment, and that there is no clinically significant difference between the IMU and RT systems when measuring either experienced or student raters conducting the tests.

Analysis of Measurement Error Sources Between RT and IMUs

Although RT has shown to be accurate and reliable enough for clinical use, it is still of interest to map out any quantifiable sources of error that might affect the system. Such information could be used to identify potential areas for improvement either in the measurement system or in the drop tests themselves. Linear mixed models were used to evaluate any effects of rater type, repetition, RTA magnitude, and out-of-sagittal plane motion on the PFRTA and QRTA sensor disagreement. The PFRTA and QRTA disagreement models demonstrated no significant relationship between RTA magnitude and sensor disagreement. This indicates there is no proportional bias of RT, which agrees with the random distributions visible in the Bland-Altman plots. The ankle inversion covariate was not significantly correlated with sensor disagreement, although it approached significance ($p = 0.058$). Geometric error was expected due to RT's use of a single camera and thus inability to accommodate out-of-plane motion. We found anecdotal evidence of geometric error (Figure 2), but the effect on the RTA disagreement was not statistically significant. The main effects of rater type and repetition were not significant in either PFRTA or QRTA models, although their interaction was significant ($p = 0.02$) in the model of PFRTA disagreement. The magnitude of this interaction is small, approximately 1° or less between the means of raters and repetitions, demonstrating that RT accuracy and reliability is consistent across rater types and repetitions. This consistency agrees with what was shown in the LoA, ICC, and MDC_{95} results. Finally, aliasing could be a source of error. In Appendix B we present calculations of the theoretical maximum aliasing error of the PFRTA and QRTA when recording typical drop tests using 30 Hz, 60 Hz, and 100 Hz sampling rates.

Limitations

While we found that the RT is an accurate and reliable method of RTA measurement, an important question regarding the utility of this device is its ability to predict recovery and/or evaluate the effectiveness of an intervention at an early stage. Such an evaluation requires a more intensive longitudinal approach and could be conducted as a future study. We used inertial motion capture as the basis of comparison for the RT. While IMUs are nearly as accurate as optical motion capture (Ricci et al., 2016; Choi et al., 2018; Lee et al., 2019), any system will have some joint kinematic error. Thus, we can only discuss the agreement between measurement systems, not the true error. While the order of ankle and knee drop tests were randomized across subjects, the order of rater type was constant with students always testing before experienced raters. As a result, the interaction effect between rater type and repetition could instead be an effect of testing order. Finally, during our analysis of error sources we could not quantify the effects of magnetic distortion or calibration error of the IMUs.

Conclusions

We developed a novel system known as the Reflex Tracker (RT) to unobtrusively and inexpensively measure reflex threshold angles in a clinical environment. We found RT performed with clinically acceptable accuracy and repeatability, with similar performance to commercial IMUs. We conclude that RT can be used as a tool to effectively measure changes in subjects' PFRTA and QRTA in a clinical environment. As RT requires only a smartphone camera and simple stickers, it provides clinicians with easy access to objective measurements of spasticity. Future work will evaluate the test-retest reliability of RT when there are multiple days between measurements. There is also a need for studies that evaluate changes in the PFRTA on a weekly basis during the course of treatment. Finally, RT could be applied to tests that elicit RTAs around other joints, such as the elbow or shoulder.

APPENDIX A: Procedures for conducting the drop tests

Ankle Clonus (Plantar Flexor) Drop Test Procedure

Tester (student or more novice clinician is first tester)
1. Position patient seated on the **narrow end of a hi-low table** (or armless chair) with both feet flat on the ground.
2. Place the **blue Dycem sheet** over the **narrow end** of the platform (or step).
3. Position the **narrow end** of the 4-inch platform (or step) underneath the bare (sockless) test foot in a rest position.
4. Tester position: stand facing narrow end of table, straddle platform to perform test
5. Apply **4 neon green ¾" circle labels** to the lower leg 1) lateral side of **knee joint line** (palpate the fibular head, place label just above it) 2) lateral **malleolus** at most prominent point 3) lateral side of posterior aspect of **heel** 4) lateral side of **5th metatarsal head** (palpate MT head just proximal to MTP joint)
6. When videographer is ready, place the ball of the test foot (forefoot) on the platform edge
7. Run a practice test without video to demonstrate test to participant and check camera set up
8. Practice Test with Videographer: (**straddle platform,** wait for videographer **"start"** command) 1) **GRASP test leg** 2 inches below the knee (**DO NOT COVER THE KNEE LABEL** at any time) 2) **LIFT the test leg up** about 4 inches, lifting the **entire foot off the platform about 2 inches** 3) **RELEASE** the test leg, **RETRACT your hands (DO NOT "Drive" the leg onto the platform)** 4) **CHECK** that the ball of the foot drops onto the platform edge
9. Trail 1: Start test when videographer says **"Start"**
10. Stop test when videographer says **"Stop"**
11. Rest 1 minute with whole foot on platform (**no stretch on plantar flexors**)
12. Trial 2: Repeat test steps 6, and 9-11
Videographer – using smart phone camera, download Google Drive mobile app onto phone
1. Adjust camera **video speed to 60 fps** (iPhone – Settings/Camera; Android – Camera Settings)
2. Lighting: maximize in test area (open shades, turn lights on, **do not shoot into the sun**, etc.)
3. Position camera up to 1 meter (3 ft) away from test leg so that test leg and markers fill the frame
4. Adjust position: fill the frame with the **test leg only,** and **all 4 green labels** visible
5. Run a practice test without video to check camera view, frame alignment, all labels visible 1) **ensure all 4 green labels** remain **in the frame throughout test** 2) **instruct tester to change hand position if needed so as not to cover up any markers**
6. Adjust as needed to **ensure all 4 circles remain in camera view throughout test**
7. Trial 1: 1) **Ask tester if ready** 2) S**tart video, watch timer** 3) **at 2 seconds g**ive **"Start"** command to tester 4) **at 15 seconds** give **"Stop"** command to tester
8. Rest 1 minute
9. Trail 2: Repeat test steps 6-8
SWITCH TESTER/VIDEOGRAPHER
1. Repeat ABOVE procedures for Practice Trial, Trial 1, and Trial 2
2. Complete Video Trials Form

Quadriceps Pendulum Test Procedure

Tester (student or more novice clinician is first tester)
1. Position patient supine (or semi-reclined to tolerance) on the **narrow end** of a **high-low mat table** 1) **Raise mat** table so **both lower legs dangle** off the mat 2) **Raise to about 12 inches of floor clearance** from feet
2. Tester position: stand facing narrow end of table to perform test, do not impede swinging leg
3. Apply 3 neon green ¾" circle labels to the leg 1) **add label** at **greater trochanter** (palpate GT and place label, over clothes is OK) 2) lateral **knee joint line** 3) lateral **malleolus** 4) **remove 2 labels**, from heel and 5th metatarsal head
4. Adjust position: 2-inch clearance between back of knee to edge of table (scoot down if needed)
5. Hold the forefoot up with knee in mid-range position until the Videographer is ready
6. When the Videographer is ready, place the test foot in dangle position
7. Run a practice test without video to demonstrate test to participant and check camera set up
8. Practice Test with Videographer: (**stand facing table**, wait for videographer **"start"** command) 1) **GRASP test FOREFOOT** (**DO NOT COVER THE MALLEOLUS LABEL** at any time) 2) **LIFT the test FOOT up, moving the knee into full extension** 3) **RELEASE** the test foot, **RETRACT your hands (DO NOT "Drive" the foot downward)**
9. Trail 1: Start test when videographer says **"Start"**
10. Stop test when videographer says **"Stop"**
11. Rest 1 minute: hold forefoot up with knee positioned in mid-range (**no stretch on quadriceps**)
12. Trial 2: Repeat test steps 6, and 9-11
Videographer – using smart phone camera
1. Position camera 1 meter (3 ft) away from test leg
2. Adjust position to 1) **fill the frame** with the **test leg only** and **all 3 green labels** visible
3. Run a practice test without video to check camera view, frame alignment, all labels visible 1) **ensure all 3 green labels** remain **in the frame throughout test** 2) **instruct tester to change hand position on foot if needed**
4. Adjust as needed to **ensure all 3 circles remain in camera view throughout test**
5. Trial 1: 1) **Ask tester if ready** 2) S**tart video, watch timer** 3) **at 2 seconds g**ive **"Start"** command to tester 4) **at 15 seconds** give **"Stop"** command to tester
6. Rest 1 minute
7. Trail 2: Repeat test steps 6-8
SWITCH TESTER/VIDEOGRAPHER
1. Repeat ABOVE procedures for Practice Trial, Trial 1, and Trial 2
2. Complete Video Trials Form

APPENDIX B: Theoretical aliasing error of RTAs due to insufficient sample rate.

In the introduction, we highlighted that a weakness of the Kinect is its maximum 30 Hz sampling rate. Aliasing error of the peak magnitude in time-series data can be predicted based on the ratio of the sampling frequency and the frequency of the measured signal. The theoretical maximum possible error of a sine wave peak is given in Equation C1:

$$Max\ Error\ \% \ = \ 100 \times (1 - cos(\pi/N\ )) \tag{4}$$

where N is defined as the number of samples taken per period of the measured signal, or the ratio of the sampling frequency divided by the signal frequency. For the signal amplitudes, we used typical joint angle changes associated with the initial release of each test: 30° of plantar flexion and 50° of knee flexion. For the kinematic signal frequencies, we used values of 6.25 Hz to model the ankle clonus drop test impact (Boyraz et al., 2015) and 1.5 Hz for the first swing of the pendulum test. For the sampling frequencies, we used the sampling frequencies of the IMUs, RT, and a Kinect. The resulting calculations using these values and Equation C1 are presented in Table C1.

Sensor	Sample Rate (Hz)	Max PFRTA Error (deg)	Max QRTA Error (deg)
IMU	100	0.58	0.06
RT	60	1.59	0.15
Kinect	30	6.20	0.62

Table 4: Theoretical maximum aliasing error of RTAs based on the sample rate of each system for typical magnitudes and frequencies of the ankle clonus drop test and quadriceps pendulum test.

As shown in Table C1, one can expect up to 6° of random aliasing error when using a 30 Hz sampling rate to measure the PFRTA. This 6° would cause consistent underestimation of the PFRTA, thereby reducing the accuracy of the system. The random nature of the aliasing error would also reduce the reliability of the PFRTA measurements by increasing the variance of repeated measurements. This analysis also reveals that the 60 Hz sampling rate of RT also contributes some error, though it is less than 2°. When measuring slower events such as the QRTA, aliasing error for each system is less than 1°, which is clinically negligible. Ultimately, this analysis demonstrates that any system designed to measure the ankle clonus drop test should use a sampling rate of 60 Hz or higher to avoid substantial aliasing error. As more smartphones come equipped with high speed video of 100 Hz and above, simply adjusting the camera app settings can allow for RT to minimize aliasing error.

REFERENCES

Adams, M., & Hicks, A. (2005). Spasticity after spinal cord injury. *Spinal Cord,* 43(10), 577–586.

Ben-Shabat, E., Palit, M., Fini, N., Brooks, C., Winter, A., & Holland, A. (2013). Intra- and Interrater Reliability of the Modified Tardieu Scale for the Assessment of Lower Limb Spasticity in Adults with Neurologic Injuries. *Archives of Physical Medicine and Rehabilitation*, 94(12), 2494–2501.

Bland, J., & Altman, D. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies,* 47(8), 931–936. (Original work published 1986)

Bohannon, R., Harrison, S., & Kinsella-Shaw, J. (2009). Reliability and validity of pendulum test measures of spasticity obtained with the Polhemus tracking system from patients with chronic stroke. *Journal of Neuroengineering and Rehabilitation*, *6*(1), 30–30.

Bonnechère, B., Jansen, B., Salvia, P., Bouzahouene, H., Omelina, L., Moiseev, F., Sholukha, V., Cornelis, J., Rooze, M., & Van Sint Jan, S. (2014). Validity and reliability of the Kinect within functional assessment activities: Comparison with standard stereophotogrammetry. Gait & Posture, 39(1), 593–598.

Choi, S., Shin, Y., Kim, S., & Kim, J. (2018). A novel sensor-based assessment of lower limb spasticity in children with cerebral palsy. *Journal of Neuroengineering and Rehabilitation*, *15*(1), 45–16.

Fee Jr, J., & Miller, F. (2004). The Leg Drop Pendulum Test performed under general anesthesia in spastic cerebral palsy. *Developmental Medicine and Child Neurology,* 46(4), 273–281.

Gracies, J. (2005). Pathophysiology of spastic paresis. II: Emergence of muscle overactivity. *Muscle & Nerve*, 31(5), 552–571.

Greenan Fowler, E., Nwigwe, A., & Wong Ho, T. (2000). Sensitivity of the pendulum test for assessing spasticity in persons with cerebral palsy. *Developmental Medicine and Child Neurology*, *42*(3), 182–189.

Guess, T., Razu, S., Jahandar, A., Skubic, M., & Huo, Z. (2017). Comparison of 3D Joint Angles Measured with the Kinect 2.0 Skeletal Tracker Versus a Marker-Based Motion Capture System. *Journal of Applied Biomechanics*, 33(2), 176–181.

Ismail Boyraz, Hilmi Uysal, Bunyamin Koc, & Hakan Sarman. (2015). Clonus: definition, mechanism, treatment. *Medicinski Glasnik*, 12(1), 19–26.

Lee, J., Shin, S., Ghorpade, G., Akbas, T., & Sulzer, J. (2019). Sensitivity comparison of inertial to optical motion capture during gait: implications for tracking recovery. *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR),* 139–144.

Liu, Y., Zhang, Y., & Zeng, M. (2019). Sensor to segment calibration for magnetic and inertial sensor based motion capture systems. *Measurement*, *142*, 1–9.

Lukežič, A., Vojíř, T., Čehovin Zajc, L., Matas, J., & Kristan, M. (2018). Discriminative Correlation Filter Tracker with Channel and Spatial Reliability. *International Journal of Computer Vision*, *126*(7), 671–688.

Maldonado-Ramírez, A., & Torres-Méndez, L. (2016). Robotic Visual Tracking of Relevant Cues in Underwater Environments with Poor Visibility Conditions. *Journal of Sensors*, *2016*, 1–16.

Manella, K., & Field-Fote, E. (2013). Modulatory effects of locomotor training on extensor spasticity in individuals with motor-incomplete spinal cord injury. *Restorative Neurology and Neuroscience*, *31*(5), 633–646.

Manella, K., Roach, K., & Field-Fote, E. (2017). Temporal Indices of Ankle Clonus and Relationship to Electrophysiologic and Clinical Measures in Persons with Spinal Cord Injury. *Journal of Neurologic Physical Therapy*, *41*(4), 229–238.

Manella, Kathleen Joy, (2011) Operant Conditioning of Tibialis Anterior and Soleus H-reflex Improves Spinal Reflex Modulation and Walking Function in Individuals with Motor-Incomplete Spinal Cord Injury. Open Access Dissertations. 679.

Mayo, M., DeForest, B., Castellanos, M., & Thomas, C. (2017). Characterization of Involuntary Contractions after Spinal Cord Injury Reveals Associations between Physiological and Self-Reported Measures of Spasticity. *Frontiers in Integrative Neuroscience, 11,* 2.

Mehrholz J, Wagner K, Meissner D, et al. (2005). Reliability of the Modified Tardieu Scale and the Modified Ashworth Scale in adult patients with severe brain injury: a comparison study. *Clin Rehabil*, 19:751-9

Mehrholz, J., Wagner, K., Meißner, D., Grundmann, K., Zange, C., Koch, R., & Pohl, M. (2016). Reliability of the Modified Tardieu Scale and the Modified Ashworth Scale in adult patients with severe brain injury: a comparison study. *Clinical Rehabilitation*, 19(7), 751–759.

Ness, L., & Field-Fote, E. (2009). Effect of whole-body vibration on quadriceps spasticity in individuals with spastic hypertonia due to spinal cord injury. *Restorative Neurology and Neuroscience*, *27*(6), 621–633.

Patrick, E., & Ada, L. (2016). The Tardieu Scale differentiates contracture from spasticity whereas the Ashworth Scale is confounded by it. *Clinical Rehabilitation,* 20(2), 173–182.

Picerno, P., Caliandro, P., Iacovelli, C., Simbolotti, C., Crabolu, M., Pani, D., Vannozzi, G., Reale, G., Rossini, P., Padua, L., & Cereatti, A. (2019). Upper limb joint kinematics using wearable magnetic and inertial measurement units: an anatomical calibration procedure based on bony landmark identification. Scientific Reports, 9(1), 14449–10.

Portney, L. G., & Watkins, M. P. (2000). Chapter 26: Statistical Measures of Reliability. In *Foundations of clinical research: Applications to practice* (pp. 557-586). Upper Saddle River, NJ: Prentice Hall.

Rábago, C., Dingwell, J., & Wilken, J. (2015). Reliability and Minimum Detectable Change of Temporal-Spatial, Kinematic, and Dynamic Stability Measures during Perturbed Gait. PloS One, 10(11).

Ramsey, P. (2007). Fisher's LSD. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (Vol. 1, pp. 360). Thousand Oaks, CA: SAGE Publications, Inc.

Ricci, L., Taffoni, F., & Formica, D. (2016). On the Orientation Error of IMU: Investigating Static and Dynamic Accuracy Targeting Human Motion. *PloS One*, *11*(9), e0161940–.

Salkind, N. J. (2010). Fisher's least significant difference test. In *Encyclopedia of research design* (Vol. 1, pp. 492-494). Thousand Oaks, CA: SAGE Publications, Inc.

van der Kruk, E., & Reijne, M. (2018). Accuracy of human motion capture systems for sport applications; state-of-the-art review. *European Journal of Sport Science*, 18(6), 806–819.

Yam WK, Leung MS. (2006). Interrater reliability of Modified Ashworth Scale and Modified Tardieu Scale in children with spastic cerebral palsy. *J Child Neurol,* 21:1031-5.